Imperial College London

Department of Computing

Dense Vision in Image-guided Surgery

Ping-Lin Chang

2014

Submitted in part fulfilment of the requirements for the degree of PhD in Computing and the Diploma of Imperial College London. This thesis is entirely my own work, and, except where otherwise indicated, describes my own research.

Copyright

Abstract

Image-guided surgery needs an efficient and effective camera tracking system in order to perform augmented reality for overlaying preoperative models or label cancerous tissues on the 2D video images of the surgical scene. Tracking in endoscopic/laparoscopic scenes however is an extremely difficult task primarily due to tissue deformation, instrument invasion into the surgical scene and the presence of specular highlights. State of the art feature-based SLAM systems such as PTAM fail in tracking such scenes since the number of good features to track is very limited. When the scene is smoky and when there are instrument motions, it will cause feature-based tracking to fail immediately.

The work of this thesis provides a systematic approach to this problem using dense vision. We initially attempted to register a 3D preoperative model with multiple 2D endoscopic/laparoscopic images using a dense method but this approach did not perform well. We subsequently proposed stereo reconstruction to directly obtain the 3D structure of the scene. By using the dense reconstructed model together with robust estimation, we demonstrate that dense stereo tracking can be incredibly robust even within extremely challenging endoscopic/laparoscopic scenes.

Several validation experiments have been conducted in this thesis. The proposed stereo reconstruction algorithm has turned out to be the state of the art method for several publicly available ground truth datasets. Furthermore, the proposed robust dense stereo tracking algorithm has been proved highly accurate in synthetic environment (< 0.1 mm RMSE) and qualitatively extremely robust when being applied to real scenes in RALP prostatectomy surgery. This is an important step toward achieving accurate image-guided laparoscopic surgery.

Dedication

To humankind and the karma.

Acknowledgements

It has been a tough 4-year PhD experience for me surviving from continuous research and financial issues. The experience is nevertheless invaluable for my future career. If this thesis happens to be any useful to anyone, the credits go to the people mentioned here.

First and foremost, I would like thank my first supervisor Dr. Philip "Eddie" Edwards who has introduced so many harsh problems to my life while being very positive and supportive to me. I still remember the first meeting when Eddie showed me the idea of camera tracking for endoscopic AR image-guided surgery. It was a blast since I had never seen such bloody images in prostatectomy and I was in fact highly doubting the feasibility of this idea. Over the years, we have gradually approached a solid solution and I think the achievement should be sourced from Eddie's original brave idea.

Great appreciation goes out to my second supervisor Prof. Andrew Davison who has helped me develop insights into visual SLAM problems. The elite team, Robot Vision Group, he has built at Imperial College London has never failed in conveying astonishing research results. I have learned ideas in plenty from other team members in the weekly reading group. His heaven-sent-rain financial aid had also released my huge burden about funding so I could had concentrated on my research in my third year.

The hardship subsidies granted by the Department of Computing for my first three years are also largely appreciated.

I would like to also thank Dr. Danail Stoyanov who has offered me a job in UCL in the last stage of my PhD while being very tolerant to allow me to finish my PhD thesis in the generous spare time allowed.

Special thanks go to Dr. Ankur Handa and Dr. Steven Lovegrove. Ankur has enlightened my mathematical skill, well, if originally I had any. I always enjoy the discussion with him about the connection between different computer vision algorithms in mathematical insights. Steve was the first person who introduced dense

methods to me. His lightweight OpenGL wrapper library, Pangolin, has turned out to be very useful for my research.

Last but not least, thanks to my family, my father Yu-Wei Chang, my mother Hue-Mei Lee, my sister Ping-Chen Chang and my wife Jung A Ho. This thesis would not be possible to finish without their continuous spiritual support. I sincerely thank their understanding and tolerance in helping me chase my stubborn dream.

Contents

List of Figures

Glossary

AD	Absolute Difference
API	Application Programming Interface
AR	Augmented Reality
AWGN	Additive White Gaussian Noise
BOBYQA	Bound Optimisation BY Quadratic Approximation
CPU	Central Processing Unit
CT	Computerised Tomography
CUDA	Compute Unified Device Architecture
DoF	Degree of Freedom
DTAM	Dense Tracking and Mapping
ESM	Efficient Second-Order Minimisation
FCA	Forward Compositional Algorithm
GLSL	OpenGL Shading Language
GPGPU	General-Purpose Computing on Graphics Processing Units
GPU	Graphics Processing Unit
ICA	Inverse Compositional Algorithm
ICP	Iterative Closest Point

IGS	Image-guided Surgery
IMU	Inertial Measurement Unit
IRLS	Iteratively Reweighted Least Squares
JBF	Joint Bilateral Filter
MAE	Mean Absolute Error
MAP	Maximum a Posteriori
MIS	Minimally Invasive Surgery
MRI	Magnetic Resonance Imaging
OpenGL	Open Graphics Library
PTAM	Parallel Tracking and Mapping
RALP	Robot Assisted Laparoscopic Prostatectomy
RMSE	Root Mean Square Error
ROF	Rudin-Osher-Fatemi
SLAM	Simultaneous Localisation and Mapping
TGV	Total Generalised Variation
TRE	Target Registration Error
TV	Total Variation
WTA	Winner-Take-All
ZNCC	Zero-mean Normalised Cross-Correlation

1 Introduction

Contents

1.1 Introduction

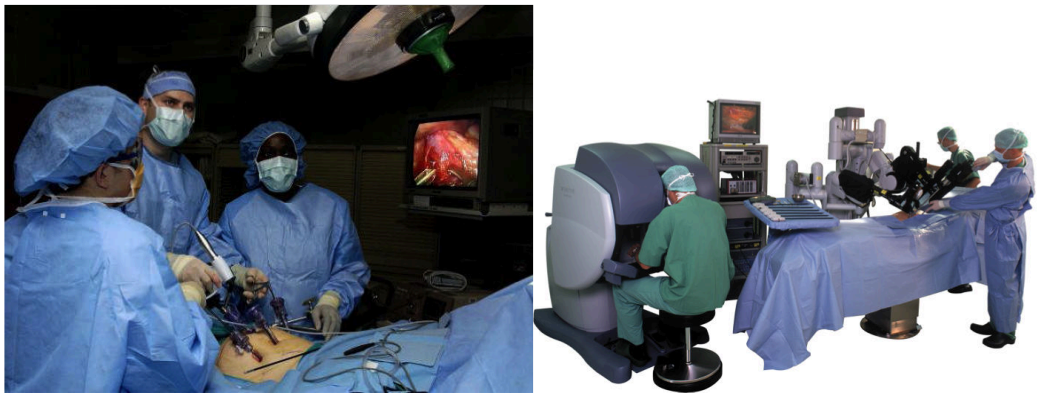
Minimally invasive surgery (MIS) is an increasingly popular treatment option due to reduced operative trauma compared to traditional open surgery, and can provide benefits of lower expense, shorter recovery, and reduced incidence of post-surgical complications. In order to perform an operation through small incisions in the skin, MIS uses endoscopic devices to indirectly observe the surgical scene. Due to the nature of live endoscopic video, however, there are severe constraints on the surgeon's spatial perception and reduced operative visual information. Figure 1.1 shows a general scene in MIS.

In the example of robot-assisted laparoscopic prostatectomy (RALP), although the da Vinci system provides a magnified 3D visualisation along with intuitive scaled manual interaction, the rates of complication from this procedure are still comparable to open surgery. Furthermore, a systematic review has shown current clinical RALP stringently requires more efforts in methodological standards for clinical research on new urologic procedures and devices (Kang et al., 2010). In this scenario, an auxiliary system providing additional visual information could be advantageous and may improve the results of RALP beyond that of open surgery.

The first image-guided surgery can be sourced back to 1895 when a patient had surgery to remove an inserted needle in her hand by surgeons using an X-ray print to help locate the needle. This can be seen as the first surgical procedure guided by an image and happened just two months after X-rays were discovered (Webb, 1988).

Nowadays surgeons in the operation theatre can easily acquire various modalities to help with surgery. Besides X-rays, high quality preoperative 3D computerised tomography (CT) and magnetic resonance Imaging (MRI) models enable surgeons to analyse the patients' anatomy, physiology and pathology. These 3D models are, however, independent of the surgeons' viewing position. In a usual scenario, surgeons refer those preoperative models off-line as images on a separate screen or light box and perform the surgery using their mental impression of the patient's anatomy and pathology.

Real-time interactive guidance modalities including ultrasound, intravascular ultrasound (IVUS) or X-ray fluoroscopy provide instant visualisation of the underlying anatomy so that the surgical tools and the operative targets are both visible in the image and the intervention proceeds until the task is finished. However, these intraoperative imaging modalities often suffer from quality issues, such as poor resolution, 2D visualisation and low signal-to-noise ratio. In general the quality of intraopera-



(a) (b)

Figure 1.1: Typical MIS (a) and robotic-assisted MIS (b).

tive imaging is poor compared to that which can be obtained preoperatively.

Different imaging modalities depict different anatomic structures, depending on their physical design principle. A desirable image-guided system would be able to integrate different modalities into a novel view which summarises information for the surgeons. Such an ideal system has to solve one big problem first - how to unify those modalities which have different coordinates. Although this is a classic image registration problem which has been studied for decades, it still remains a difficult challenge for image guidance which may have issues such as failed correspondence matching, deformation modelling and so on.

An alternative is to use a navigation system which can track the surgeon's viewing frame so it can augment virtual information or preoperative models into the right place. Once the virtual objects are locked in, the navigation system takes over the alignment task frame-by-frame. As a consequence, the augmented objects will stick to the correct pose in 3D no mater how surgeon's viewing angle changes.

1.2 Vision-based tracking for image-guided surgery

In the early days, stereotactic frames were regarded as the most accurate tracking method for surgery guidance (Jensen et al., 1996). In this method, a stereotactic frame is rigidly attached to the patient to provide a reference frame for targeting points as shown in Figure 1.2a. Specifically, the frame is screwed into the skull prior to brain imaging, which is generally performed immediately before the operation. Reference points on the frame are identified in these images, as is the target point within the lesion and entry path. A system consisting of two accurately gauged

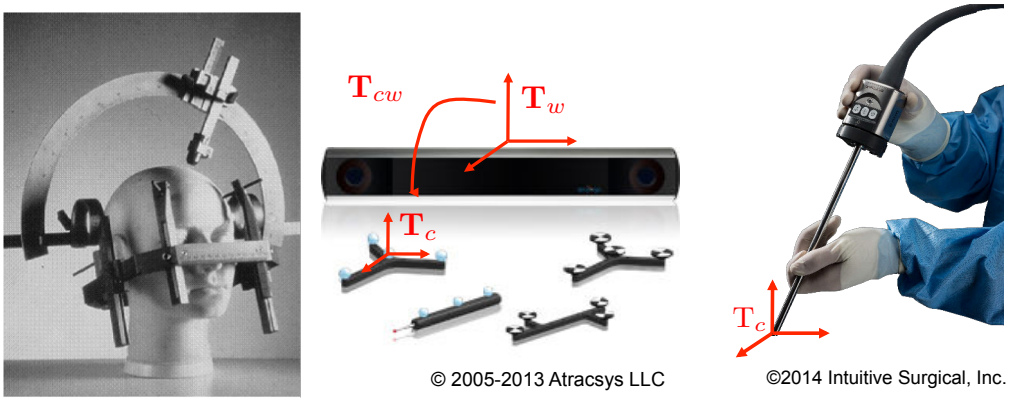
protractors allows the alignment of a biopsy needle along the desired approach and insertion to a predetermined depth. The frame installation is unavoidably an invasive procedure and a somewhat unpleasant experience for the patient. The size of the frame also creates problems for more open surgery and is infeasible for MIS. Therefore such a fixed frame system is limited. A frameless system would clearly be preferable.

A frameless option requires a tracking system to define a coordinate system within the operating theatre. The most popular tracking systems for IGS use optical images. The basic idea is to find a point in the tracker's multiple 2D camera views, from which simple triangulation will give the 3D position of the point relative to the cameras. The localised points can be either active (infra-red emitting diodes, IREDs) or passive (highly reflecting spheres or high contrast points) as shown in Figure 1.2b. The main difficulty with optical tracking is that line-of-sight between the tracker's cameras and tracked objects must be maintained. An additional possible error source with the reflecting sphere system is partial occlusion of the sphere or contamination with blood, which may shift the measured location. In addition, the accuracy of tracking a target diminishes as you get further away from the markers due to rotational error Fitzpatrick et al. (1998). As a result, this method is not suitable for accurately tracking the endoscope or laparoscopic camera in MIS since any external markers must be a long way from the end of the scope to be visible.

A particular goal of this thesis is to approach endoscopic/laparoscopic camera tracking in a different way. We aim to provide a relatively practical pure vision-based tracking system for image-guided surgery. As shown in Figure 1.2c, while the camera exposes the scene, it also locates itself on-the-fly. Such a tracking method is termed *vision-based ego-motion estimation* or *visual odometry* in the field of robotic vision. Furthermore, if the camera can track self-pose while maintaining an environment map for correcting drift, the method is known as *visual simultaneous localisation and mapping* (SLAM).

1.3 Why dense vision?

Camera ego-motion estimation is a widely studied problem in the computer vision and robotic vision community. It turns out that the essence of a tracking problem is equivalent to image registration, in which consequent camera frames are aligned with a reference frame by solving some mathematical models (Section 2.4). Based on this principle, the algorithms for the registration task can be categorised into two main approaches - *Dense methods* and *feature-based methods*.



(a) (b) (c)

Figure 1.2: Different tracking methods for IGS. (a) Stereotactic frame (Leksell). (b) Optical tracking system (Atracsys). (c) The preferable ego-motion tracking method.

Feature-based methods obtain the registration transformation by feature matching as a front-end step. A feature point in an image is detected and represented by a feature descriptor. Given two images with their set of feature points, the matching task is to match the most similar features in the two images according to the feature descriptor. The transformation between the two images can then be found by solving an overdetermined linear system formed by the feature correspondences (Davison, 2003; Klein and Murray, 2007).

As a counterpart, dense methods directly obtain the transformation from image intensities. A pixel-wise cost function is devised and optimised as the degree of similarity between all pixels in the two images, together with a certain motion parameterisation for warping either image in order to align the two. The optimisation starts from an appropriate initial estimate (within the convex basin) so that a gradient-descent-based algorithm can find the global extreme of the cost function (Comport et al., 2010; Newcombe and Davison, 2010; Newcombe et al., 2011b,a).

Which methodology is superior to the other is still an open debate. Feature-based methods globally match the descriptors so that they allow very wide baseline pose estimation such as in a large scale *structure-from-motion* task. In contrast, dense methods, by their nature, are ill-suited for wide baseline image matching. However, while feature-based methods apparently require a texture abundant scene in order to collect enough robust correspondences, dense methods that exploit all pixel intensities for the optimisation are more suitable for tracking in a texture-poor scene. It is clear that we can at least conclude that the method with best performance really depends on the target application. Decent discussions and comparisons

of the two methods in the literature can be found in (Irani and Anandan, 2000; Tri, 2000). Handa also presented a systematic summary for the advantages of dense methods (Handa, 1999, Section 4.1).

The state of the art feature-based method assumes good scene conditions such that the scene is static, there are a sufficient number of feature points to track and the lighting is constant (Klein and Murray, 2007). However in endoscope/laparoscope images, a texture-poor, deforming, smoky and specular scene is expected. When applying the feature-based method to such a scene, it is doomed to failure due to all the bad conditions that will appear in the endoscopic/laparoscopic scenes. In addition, the small baseline assumption can be perfectly met in a camera tracking application, where the frame-rate and motion speed are such that consecutive frames are very close to one another. Furthermore, using the da Vinci's stereoscopic camera enables us to reconstruct an instant 3D scene structure on-the-fly so there is no need for a pose/map initialisation step (Klein and Murray, 2007; Newcombe et al., 2011b). We therefore have focused on dense methods for realising an endoscopic/laparoscopic tracking system.

1.4 Contributions

This thesis aims to provide a robust camera ego-motion estimation system for augmented reality in IGS particularly for RALP. The original thought was to use the-state-of-the-art visual SLAM system PTAM, to bootstrap a number of camera poses for subsequent 3D to 2D preoperative models registration followed by dense tracking using the registered models. The preliminary research described in Chapter 3 has found that PTAM cannot provide sufficiently accurate pose estimations and that discrepancies between the preoperative model and the scene cause registration failure and camera tracking drift. Nevertheless, in the initial studies we have also found that dense tracking is extremely robust by itself even in an endoscopic scene. It is therefore promising to track the endoscopic camera using dense methods.

Starting from Chapter 4, we have redirected our approach to the problem. Instead of registering the preoperative models into the endoscopic scenes, a state-of-the-art stereo reconstruction algorithm using the stereoscopic camera equipped on the da Vinci robotic platform has been proposed. According to comprehensive validation studies, the proposed variational-based dense method, which uses ZNCC as the data term and Huber norm as the regulariser term, has shown the best performance among other previous algorithms and imaging sensors.

In Chapter 5, based on the early work of Comport et al. (2010), a stereoscopic

dense visual odometry system is proposed. It exploits all the stereo image data for visual odometry, by using the quadrifocal geometry constraint together with the dense stereo reconstructed model. The dense methods have turned out to be extremely robust even within a terribly ill-conditioned scenes which involves instrument motion, tissue deformation, specular highlights, and smoke. Consequently the proposed system has overcome all the problems we met in Chapter 3 and constitutes a state-of-the-art camera tracking system in IGS.

1.5 Publications

This thesis has led to the following publications:

1. **Ping-Lin Chang**, Dongbin Chen, and Philip "Eddie" Edwards. **Registration of a 3D preoperative model with 2D endoscopic images using parallel tracking and mapping (PTAM) with colour-consistency**. In Online Proceedings of the Medical Image Analysis and Understanding (MIUA), 2011.

2. **Ping-Lin Chang**, Dongbin Chen, Daniel Cohen, and Philip "Eddie" Edwards. **2D/3D registration of a preoperative model with endoscopic video using colour-consistency**. In Proceedings of the Augmented Environments for Computer-Assisted Interventions (AE-CAI) in Conjunction with MICCAI, volume 7264, pages 1-12, 2012.

3. Qinquan Gao, **Ping-Lin Chang**, Daniel Rueckert, S Mohammed Ali, Daniel Cohen, Philip Pratt, Erik Mayer, Guang-Zhong Yang, Ara Darzi, Philip "Eddie" Edwards. **Modeling of the bony pelvis from MRI using a multi-atlas AE-SDM for registration and tracking in image-guided robotic prostatectomy**. Computerized Medical Imaging and Graphics (CMIG), 37(2):183-194, 2013.

4. **Ping-Lin Chang**, Danail Stoyanov, Andrew J. Davison, and Philip "Eddie" Edwards. **Real-time dense stereo reconstruction using convex optimisation with a cost-volume for image-guided robotic surgery**. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 8149, pages 42-49, 2013.

5. L. Maier-Hein, A. Groch, A. Bartoli, S. Bodenstedt, G. Boissonnat, **P.-L. Chang**, N.T. Clancy, D. S. Elson, S. Haase, E. Heim, J. Hornegger, P. Jannin, H. Kenngott, T. Kilgus, B. Müller-Stich, D. Oladokun, S. Röhl, T. R. Dos

Santos, H.P. Schlemmer, A. Seitel, S. Speidel, M. Wagner and D. Stoyanov. **Comparative validation of single-shot optical techniques for laparoscopic 3D surface reconstruction**. To appear in IEEE Transactions on Medical Imaging (TMI), 2014.

6. **Ping-Lin Chang**, Ankur Handa, Andrew J. Davison, Danail Stoyanov, and Philip "Eddie" Edwards. **Robust real-time visual odometry for stereo endoscopy using dense quadrifocal tracking**. In Proceedings of the Information Processing in Computer-Assisted Interventionas (IPCAI), pages 11-20. 2014.

1.6 Thesis structure

Relevant background reviews will be introduced at the beginning of each chapter. This thesis is organised as follows

Chapter 2 introduces the materials and mathematical tools that will be used through out this thesis.

Chapter 3 describes our preliminary trail to approach the camera tracking problem in IGS.

Chapter 4 provides a novel stereo reconstruction algorithm which outperforms the previous methods.

Chapter 5 presents the overall dense visual odometry system that is the-state-of-the-art for endoscopic camera tracking in RALP.

Chapter 6 concludes this thesis and gives some future directions for the possible extensions of this thesis.

2 Preliminaries

Contents

2.1 Introduction

This chapter describes materials and mathematical concepts that will be used throughout this thesis. Notation used here will be complied with as much as possible in the subsequent chapters. One key factor to achieve an efficient image guidance system is the use of 3D graphics programming which has been significantly improved in terms of both algorithms and hardware architectures in the last few decades by researchers from the computer graphics community. Implementation details for a real-time image guidance system are also described here. I recommend that readers study this chapter before proceeding the remainder of the thesis.

2.2 Coordinate system

Coordinate systems or coordinate frames are the foundation of analytic geometry. Works in this thesis use a Cartesian coordinate system to span $\mathbb{R}^n$ to represent n-dimensional data. A Cartesian coordinate system specifies each point uniquely in $\mathbb{R}^n$ space by a n-vector in which each entry in the vector is the signed distance from the point to n mutual perpendicular axes (basis vectors), measured in the same unit of length. The point where all axes intersect is the world origin which is denoted by a n-dimensional zero vector $\mathbf{0}$.

In a $\mathbb{R}^n$ space, points are defined with respect to the *reference frame* which consists of the basis vectors. For example, points $a = (1, 1)$, $b = (3, 2)$ and $c = (-1, -4)$ in $\mathbb{R}^2$ can be shifted to $a = (0, 0)$, $b = (2, 1)$ and $c = (-2, -5)$ if the point a is regarded as the world origin of the reference frame. Rigid transformation between each local coordinate system is possible. When interacting with different local frames, careful attention to coordinate definition is required, particularly when we are dealing with 3D graphical data which may be defined in *world coordinates*, *camera coordinates* and *model coordinates*. It becomes surprisingly difficult to figure out what is wrong once the coordinate systems are messed up.

3D coordinate systems can be expressed by the special Euclidean $\mathbb{SE}(3)$ Lie group. Its minimum parameterisation is $\mathbf{x} \in \mathfrak{se}(3)$ Lie algebra in which it is a 6-vector $\mathbf{x} = (\nu, \omega) \in \mathbb{R}^6$ consisting of $\nu \in \mathbb{R}^3$ for the linear velocity and $\omega \in \mathfrak{so}(3)$ for the angular velocity of the motion. The smooth and invertible rigid-body 4×4 transformation matrix $\mathbf{T} \in \mathbb{SE}(3)$ can be obtained by the exponential map of $g(\mathbf{x})$:

$$\mathbf{T}(\mathbf{x}) = \exp(g(\mathbf{x})) = \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^\top & 1 \end{pmatrix} \in \mathbb{R}^{4\times 4},$$

where $\mathbf{R} \in \mathbb{SO}(3)$ and $\mathbf{t} \in \mathbb{R}^3$. See Appendix A for details of Lie groups.

2.2.1 3D coordinate convention

Without loss of generality, we define the camera coordinate system with OpenGL's convention, which is a right-handed coordinate system with its origin located at the optical center as shown in Figure 2.1a. The positive x-axis points right. The positive y-axis is defined as the up vector of the camera. The negative z-axis has the same direction as the camera faces and coincides with the optical axis. The 2D image coordinate has an origin at top-left corner where the u-axis points right and v-axis points down.

2.2.2 Transformation between coordinates

We use a notation $\mathbf{T}_{\{\cdot\}}$ to denote an individual frame and $\mathbf{T}_{\{a\}\{b\}}$ to indicate transformation from b frame to a frame. With this notation, it is crystal clear that a world frame $\mathbf{T}_w$ transformed to a model frame $\mathbf{T}_m$ by $\mathbf{T}_{mw}$, followed a transformation $\mathbf{T}_{cm}$ from the model frame to a camera frame, is equivalent to the transformation $\mathbf{T}_{cw}$ from the world frame to the camera frame. Figure 2.1b shows the relationship. Sometimes we need to transform coordinates in the forward or backward direction. The inverse of $\mathbf{T}$ has a closed form

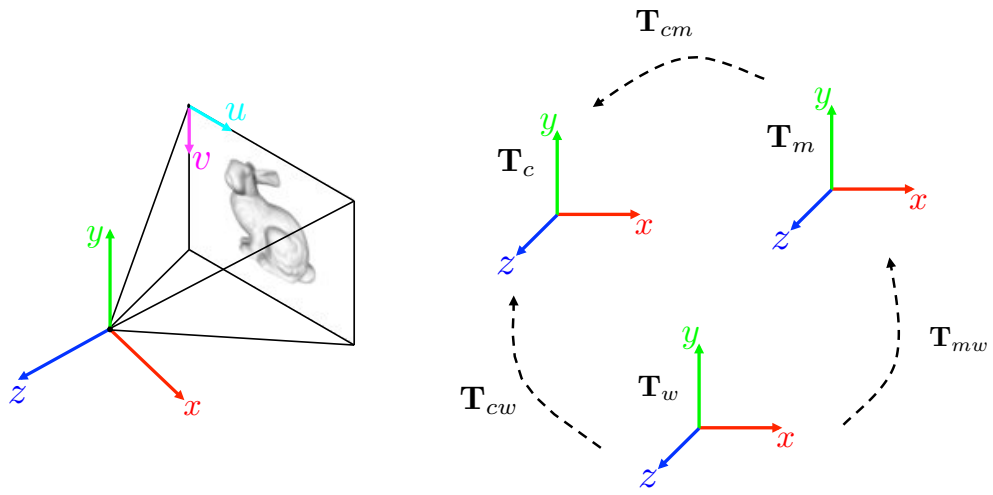
$$\mathbf{T}^{-1} = \begin{pmatrix} \mathbf{R}^\top & -\mathbf{R}^\top\mathbf{t} \\ \mathbf{0}^\top & 1 \end{pmatrix}.$$

Given a transformation $\mathbf{T}_{cw}$, the optical center position of a camera frame with respect to the world frame can be inferred as $-\mathbf{R}_{cw}^\top\mathbf{t}_{cw}$.

A 3D point $\mathbf{P}_w = (x, y, z)^\top$ in the world frame can be transformed to the camera frame using the homogeneous representation

$$\mathbf{P}_c = \pi(\mathbf{T}_{cw}\dot{\mathbf{P}}_w),$$

where the $\dot{\{\}}$ operator simply appends 1 in the augmented vector, so $\dot{\mathbf{P}}_w =$



(a) (b)

Figure 2.1: (a) Coordinate convention used in this thesis. (b) Coordinates transformation notation.

$(x, y, z, 1)^\top$ and the projection function $\pi : \mathbb{R}^n \to \mathbb{R}^{n-1}$ is defined as

$$\pi(a) = \frac{1}{a_n} \begin{pmatrix} a_1 \\ \vdots \\ a_{n-1} \end{pmatrix}.$$

2.3 Pinhole camera model

The most common camera projection model is the *pinhole camera model*, which projects a 3D point $\mathbf{P}$ in the camera frame to a 2D point $\mathbf{p}$ using a camera intrinsic matrix $\mathbf{K}$ such as

$$\mathbf{p} = \pi(\mathbf{KP}),$$

where the $\mathbf{K}$ is

$$\mathbf{K} = \begin{pmatrix} f_u & \alpha & u_0 \\ 0 & f_v & v_0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The 2D coordinates u_0 and v_0 is the *principal point* position of the camera in 2D image space which is the center of projection or the nearest point on the image plane to the pinhole. In general this position does not need to be at exact and is often set to the image center. The f_u and f_v are the horizontal and vertical focal length scaling the projection pixel size $p_w \times p_h$. They assume the original constant bidirectional focal length f that is related to

$$f_u = \frac{f}{p_w} \quad \text{and} \quad f_v = \frac{f}{p_h},$$

and in most camera sensors it gives $p_w \approx p_h$ (square pixel). The α is a skew parameter for shearing the image plane, which is usually ignored and set to zero. As will be used in Chapter 5, the inverse of $\mathbf{K}$ ($\alpha = 0$) is

$$\mathbf{K}^{-1} = \begin{pmatrix} \frac{1}{f_u} & 0 & -\frac{u_0}{f_u} \\ 0 & \frac{1}{f_u} & -\frac{v_0}{f_v} \\ 0 & 0 & 1 \end{pmatrix}.$$

We calibrate the camera intrinsic matrix $\mathbf{K}$ with a publicly available Matlab camera calibration toolbox[1] which also provides a function to do stereo camera rectification.

2.4 Mathematical models for dense vision

In the literature dense methods are also referred as *direct* methods since a vision problem is solved by analysing the intensity of each pixel directly. The most early examples of dense methods can be sourced back to Horn and Schunck (1981) and Lucas and Kanade (1981) in which they introduced dense algorithms to tackle different vision problems in concurrent streams, but they surprisingly ended up with similar principles and theoretical bases.

2.4.1 Taxonomy of models

In the early eighties, Horn and Schunck (1981) proposed to estimate general motions in 2D images at the pixel level, which is known as *optical flow* in the literature. As such an inverse problem is ill-posed due to homogeneous areas and the *aperture*

[1] Camera Calibration Toolbox for Matlab: http://www.vision.caltech.edu/bouguetj/calib_doc

problem[2], as well as a sum of squared data term, they introduced an explicit regulariser term to smooth pixel flow vectors within a local neighbourhood and better constrain the problem. The regularised cost function was minimised by solving the Euler-Lagarnge equations. Their original work has had a significant impact on the development of variational approaches in dense vision. Subsequent work from others has mainly been dedicated to looking for a better regularisation scheme.

In contrast, Lucas and Kanade (1981) were trying to solve a stereo matching problem without an explicit regulariser. In their original paper, the photometric cost function was formulated pixel-wise as the sum of squared difference between a parametrically warped patch in one image and the target template patch in the other image. The optimisation was performed by locally linearising the cost function with respect to the parametric patch motion model and moving the parameters in the gradient descent direction using the Newton-Raphson method. With a rectified stereo rig, the correspondence search can be found along a 1D search line, whereas when the matching problem goes beyond 1D such as in a non-rectified case, the warping-patch scheme can be much more efficient than the naive exhaustive search. They also showed that for 2D optical flow estimation, the correspondences can be found by warping the patch to find the best displacement of (u, v) in sub-pixel accuracy.

At a glance the two research streams look very different but in fact they share common assumptions and theoretical principles. They are both algorithms for image registration. They both used the sum of squared differences cost function to model the problems together with a Lambertian reflectance model to ensure lighting consistency between two images. In addition, although the original work of Lucas and Kanade (1981) did not explicitly model the local smoothness assumption, Lucas (1984) later found that the local patch indeed implicitly behaves as a smooth regulariser in response to the window size. More interestingly, when we look them from Bayesian inference point of view, they are both a *maximum a posteriori* (MAP) estimator.

Furthermore, Bergen et al. (1992) provided a clear categorisation for different models used in direct methods:

[2] Aperture problem is a phenomenon that when we observe a moving slant line through an aperture, its local motion has ambiguity up to the two perpendicular directions of the perceived motion.

Fully parametric models

Lucas-Kanade-like methods solve a model fitting problem assuming all pixels undergoing a global parametric motion model such as translation, rotation, affine, homography, etc.

Quasi-parametric models

Between fully parametric and non-parametric, there are methods lying in quasi-parametric models where there is a prior 3D rigid object to be parameterised in order to constrain the pixel motion in the 2D image. These models can be applied to a 3D to 2D rigid image registration problem (Chang et al., 2012; Prisacariu and Reid, 2012).

Non-parametric models

Horn-Schunck-like methods do not assume any specific global motion model but let each pixel have its own motion. Since the problem is highly under-constrained, in general the models require an explicit smooth regulariser term to be estimated together with a global variational function.

2.4.2 Bayesian inference interpretation

Since we are working with sensors to imagine scenes, the measurements always come along with some kind of uncertainties such as noise. Interpreting the problem in a statistical way is therefore beneficial. Assume that we would like to estimate a state vector $\boldsymbol{\xi}$ given a measurement vector $\boldsymbol{\mu}$. Treating the $\boldsymbol{\xi}$ and $\boldsymbol{\mu}$ as random variables enables us to analyse different hypotheses $\boldsymbol{\xi}$ so that the hypothesis with highest probability given the observation $\boldsymbol{\mu}$ will be the most likely solution we are looking for.

The hypothesis can be analysed by *maximum a posteriori* (MAP) estimation. Mathematically speaking this is equivalent to maximising the *posterior*:

$$\mathring{\boldsymbol{\xi}} = \arg\min_{\boldsymbol{\xi}} p(\boldsymbol{\xi}|\boldsymbol{\mu}). \tag{2.1}$$

According to Bayesian inference theory, the conditional probability $p(\boldsymbol{\xi}|\boldsymbol{\mu})$ can be written as

$$p(\boldsymbol{\xi}|\boldsymbol{\mu}) = \frac{p(\boldsymbol{\mu}|\boldsymbol{\xi})p(\boldsymbol{\xi})}{p(\boldsymbol{\mu})}, \tag{2.2}$$

where $p(\boldsymbol{\xi})$ is the *prior* and $p(\boldsymbol{\mu}|\boldsymbol{\xi})$ the *likelihood* (also called *data model*) expressing how well the measurements $\boldsymbol{\mu}$ can be explained by the state $\boldsymbol{\xi}$. The denominator $p(\boldsymbol{\mu})$ is known as the *evidence* which normalised the conditional probability. Since it does not depend on $\boldsymbol{\xi}$, it can be ignored during the optimisation.

MAP estimation for non-parametric models

From Equation 2.2 we can see that the performance of the MAP estimator is determined by the generative data model $p(\boldsymbol{\mu}|\boldsymbol{\xi})$ and the prior $\boldsymbol{\xi}$. In principle, if we had perfect data model and prior model, the MAP estimation would find us a global optimum. However, the perfect models are unlikely to be achieved in practice since an ill-posed non-parametric model problem is in general very complicated.

As we will be dealing with a reconstruction problem in Chapter 4, let us assume $\boldsymbol{\xi} = \mathcal{D}$ where the state vector $\boldsymbol{\xi}$ is the disparity map we would like to estimate, and the pixel-wise measurement $\boldsymbol{\mu}$ as the initial disparity estimation from a simple stereo matching algorithm such as plane sweep, followed by a WTA scheme. Assuming the initial disparity $\boldsymbol{\mu}$ has pixel-wise additive Gaussian noise and each measurement is *independent and identically distributed* (i.i.d), the likelihood is

$$p(\boldsymbol{\mu}|\boldsymbol{\xi}) = \prod_{\boldsymbol{\xi}} \frac{1}{\sqrt{2\pi}\alpha} \exp(-\frac{(\boldsymbol{\mu}-\boldsymbol{\xi})^2}{2\alpha^2}). \tag{2.3}$$

The prior knowledge of a disparity map could be that the disparities should be smooth in a local region in terms of the first order quadratic smoothness assumption:

$$|\nabla\boldsymbol{\xi}| = \sqrt{\left(\frac{\partial\boldsymbol{\xi}}{\partial u}\right)^2 + \left(\frac{\partial\boldsymbol{\xi}}{\partial v}\right)^2}. \tag{2.4}$$

Again we assume the prior is also a Gaussian distribution:

$$p(\boldsymbol{\xi}) = \prod_{\boldsymbol{\xi}} \frac{1}{\sqrt{2\pi}\beta} \exp(-\frac{|\nabla\boldsymbol{\xi}|^2}{2\beta^2}). \tag{2.5}$$

Incorporating Equation 2.3 and 2.5 into 2.2, we arrive at

$$p(\boldsymbol{\xi}|\boldsymbol{\mu}) = \prod_{\boldsymbol{\xi}} \frac{1}{2\pi\alpha\beta} \exp(-\frac{|\nabla\boldsymbol{\xi}|^2}{2\beta^2} - \frac{(\boldsymbol{\mu}-\boldsymbol{\xi})^2}{2\alpha^2}). \tag{2.6}$$

Performing MAP estimation on Equation 2.6 is equivalent to minimising its negative-log form as a cost function:

$$E(\boldsymbol{\xi}) = \sum_{\boldsymbol{\xi}} \left\{ \frac{1}{2}|\nabla\boldsymbol{\xi}|^2 + \frac{1}{2\lambda}(\boldsymbol{\mu}-\boldsymbol{\xi})^2 \right\}, \tag{2.7}$$

where the λ is $\frac{\alpha^2}{\beta^2}$. Equation 2.7 turns out to be a discrete configuration of Tikhonov regularisation (Tikhonov, 1943) which is a special form of Equation 4.16. The prior acts as a regulariser and the likelihood is nothing but the data term. The λ originally consists of the variances now acts like a parameter controlling the strength of the regulariser and data term.

The derivation shows that the original Tikhonov variational model used in Horn-Schunck's work makes the assumption that the underlying data model and smoothness prior has a Gaussian distribution. However, this is generally not true. The probability distributions do not necessarily need to be Gaussian and we will discuss a few other better choices to model the inverse problem in Section 4.3.3.

MAP estimation for parametric models

Parametric models can be solved by the MAP estimation in the similar way. As we will be dealing with the camera tracking problem in Chapter 5, let us assume $\boldsymbol{\xi} = \mathbf{x}$ where the state vector $\boldsymbol{\xi}$ is a 6-vector in $\mathfrak{se}(3)$. The n-pixel measurement is defined by the photometric residual between the warped current image $\mathcal{I}$ and the reference image $\mathcal{I}^*$:

$$\mu_i(\boldsymbol{\xi}) = \mathcal{I}(w(\boldsymbol{\xi}, \mathbf{p}_i)) - \mathcal{I}^*(\mathbf{p}_i), \quad \text{where} \quad i = 1, ..., n. \tag{2.8}$$

The warping function w takes the global motion model $\boldsymbol{\xi}$ to transform all pixel $\mathbf{p}_i$. With the i.i.d property and assume the pixel-wise residual is a Gaussian distribution, the likelihood is

$$p(\boldsymbol{\mu}|\boldsymbol{\xi}) = \prod_i \frac{1}{\sqrt{2\pi}\gamma} \exp(-\frac{\mu_i^2}{2\gamma^2}), \quad \text{where} \quad i = 1, ..., n. \tag{2.9}$$

The parametric model MAP estimation can be conducted without considering prior knowledge, in which it is equivalent to only maximising the likelihood. The cost function form of Equation 2.9 is therefore

$$E(\boldsymbol{\xi}) = \frac{1}{\sqrt{2\pi}\gamma} \sum_i \Big(\mathcal{I}(w(\boldsymbol{\xi}, \mathbf{p}_i)) - \mathcal{I}^*(\mathbf{p}_i)\Big)^2, \quad \text{where} \quad i = 1, ..., n. \tag{2.10}$$

Equation 2.10 is exactly the same cost function used in Equation 5.2. Note that the coefficient term $\frac{1}{2}$ or $\frac{1}{\sqrt{2\pi}\gamma}$ does not affect the position of global minimum. The prior $p(\boldsymbol{\xi})$ can be introduced over the estimated camera motion of course. Possible choices include an additional sensor such as an IMU (uniform prior), or prediction from a Kalman filter (Gaussian prior).

2.5 Graphics and parallel computing

When implementing a real-time IGS system, it is crucial to leverage the computation resources for computer vision algorithms and AR graphics. The work in this thesis relies heavily on novel graphics technologies in both software and hardware.

Dense methods can significantly benefit from the increasing sophistication of computer graphics algorithms and graphics hardware. For example, when one wants to visualise the results with complex 3D objects, modern rendering algorithms provide optimised drawing-on-screen pipeline with hardware acceleration. Hence with proper graphics programming the IGS system will not stall because of the complexity of visualisation tasks. In addition, when solving dense methods for computer vision tasks, parallel computation is the most intuitive and efficient method since most of operations are pixel-wise and given that the computational power of GPU outperforms CPU when the serial operations can actually be parallelised and all the element-wise operations proceed at once. The design principle is that the GPUs have a parallel throughput architecture that in general emphasises executing many concurrent threads slowly, rather than executing a single thread very quickly.

Furthermore, the gap between computer vision and computer graphics is tending to reduce. Sophisticated dense visual SLAM systems such as DTAM have shown that by incorporating graphics algorithms into the computer vision tasks, the built

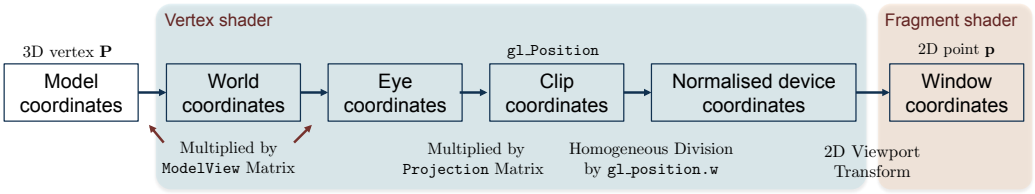


Figure 2.2: OpenGL rendering pipeline.

dense map and tracking can seamlessly improve one another. Specifically, on one hand, the dense tracking needs the textured surface to be efficiently rendered in hyper real time, and on the other hand, the multi-view stereo reconstruction also requires the camera to be well tracked with the rendered keyframe textures.

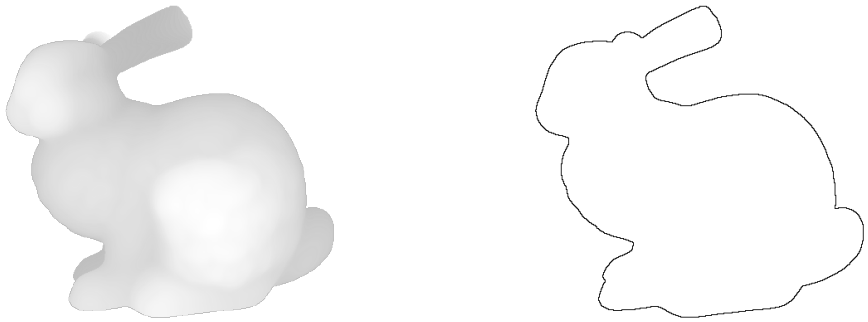
2.5.1 Open graphics library (OpenGL)

OpenGL is a cross-platform application programming interface (API) for rendering 2D and 3D vector graphics. The API is designed to access the GPU, to achieve hardware-accelerated rendering. The generic geometry primitives that OpenGL provides include points, lines, triangle meshes, etc.

OpenGL defines and implements a highly optimised rendering pipeline to efficiently render 3D geometries onto 2D windows on the screen. The pipeline is shown in Figure 2.2. 3D geometry data are input to *vertex shader* and transformed, projected and rasterised before arriving at the *fragment shader* and in the end the pixel colours are assigned in the fragment shader and will be filled into the *framebuffer*.

Note that OpenGL combines the transformation of model and camera coordinates together in OpenGL's `ModelView` matrix. In practice we will never need to literally transform object vertices into world coordinates using $\mathbf{T}_{wm}$ since this can be computationally expensive if the number of vertices is large. Instead, by incorporating the model-to-world transformation matrix $\mathbf{T}_{wm}$ into the world-to-camera transformation $\mathbf{T}_{cw}$, we can obtain a `ModelView` matrix $\mathbf{T}_{cm}$ which simply transforms the 3D points from the model coordinates to camera coordinates, bypassing the world coordinates.

Beside the default on-screen framebuffer, OpenGL also provides off-screen framebuffer rendering. The frame buffer object (FBO), vertex buffer object (VBO) and pixel buffer object (PBO) are particularly useful functions when we want to perform a render-to-texture operation and use the rendered textures to interact with computer vision tasks.



(a) (b)

Figure 2.3: GLSL rendering examples. The Stanford bunny in depth (a) and contour (b).

2.5.2 OpenGL shading language (GLSL)

OpenGL is a kind of state machine which provides high performance rendering functions for 3D graphics, in which users have to set flags for a given rendering task. The OpenGL APIs such as `glEnable()`, `glViewport()`, `glClear()`, etc, can set the rendering states but they do not allow users to control rendering approaches. This is where GLSL plays a role. GLSL is a high-level shading language based on the syntax of the C programming language. With GLSL the original functions of the vertex shader and fragment shader become programmable. This enables users to implement task-driven rendering approaches without scarifying any performance.

Figure 2.3 shows rendering examples which are not possible to achieve by using the generic OpenGL API. Although in general GLSL is used for better graphics rendering such as to realise Phong shading, programming shaders can also be very useful for computer vision tasks. For example, depth map can be used as assigning depth value to each pixel and a normal map provides a dense interpolated normal for each pixel.

2.5.3 Compute unified device architecture (CUDA)

Compared with GLSL, CUDA is a more general programming language which allows users to program GPGPU to exploit the parallel computation power of the GPU. In a modern CUDA architecture, a CUDA core consists of several functional components as shown in Figure 2.4.

The most basic computation unit is a *thread* which lies inside a *block*. A number

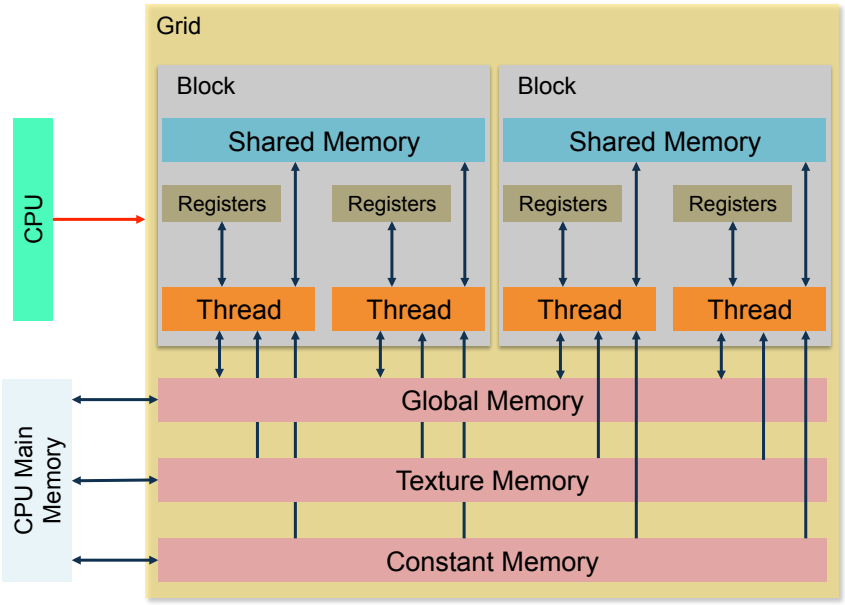


Figure 2.4: Modern CUDA architecture. The texture and constant memory on each grid is read-only. After CUDA compute capability 2.0, it is possible to write the texture by using *surface object*.

of blocks form a *grid*. In a grid there are *global memory*, *texture memory* and *constant memory* which are globally accessible (for all threads). In each block, there is a local memory called, *shared memory*, which allows access for threads in the same block.

When programming a CUDA application, the processing data has to be uploaded from CPU main memory to GPU global memory so each thread in each block has access to it. One can see that it is not possible for threads across different blocks to share processing data. When implementing algorithms where a summation operation is involved, such as the vector inner product or the solver of the normal equations, a *reduction* algorithm has to be considered (Sanders and Kandrot, 2010).

2.6 Summary

This chapter has describes basic geometric and mathematical concepts about dense methods. The practical programming tools used in this thesis have also been introduced.

3 Mapping and Tracking Using Preoperative Models

Contents

3.1 Introduction

The first approach we took to pure vision-based image-guided surgery was to exploit the preoperative CT/MRI scanned models. The idea is that preoperative models are not only of interest for the surgeons to perceive the location of underlying anatomical structures in the endoscopic images, but also that the geometric shape of anatomy from the scans can provide information for registration to the right pose. We make the assumption that the scenes are rigid so that we can perform rigid image registration to align a model with endoscopic images. In RALP, the pelvis bone should be rigid and the pubic arch is visible in the endoscopic scene, so this is potentially useful anatomy for registration.

We propose approaches to register a non-textured rigid 3D model with a number of 2D images followed by dense monocular camera tracking using the registered model in order to continuously align the preoperative models for AR image-guided surgery. This framework is similar to visual SLAM but the map is now replaced with the dense preoperative model and we just have to register the dense map in the right place. Dense visual tracking has recently been shown very robust and can achieve real-time performance using GPU computation (Newcombe et al., 2011a,b). The lack of a ground truth in 3D medical image registration has led to the suggestion of simulations for algorithm testing (Jannin et al., 2002). We have developed a ground truth simulation to validate the performance of the proposed registration algorithm. Additionally, a fast method for calculation of the visible 3D surface points for colour-consistency calculation is proposed.

The initial idea was to use PTAM, which is a real-time sparse feature-based monocular visual SLAM system (Klein and Murray, 2007) to provide an initial set of 2D images with their camera poses, followed by photo-consistency-based multi-view 3D to 2D image registration (Clarkson et al., 2001). The method turned out to work well with synthetic rigid scenes where the registration has a capture range of ± 9mm/° and a target registration error (TRE) less than 2 mm. However, in real endoscopic scenes, we found that PTAM could not provide accurate camera poses since most of the feature points were extracted from deforming regions or from specular highlights. Therefore, we abandoned PTAM and turned to stereo vision in which we can have two rigid views to register the model.

Dense tracking using the registered and textured 3D model turns out to be extremely accurate by itself, providing that the scene is rigid and the textured model agrees with the true geometry. However, the geometric shape of the preoperative

model is not exactly the same as what is observed in the real endoscopic scene, due to a thick covering of tissue over the bone and model segmentation errors. A stereo registration for a particular frame can be obtained by optimisation so that we can perspectively back-project the 2D image to texture the model. However, the dense tracking drifts rapidly in subsequent endoscopic video frames. This is because the problem is ill-conditioned due to the lack of correspondence between the model and the viewed surface. To resolve this problem, Chapter 4 and 5 provide algorithms for real-time stereo dense reconstruction and tracking for endoscopic scenes in a sound theoretical framework.

The method proposed in this chapter using 3D to 2D non-textured model image registration with PTAM has been published (Chang et al., 2011, 2012). The method for dense tracking using an automatically segmented preoperative model has been published in a co-authored paper (Gao et al., 2013).

3.2 Background

A significant issue when reconstructing a 3D organ is tissue deformation. Algorithms for deformable 3D surface reconstruction can be separated into template-based and non-rigid structure from motion reconstruction. Both approaches have shown success in deformable 3D surface reconstruction (Salzmann and Fua, 2010). However, when there are too few features that can be detected in the scene, neither class of approach performs well, which prevents them from being used in practice. Nevertheless, a number of techniques have been published which applied feature-based 3D reconstruction in endoscopic sequences. Stoyanov et al. (2004) presented a method for dense depth recovery from stereo laparoscopic images of deformable soft-issue. Mourgues et al. (2001) proposed a correlation-based stereo method for surface reconstruction and organ modelling from stereo endoscopic images.

Quartucci Forster and Tozzi (2000) applied a shape-from-shading technique to estimate the surface shape by recovering the depth information from the surface illumination. Mountney et al. (2007) proposed a probabilistic framework for selecting the most discriminative descriptors by Bayesian fusion method to compare twenty-one different descriptors. Wang et al. (2008) used scale-invariant feature transform (SIFT) features for endoscopy sequences and used the adaptive scale kernel consensus for robust motion estimation. Wu et al. (2008) also tracked SIFT features and utilised an iterative factorisation method for structure estimation. Mountney et al. (2006) presented a technique to construct a 3D map of the scene for MIS endoscopic surgery while recovering the camera motion based on SLAM from a stereo

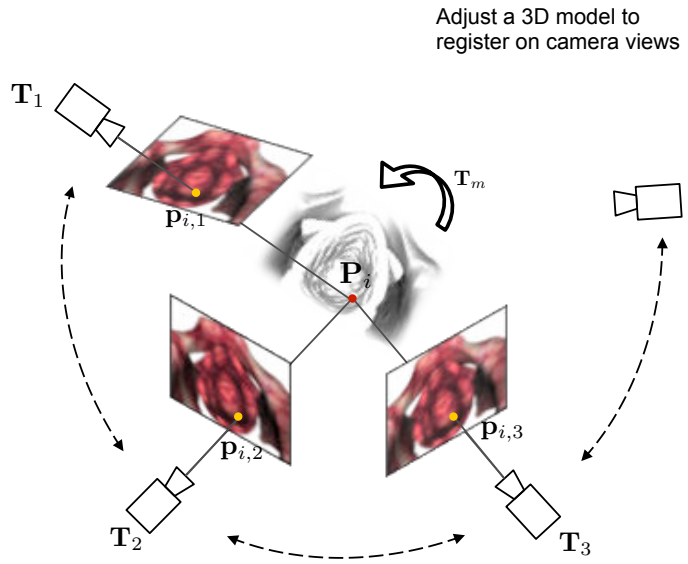


Figure 3.1: Colour-consistency with PTAM's camera tracking for 3D to 2D images registration.

endoscope, but their main focus was surface reconstruction.

It is worth noting that state-of-the-art sparse feature-based visual SLAM systems such as PTAM have achieved significant success in real-time camera tracking and mapping for real scenes with decent accuracy (Klein and Murray, 2007). Newcombe and Davison (2010) also utilised PTAM to obtain camera poses for dense surface reconstruction. However, with the features available in endoscopic images, the performance of 3D reconstruction or registration is much worse due to deformation, specularity and textureless regions.

3.3 Multi-view 3D to 2D image registration

When registering a 3D model to multiple 2D views, many robust features are tracked in the scene and the camera tracking and scene reconstruction are calculated in separate parallel threads. The main role of PTAM is to provide the camera poses for a number of video frames. The idea of incorporating colour-consistency with PTAM's camera tracking for 3D to 2D image registration is shown in Figure 3.1.

3.3.1 Colour-consistency cost function

The i-th 3D point $\mathbf{P}_i \in \mathbf{\Psi}$ in the visible point set $\mathbf{\Psi}$ of a 3D model and its corresponding 2D points $\mathbf{p}_{i,j}$ in the j-th camera projection image $\mathcal{I}_j \in \mathbf{\Omega}_j$ should have a similar colour. This property is called colour-consistency. Lighting and reflectance of

the surface are significant factors in making the colour-consistency assumption. For preliminary studies, we take the Lambertian model that considers purely ambient lighting so we have an ideal environment to examine the proposed approach.

The colour-consistency cost function is defined by projecting all of the visible points $\mathbf{P}_i$ into each camera image $\mathcal{I}_j$ using the camera pose $\mathbf{T}_j \in \mathbb{SE}(3)$ and the model pose $\mathbf{T}_m \in \mathbb{SE}(3)$:

$$E_{cc} = \sum_{j} \sum_{\mathbf{P}_i \in \mathbf{\Psi}} \left(\mathcal{I}_j(\pi(\mathbf{K}\mathbf{T}_j\mathbf{T}_m\mathbf{P}_i)) - \bar{\mathcal{I}}(\mathbf{T}_m\mathbf{P}_i) \right)^2 . \tag{3.1}$$

Taking all the camera images into account, the function $\bar{\mathcal{I}}$ averages the colour of the point $\mathbf{P}_i$ with a model pose $\mathbf{T}_m$. The projection function π projects a 3-vector homogeneous point to 2D space. We assume that the camera is calibrated in advance so the intrinsic matrix $\mathbf{K}$ is known. Note that we stack the RGB intensities to form a long image vector $\mathcal{I}$.

The term inside the squared function in Equation 3.1 is essentially the colour variance of each point $\mathbf{P}_i$. In other words, the colour-consistency is established by minimising the colour variance over all the visible points with respect to the model pose $\mathbf{T}_m$. The camera pose estimates $\mathbf{T}_j$ are provided by PTAM so they are constant during the registration.

Visibility detection

In Equation 3.1, only 3D points that are visible in at least two camera images are taken into account. For each point on the model surface, we first need to calculate whether it is visible in each keyframe (Spoerk et al., 2012). To achieve this we set a surface colour for each vertex, where the colour $[R, G, B]^\top$ has been set to the position $[x, y, z]^\top$ as shown in Figure 3.2c. By rendering the object from each keyframe position we can limit ourselves to the visible front face simply by checking that the projected colour matches the colour of the vertex. As shown in Figure 3.2d, the colour of the vertex at the front side is different from the back side. This turns out to be much more efficient than z-buffer methods which require us to calculate the distance to the vertex in each of the images.

Validation of colour-consistency assumption

To investigate whether the designed colour-consistency is optimisable, we conducted a synthetic test using textured bladder and pelvis phantom models shown in Fig-

ure 3.2a. The phantom models were obtained by CT and we textured them with real RALP endoscopic images. The synthetic textured phantom models are shown in Figure 3.2b. With the textured models, we make a simulation video where the true camera poses $\mathbf{T}_j$ are known. The synthetic video can be watched online[1].

The goal of minimising the Equation 3.1 with respect to the model pose $\mathbf{T}_m$ is to accurately align the model onto every 2D projection image. Using the simulation video and the camera poses, we devised a cost function test which intentionally moves the model pose $\mathbf{T}_m$ with respect to each motion parameter apart from the ground truth in $[-20, 20]$ mm and $[-30, 30]$ ° with 0.1 interval for translation and rotation respectively. Figure 3.3 reveals that the cost function has a clear global minimum in the investigated ranges, and with more camera frames included the cost function tends to be smoother.

Optimisation

To optimise the pose $\mathbf{T}_m$ of the 3D model, a derivative-free algorithm is adopted. Having tried classical derivative-free approaches, we found Powell's bound optimisation by quadratic approximation (BOBYQA) has better performance than the principal axis (PRAXIS) and simplex methods (Powell, 2009). In addition, BOBYQA provides bound-constrained optimisation which can restrict the search within a reasonable capture range. The optimisation is conducted for the 6-vector $\mathbb{SE}(3)$ of the rigid body transformation $\mathbf{T}_m$.

3.3.2 Empirical studies

Experiments were run on an Intel(R) Core(TM) 2 Quad 2.5 GHz CPU with 4 GB physical memory and a Nvidia GeForce GT 330 graphic card with 1 GB global memory. The PTAM system is provided by the original work of Klein and Murray (2007)[2], which is used to evaluate camera poses. The BOBYQA algorithm for registration is provided by an open source library[3]. The rest of programs are implemented in C++ and CUDA together with VTK for visualisation.

Evaluation of PTAM's camera tracking

PTAM tracks camera positions on the fly by simultaneously tracking sparse features and mapping scenes. To validate the accuracy of using PTAM's camera pose, we

[1]Synthetic video: http://youtu.be/UyLnC6De1kw
[2]PTAM: http://www.robots.ox.ac.uk/~gk/PTAM
[3]NLopt: http://ab-initio.mit.edu/wiki/index.php/NLopt_Algorithms

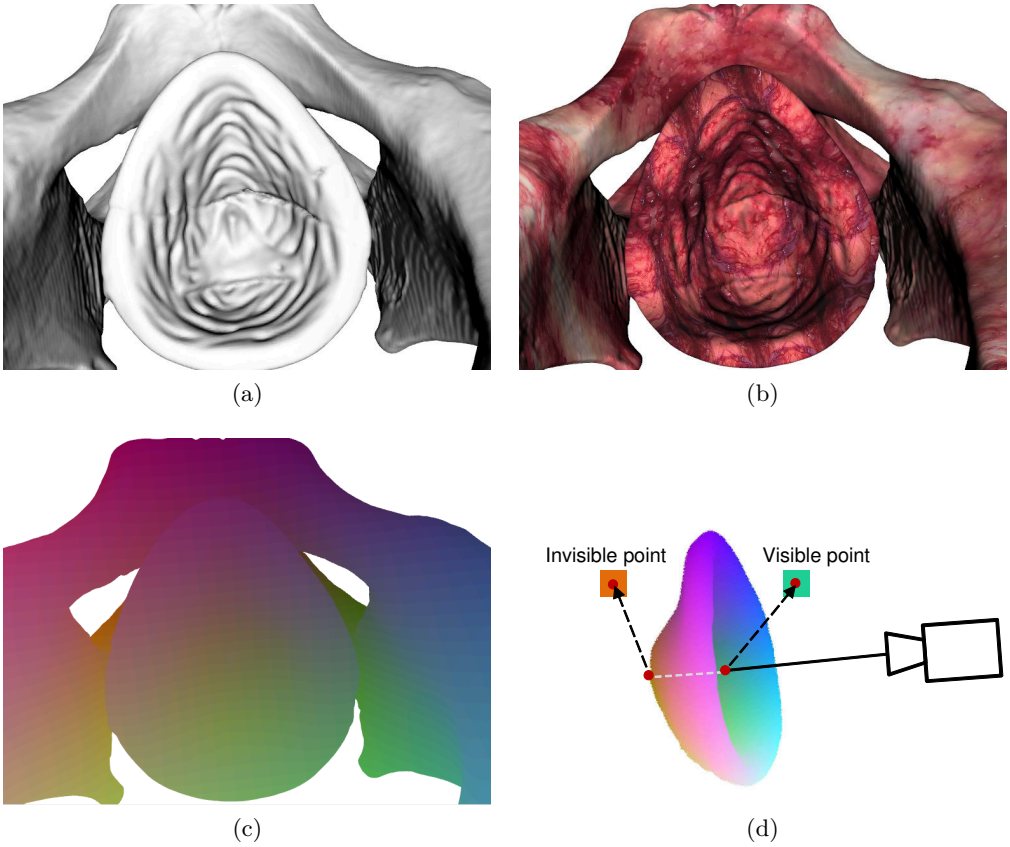


Figure 3.2: (a) The ground truth phantom models of bladder and pelvis. (b) Their appearance after being textured. (c) The colourised model for detecting visible front face vertices. (d) The proposed fast visible point detection algorithm. Both cyan (front) and orange (back) points are projected onto the same position in the camera image, and can be distinguished by the vertex colour.

ran PTAM on the simulation video to obtain the tracked camera poses as shown in Figure 3.4a and 3.4b, and then, such estimated camera poses were compared with the simulation video's camera poses. Since PTAM's coordinate system is defined by using a manual stereo initialisation, to fairly compare them, we used Procrustes analysis algorithm (Schönemann, 1966; Horn, 1987; Bookstein, 1997) to carry out coordinate system registration with respect to the rigid and scale transformation. Figure 3.4c and 3.4d show the result after PTAM's camera positions are transformed into the ground truth coordinate system. The overall RMSE is 0.5 mm. Note that although the pose errors are small, they will be propagated to the registration and break the colour-consistency assumption.

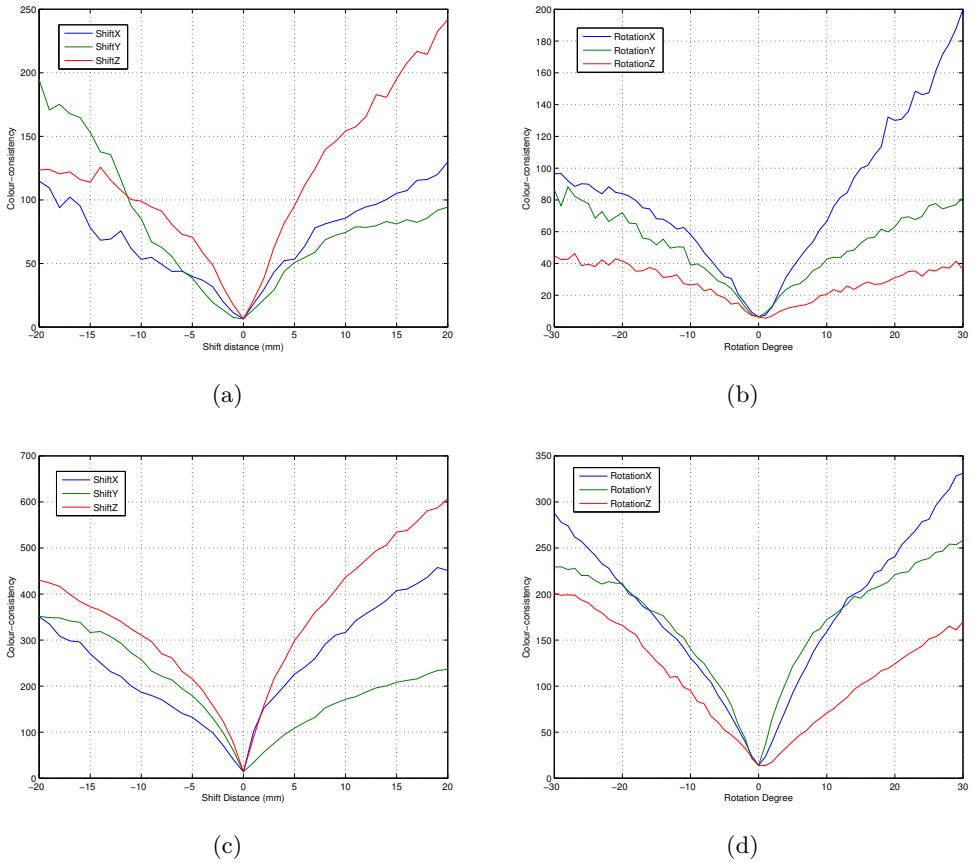


Figure 3.3: The validation results produced by rotating and shifting the 3D model with ground truth camera poses using 2 camera images in (a) and (b) and using 10 camera images in (c) and (d).

Evaluation of the model registration

We conducted an experiment which changes the ground truth pose by using additive white Gaussian noise (AWGN) with different standard deviations. Under each standard deviation, we ran the registration 500 times using a random set of 2, 5, 10 and 20 PTAM camera keyframes. A three-layer pyramid suggested by (Maes et al., 1999) was used for the derivative-free optimiser (see Section 5.5.6). If the target registration error (TRE), which is defined as RMSE of the entire vertices, is less than 2 mm, we regarded the registration process as having converged.

Figure 3.5a and 3.5b show the results of the frequency of convergence when the projection point colours are calculated by nearest neighbour and by bilinear interpolation respectively. The performances do not show much difference, and curves in each case share common trends. One can see that when only 2 keyframes are used,

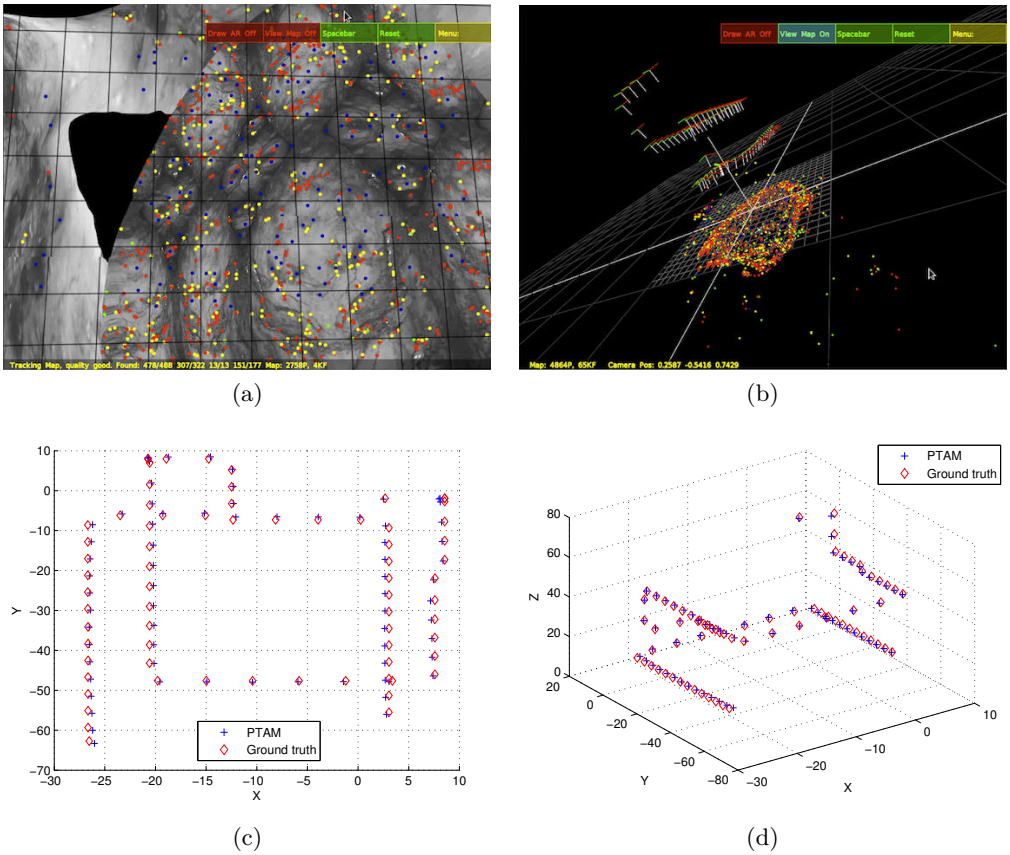
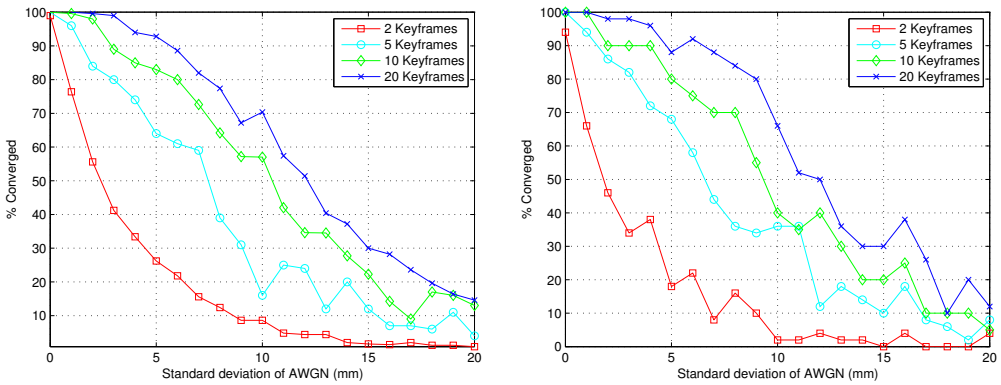


Figure 3.4: (a) Using PTAM for camera tracking on the simulation video. (b) PTAM's camera keyframes and the extracted features in a 3D view. (c) The camera poses of PTAM's keyframes compared with the ground truth. (d) The 3D view of (c). The overall pose RMSE is 0.5 mm.

the frequency of convergence drastically decreases. When using 20 keyframes, the frequency of convergence starts to drop down after 3 mm standard deviation which is about the range in ±9 mm/°. Involving more keyframes may result in much better performance, but it also introduces more computation effort. Table 3.1 shows the statistics of the required optimising iteration and the running time for the converged cases. Note that the iteration time is accumulated by the number of iterations the optimiser takes in each layer. The average numbers of iteration are almost consistent no matter how many keyframes are used. This is because the required iteration of a derivative-free optimiser is largely affected by the number of parameters, and here we have 6 parameters in all cases.

Table 3.1: Average running time and required iteration time for optimization with only the model pose

Keyframe #	Avg. iteration	Avg. running time per iter. (sec.)	Avg. running time (sec.)
2	130	0.05	6.75
5	149	0.12	14.32
10	151	0.20	30.48
20	154	0.40	62.06



(a) (b)

Figure 3.5: Optimisation convergence frequency of the proposed approach using nearest neighbour interpolation (a) and using bilinear interpolation (b) for the projection point colour.

Registration in real endoscopic sequences

We manually align the 3D model with the camera images to initialise the model pose. Although in the simulation scene PTAM has shown an accurate tracking result (Figure 3.4), it turns out that PTAM cannot provide accurate camera poses when the scenes are more challenging, where there are many specular highlights or tissue deformation occurs. In addition, the invasion of surgical instruments will make the tracking fail. Therefore, we conclude that PTAM is not robust enough to provide multi-view camera poses in the endoscopic scenes.

Given that the da Vinci robotic platform equipped with a stereo endoscopic camera, we can potentially use only two views to register the model. The stereo vision can provide an instant static snapshot for registration. As the convergence frequency may be poor when using only two images (Figure 3.5), we will have to investigate the final registration result and re-run the registration with different initial model poses. This, however, is not practical for an augmented reality system

for clinical use since the response time can be extremely critical during surgery.

The registration of a preoperative model with real clinical endoscopic scenes was eventually unsuccessful. To approach augmented reality, this methodology is implausible primarily due to the lack of geometry consistency and the difficulty of performing robust camera tracking to provide accurate camera poses in advance in endoscopic scenes. The computation time as shown in Table 3.1 is also far from real-time performance. Nevertheless, this does not prevent us from using the registered model to perform dense camera tracking.

3.4 Dense monocular camera tracking

We have shown that a non-textured 3D model can be registered with a number of 2D camera images. Once the model is properly registered, it can be textured by *projective texturing* by using OpenGL. Given a textured 3D model, a real-time dense 2.5D camera tracking can then be performed. Note that this also makes the assumption of a rigid-scene, which is a common limitation of general visual SLAM systems such as PTAM. But with a considerable number of measurements being present, such a dense tracking methodology turns out to be much more robust than a sparse feature based tracking system, even without compromising with real-time performance thanks to the availability of novel GPGPU architectures (Newcombe et al., 2011b).

3.4.1 Photometric cost function

The dense camera tracking can be achieved by projecting the textured model and registering the projection image with the current video image. It follows that the textured 3D model is projected into a virtual image $\mathcal{I}_v$ and the current image $\mathcal{I}_c$ is warped with respect to the camera pose $\mathbf{T}_{cv}$ to register with the virtual image. The photometric cost function is thus defined as

$$E_{dt} = \sum_{\mathbf{P} \in \mathbf{\Psi}_v} \left(\mathcal{I}_c(\pi(\mathbf{K}\mathbf{T}_{cv}\mathbf{P})) - \mathcal{I}_v \right)^2 . \tag{3.2}$$

Once the photometric cost function is minimised, the optimal camera pose $\mathring{\mathbf{T}}_{cv}$ can be found that best registers the live image with the virtual image. We use the Lucas-Kanade style 2.5D forward compositional algorithm (FCA) to minimise Equation 3.2 with respect to the $\mathbf{T}_{cv}$ using its minimal parameterisation $\mathfrak{se}(3)$ 6-

vector (see Section 5.5 for the algorithm details).

The visible point set $\mathbf{\Psi}_v$ will vary with the camera motion. Therefore we will have to update the virtual image at a certain stage to ensure there are enough visible points to form the virtual image. In practice the 3D points do not need to be from the model vertices. Instead, we exploit the OpenGL rendering pipeline to get a dense 3D point set on the interpolated smooth surface.

3.4.2 Empirical studies

Experiments were run on an Intel(R) Core(TM) 2 Quad 2.5 GHz CPU with 4GB physical memory and a Nvidia GeForce GT 330 graphics card with 1 GB global memory. The dense tracking algorithm (FCA) is implemented in CUDA, which is able to run at 100ms per 720×576 frame.

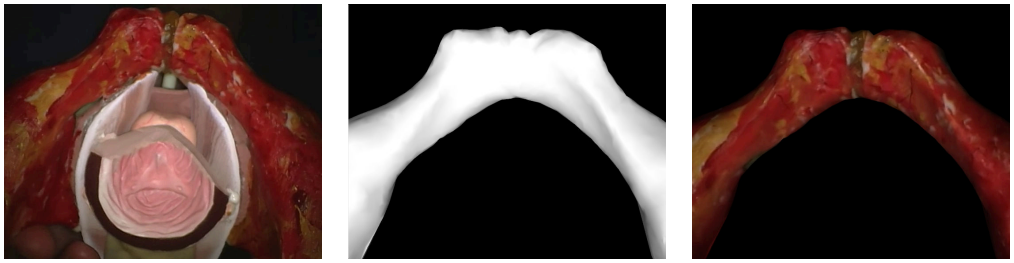
Synthetic ground truth study

In the RALP, the arch of the pubic bone is the largest rigid structure that can be seen in the endoscopic images. The pubic bone model is manually segmented from the preoperative CT scan and registered using the algorithm proposed in the previous section. To investigate the accuracy of dense tracking, we again conducted a synthetic tracking test where we simulate a video sequence using a textured pubic bone segmented from a textured phantom model shown in Figure 3.6.

In order to obtain a realistic camera trajectory, we first tracked the real phantom video using PTAM. This trajectory was then taken as the ground truth to generate the synthetic camera trajectory for the textured model as shown in Figure 3.6c to generate a simulation video. Dense tracking was then performed on the simulation video and the result was compared with the ground truth, as shown in Figure 3.7. The RMSE is 0.33 mm overall. One can also qualitatively judge that the tracking path is very close to the ground truth. This shows that with ideal conditions, the dense tracking can be extremely accurate.

Real sequence study

We manually align the pubic bone model with the camera images to initialise the model pose and perform registration using the method proposed in the previous section. Figure 3.8a shows a phantom alignment example, where the camera tracking is done by aligning the pubic bone frame-by-frame. The tracking result turns out to be quite robust with this rigid phantom. On the other hand, Figure 3.8b shows the result in real endoscopic scene. The tracking can still work but the visibility of the



(a) (b) (c)

Figure 3.6: The phantom model used for the synthetic tracking test. The real left camera scene (a); its CT scan model after manual segmentation (b); the model after being projectively textured by using the image (a).

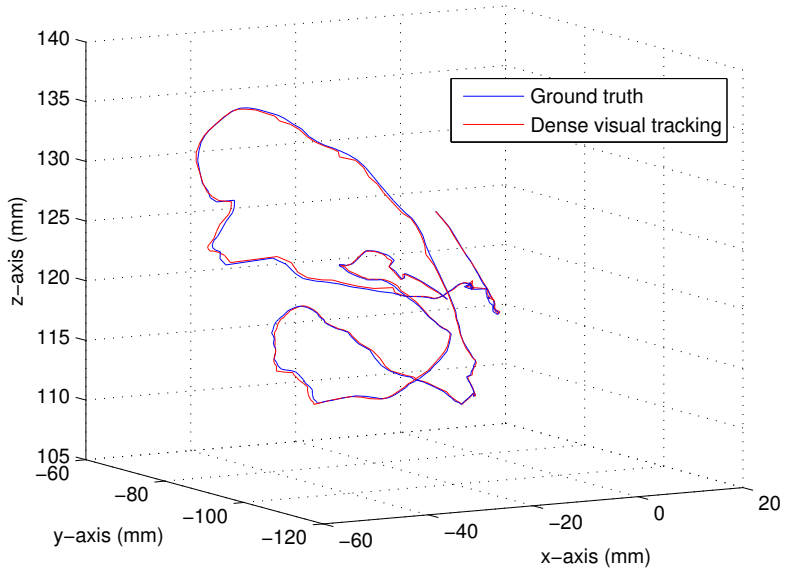
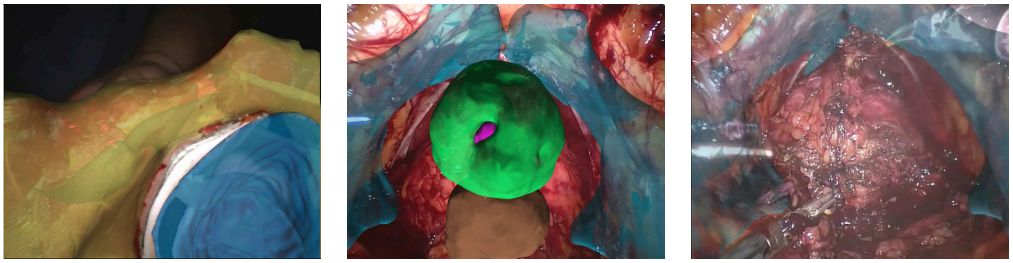


Figure 3.7: Synthetic dense tracking test showing a RMSE 0.33 mm.

pubic bone is limited. Sometimes it is even barely seen. In addition, the pubic bone model has an obviously different geometric shape compared to the real endoscopic appearance due to the presence of layers of soft tissue over the bone. These issues make the tracking very unstable and the camera will drift quickly. Figure 3.8c also shows a failed tracking case where there are motions from surgical instruments. Since the model is textured by the camera images, if the scene is not static, the texture will not be consistent between the model and the scene in different frames and dense tracking consequently fails. For better understanding of the AR demos,



(a) (b) (c)

Figure 3.8: (a) The phantom video with the augmented bladder and pelvis models; (b) The real video of RALP with the augmented models of pelvic rim, prostate, urethra and rectum. (c) A failed tracking example where the tracked pubic bone model moves with the instrument motion.

please watch the online video[4].

3.5 Discussions and conclusions

In this chapter, we have presented an approach to registration of a preoperative 3D model to intraoperative endoscopic video which combines PTAM tracking with colour-consistency registration incorporating a fast calculation of the visible 3D surface, followed by a preliminary attempt for dense monocular camera tracking. To validate the method we developed simulation test datasets with ground truth camera trajectories and textured models. Although the monocular dense tracking showed promisingly robust and real-time performance, the proposed registration scheme is however problematic in real clinical cases in terms of both practice and theory.

Registration between two different geometries

The 3D to 2D registration using colour-consistency is an over-optimistic idea. In practice, a segmented preoperative model does not necessarily have a geometric shape exactly the same as what can be seen in the endoscopy. This can be due to errors sourced from the CT/MRI scan and the segmentation algorithm. The errors will be accumulated and propagated to the final 3D model. If the surface geometries are not the same, the multi-view colour-consistency assumption between the 3D model and the 2D images will not be hold.

A promising algorithm for relaxing the geometry constraint for a 3D to 2D non-textured model registration is to use a soft probabilistic method in which the 3D

[4] http://youtu.be/fPWp8XiurCI

model plays a role to provide foreground/background masks to define the statistical appearance models, and the registration is performed by pixel-wise MAP estimation (Prisacariu and Reid, 2012). But this scheme requires the foreground and background statistical models to have distinct probability distributions, which may not be the case for our surgical images. This is something we will investigate in the near future.

The impracticality of registering rigid models using PTAM for real-time augmented reality

We have also shown that PTAM can provide accurate camera tracking only within ideal scenes where there is no deformation and there are enough good features to track. Such ideal scenes are practically not available for real endoscopic scenes. Furthermore, the instrument motions make PTAM easily fail. As a result, the sparse feature version of visual SLAM systems are not suitable for use in RALP endoscopic scenes.

Even if we have a number of decently tracked camera poses, the registration process is still far from real-time performance as shown in Table 3.1. Although the cost function in Equation 3.1 can instead be locally linearised and be optimised by efficient Lucas-Kanade-style algorithms which will be introduced in Chapter 5, the unestablished colour-consistency assumption will anyway make the registration failed in the end. In short, we had just ill-conditioned the problem.

Redirection

Instead of registering the preoperative models with the camera images, we can directly reconstruct the 3D structure of the scene using a stereo endoscopic camera. In this way, the stereo reconstruction provides an instant 3D textured model that is useful for further dense tracking. Given that the-state-of-the-art stereo reconstruction algorithms have demonstrated high quality and real-time performance, this is a promising approach to achieve an efficient and effective tracking system for image-guided endoscopic scenes. We therefore redirected our approach to the problem. Chapter 4 will introduce a real-time dense stereo reconstruction algorithm. The reconstructed model is then used for stereo camera tracking as described in Chapter 5.

4 Real-Time Dense Stereo Reconstruction

Contents

4.1 Introduction

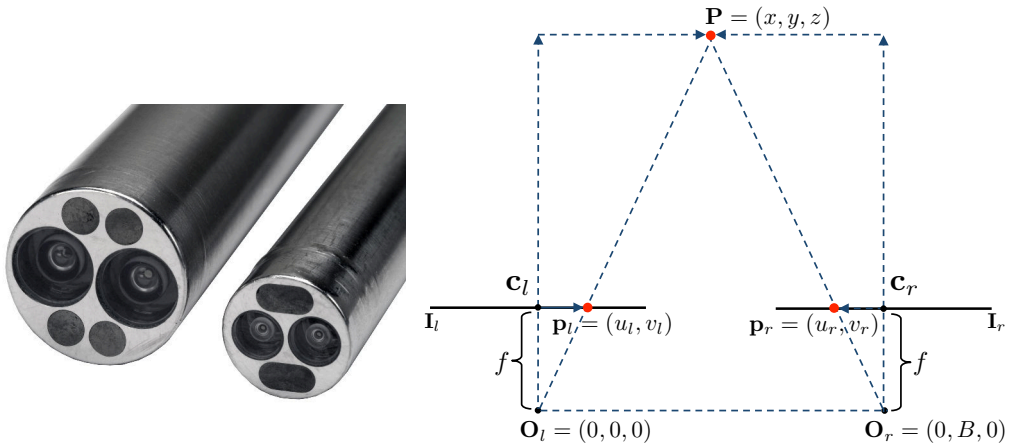
We have seen that the colour-consistency-based registration between a 3D preoperative model and 2D surgical endoscopic images is unstable if the surface geometry of the preoperative model and the surgical scenes are very different. This situation is however depressingly common due to the fact that the segmentation from a CT/MRI scan model will never be perfect. The inaccurate registration will make the subsequent dense camera tracking prone to drift. Instead, we can estimate the 3D structure of the observed surgical scene directly. The recovered 3D scene can provide a rich source of information for visualisation and interaction, enabling dense camera tracking and registration to a preoperative model for surgical navigation (Mirota et al., 2011; Mountney et al., 2010).

Given that the surgical site is equipped with stereoscopic laparoscope as shown in Figure 4.1a, a straightforward approach for estimating the 3D structure is by dense stereo reconstruction. This chapter elaborates the dense stereo reconstruction problem and describes the proposed reconstruction algorithm which has yielded the best result in a number of publicly available ground truth datasets. The results have also been published in MICCAI 2013 (Chang et al., 2013) followed by a co-author work published in IEEE Transaction on Medical Image 2014 for further comprehensive evaluation (Maier-Hein et al., 2014).

The reason that the proposed method stands out is mainly due to the use of patch-based data term (ZNCC). Compared with the conventional pixel-to-pixel data term, patch-based similarity measure is able to correlate larger regions, and this is particularly important for stereo matching in texture-poor areas. A more sophisticated regulariser term (Huber norm) is also introduced to tackle with the stair-casing effect caused by the conventional L^1 norm. Details of the underlying mathematical theories over different approaches are described in Section 4.3.

4.2 Problem statement

From two views of a textured scene in which the camera poses are known, dense structure can be estimated by a process known as *dense stereo reconstruction*. In general we would like to make the two images horizontally line up by a process termed *stereo rectification*, so that the correspondence search on the epipolar line can be efficiently performed in only one dimension. The horizontal displacement between the pixel and its correspondence is named *disparity*. As shown in Figure 4.1b shows that, to estimate the depth z of each pixel, we can simply apply a triangulation



(a) (b)

Figure 4.1: (a) Stereo endoscopes (©2014 Intuitive Surgical, Inc). (b) Rectified stereo geometry illustrated by top view.

using the disparity d:

$$u_l = \frac{xf}{z} \quad \text{and} \quad u_r = \frac{(x-B)f}{z} \quad \Rightarrow \quad z = \frac{fB}{u_l - u_r} = \frac{fB}{d}, \tag{4.1}$$

where $d > 0$. Providing that the stereo camera is usually calibrated in advance and hence the principal points $\mathbf{c}_l$ and $\mathbf{c}_r$, the focal length f and baseline B are known, the reconstruction process is mainly to estimate the disparity d for each pixel to get the disparity map.

Consider an image function $\mathcal{I}(\mathbf{p}) : \mathbf{\Omega}_\mathcal{I} \to \mathbb{R}$ where $\mathbf{p} = (u, v)$ is the pixel location in the domain $\mathbf{\Omega}_\mathcal{I} \subseteq \mathbb{R}^2$. Given a rectified stereo rig, a point $\mathbf{p}_l$ in the left image has its correspondence $\mathbf{p}_r = (u - \mathcal{D}(\mathbf{p}_l), v)$ in the right, or inversely, a point $\mathbf{p}_r$ in the right has its correspondence $\mathbf{p}_l = (u + \mathcal{D}(\mathbf{p}_r), v)$ in the left, depending on which image is regarded as the reference frame. The goal of the dense stereo reconstruction is to estimate the disparity function $\mathcal{D}(\mathbf{p}) : \mathbf{\Omega}_\mathcal{I} \to \mathbb{D}$ where $\mathbb{D}$ can be a discrete or continuous set and is in general positive, i.e., $\mathbb{D} \subseteq \mathbb{R}_+$. Such a problem is also termed *stereo matching* or *stereo correspondence*.

4.3 Background

The stereo correspondence problem has been extensively studied and is still an active research topic in the computer vision community. The first taxonomy and evaluation

work was done by Scharstein and Szeliski (2002). In their work, the solution space of disparity map is termed disparity space image (DSI) which is basically a function that maps a 3-vector to a cost value:

$$\mathcal{C}(\mathbf{p}, \mathcal{D}(\mathbf{p})) = \rho(\mathcal{I}_l, \mathcal{I}_r, \mathbf{p}, \mathcal{D}(\mathbf{p})), \tag{4.2}$$

where $\mathcal{C} : \mathbf{\Omega}_\mathcal{C} \to \mathbb{R}$ and $\mathbf{\Omega}_\mathcal{C} \subseteq \mathbb{R}^3$. The function ρ can be an arbitrary similarity measurement. Without loss of generality, instead of using the name DSI, we would rather call the function $\mathcal{C}$ a *cost volume* throughout this thesis. This is in consideration of the context of multi-view stereo reconstruction where the stereo pairs need not to be rectified, so the cost volume in this case is constructed by inverse depth instead of the disparity. Figure 4.2b shows a discrete cost volume with a size $|\mathbf{\Omega}_\mathcal{I}| \times |\mathbb{D}|$. Scharstein and Szeliski also summarised that a stereo correspondence algorithm in general conducts four steps:

i Matching cost computation;
ii Cost aggregation;
iii Disparity computation/optimisation; and
iv Disparity refinement.

Which steps to adopt depends on a specific algorithm design. Past works have been seeking for a discontinuity-preserving, occlusion-handling, radiometric-difference-tolerant, slanted-surface-aware and sub-pixel-accurate algorithm which can result in precise stereo matching.

Approaches designed with focus on the step **i** and **ii** are categorised into *local* algorithms. In this case the correspondence computation considers only either pixel-to-pixel or patch-to-patch photometric similarity to construct the cost volume $\mathcal{C}$, following by a simple *winner-take-all* (WTA) scheme for the step **iii**, which simply extracts the disparity that has the minimum cost per pixel in the cost volume. In contrast, *global* algorithms spend most of the effort directly in step **iii** to optimise the cost function C with a smooth prior driven by the step **i** or **ii**. The step **iv** usually includes a post-processing which can be optionally conducted for those methods working in the discrete space in order to achieve sub-pixel accuracy. In addition, a left-right consistency check can also be performed in this final stage, in which the disparity maps based on the left and the right image are compared with each other, so the inconsistent regions caused by self-occlusion and illuminance difference can be handled accordingly.

The computation of the local cost matching and aggregation is usually fast in

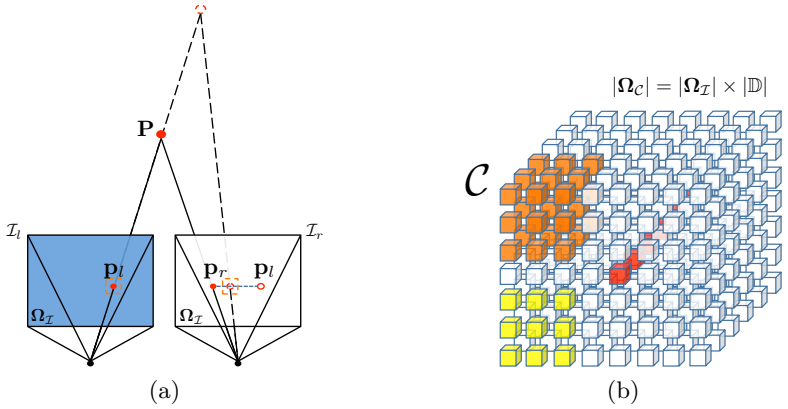


Figure 4.2: (a) To construct a cost volume, the matching cost computation calculates the similarity between patches around a pixel $\mathbf{p}_l$ in the left and a pixel $\mathbf{p}_r$ in the right. (b) A cost volume is constructed by a reference frame with resolution $|\mathbf{\Omega}_\mathcal{I}|$ using a disparity range $|\mathbb{D}|$. The red cells illustrates the search space of the WTA scheme. The yellow and orange cells show respectively a 2D and 3D convolution kernel for cost aggregation.

which a pixel-wise method has the complexity $\mathrm{O}(N{\cdot}|\mathbb{D}|)$ and a patch-wise $\mathrm{O}(s^2N{\cdot}|\mathbb{D}|)$ for the number of pixels N. On the other hand, global approaches are generally very slow since the number of variables to solve is enormously large. For example, the number of possible solutions of a 640×480 image with each pixel having 32 possible discrete disparities is 32^{307200}. Search for the global optimum in this space is actually a NP-hard problem. One may instead seek for an approximate solution using an efficient and effective optimisation scheme.

4.3.1 Matching cost computation

With a rectified stereo pair, the matching cost computation can be performed along the horizontal epipolar line as shown in Figure 4.2a. This initial process for constructing the cost volume is crucial, and subsequent steps all rely on its result. If the cost volume is built by using a vulnerable matching assumption, the reconstruction is doomed to fail. Therefore adopting a robust matching measurement is very important.

Pixel-to-pixel matching

The simplest matching cost computation can be pixel-to-pixel photometric difference by using absolute difference (AD, Kanade et al., 1995) or squared difference (SD, Matthies et al., 1989). The corresponding cost function ρ taking two images $\mathcal{I}_l$ and

$\mathcal{I}_r$, a disparity d and a pixel $\mathbf{p}_l$ is:

$$\rho_{\text{AD}} = |\mathcal{I}_l(\mathbf{p}_l) - \mathcal{I}_r(\mathbf{p}_r)| \quad \text{and} \tag{4.3}$$

$$\rho_{\text{SD}} = (\mathcal{I}_l(\mathbf{p}_l) - \mathcal{I}_r(\mathbf{p}_r))^2. \tag{4.4}$$

To improve matching under different luminance conditions, *truncated absolute difference* (TAD) is proposed:

$$\begin{aligned}\rho_{\text{TAD}} =& (1-\alpha) \cdot \min\left(|\mathcal{I}_l(\mathbf{p}_l) - \mathcal{I}_r(\mathbf{p}_r)|, \tau_1\right) + \\ & \alpha \cdot \min(|\nabla_u \mathcal{I}_l(\mathbf{p}_l) - \nabla_u \mathcal{I}_r(\mathbf{p}_r)| + |\nabla_v \mathcal{I}_l(\mathbf{p}_l) - \nabla_v \mathcal{I}_r(\mathbf{p}_r)|, \tau_2),\end{aligned} \tag{4.5}$$

where ∇_u and ∇_v are the gradient operators with respect to u and v dimension. The α controls the weights between the colour and gradient terms. The τ_1 and τ_2 are truncation thresholds. The truncated absolute difference has been shown robust to illumination changes and has become popular in stereo correspondence and optical flow studies (Slesareva et al., 2005; Steinbrücker et al., 2009; Brox and Malik, 2011; Rhemann et al., 2011). Pixel-to-pixel photometrics can be computed very fast but however pixels in homogeneous colour regions will be easily mismatched.

Correlation-based matching

One may adopt the sum of absolute differences (SAD) or sum of squared differences (SSD) as to perform *block matching* to account the correlation between left and right image patches. Specifically, the correlation is measured by computing the intensities difference in a patch $\mathbf{w}_{\mathbf{p}_l}$ at $\mathbf{p}_l$:

$$\rho_{\text{SAD}} = \sum_{\mathbf{p}_l \in \mathbf{w}_{\mathbf{p}_l}} |\mathcal{I}_l(\mathbf{p}_l) - \mathcal{I}_r(\mathbf{p}_r)| \quad \text{and} \tag{4.6}$$

$$\rho_{\text{SSD}} = \sum_{\mathbf{p}_l \in \mathbf{w}_{\mathbf{p}_l}} (\mathcal{I}_l(\mathbf{p}_l) - \mathcal{I}_r(\mathbf{p}_r))^2. \tag{4.7}$$

The summation can be computed efficiently by using *integral image* (Crow, 1984; Veksler, 2003).

However, simply taking the intensity differences and summing them up is vulnerable to different environment luminance and camera gain. Inspired by the Pearson correlation coefficient in statistics, which is for finding the linear correlation of two

random variables, we can match two image patches by using zero-mean normalised cross-correlation (ZNCC):

$$\rho_{\text{ZNCC}} = 1 - \langle\, \hat{\mathbf{n}}_l, \hat{\mathbf{n}}_r \,\rangle\,; \tag{4.8}$$

$$\hat{\mathbf{n}}_l = \frac{\mathcal{I}_l(\mathbf{p}_l) - \overline{\mathcal{I}}\,(\mathbf{w}_{\mathbf{p}_l})}{\sqrt{\sum_{\mathbf{p}_l \in \mathbf{w}_{\mathbf{p}_l}} \left(\mathcal{I}(\mathbf{p}_l) - \overline{\mathcal{I}}\,(\mathbf{w}_{\mathbf{p}_l})\right)^2}}, \quad \forall \mathbf{p}_l \in \mathbf{w}_{\mathbf{p}_l};$$

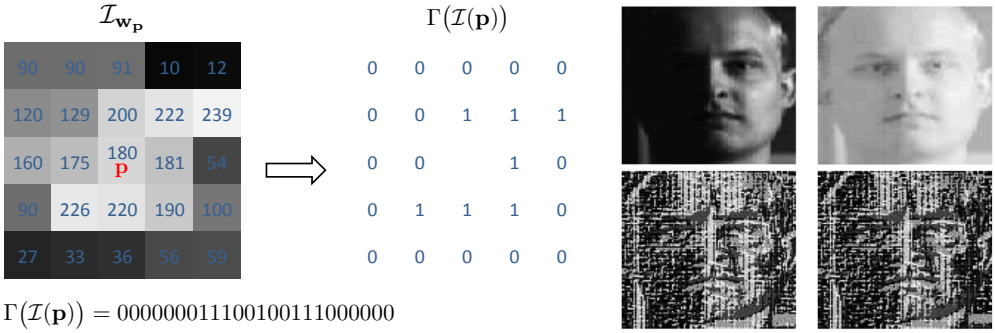
$$\hat{\mathbf{n}}_r = \frac{\mathcal{I}_r(\mathbf{p}_r) - \overline{\mathcal{I}}\,(\mathbf{w}_{\mathbf{p}_r})}{\sqrt{\sum_{\mathbf{p}_r \in \mathbf{w}_{\mathbf{p}_r}} \left(\mathcal{I}(\mathbf{p}_r) - \overline{\mathcal{I}}\,(\mathbf{w}_{\mathbf{p}_r})\right)^2}}, \quad \forall \mathbf{p}_r \in \mathbf{w}_{\mathbf{p}_r},$$

where $\hat{\mathbf{n}}$ is a normalised vector consists of centralised patch intensities and the function $\overline{\mathcal{I}}$ calculates the intensity mean of that patch (Hannah, 1974; Lewis, 1995; Heo et al., 2011). Equation 4.8 can be interpreted as the dissimilarity between two normalised vectors. While ZNCC is resilient to linear transformation on intensities, it has poor performance around the object edges. This is because near such an object boundary, the pixels have at least two distinct intensity populations within the window, in which some of the pixels belong to the object and some to other parts of the scene. Therefore, using a single correlation hypothesis cannot represent the data well.

The aforementioned traditional methods and the correlation-based methods in fact implicitly use parametric distribution with single mean and variance. Such a 1D unimodal distribution does not behave well if the intensity distribution within a patch is actually multimodal. This problem is referred as *factionalism* by Zabih and Woodfill (1994). If the disparity matching for one pixel results in several equivalent cost peaks, the subsequent disparity optimisation (i.e. WTA, dynamic programming, graph cuts, variational approach, etc.) will fail unless it introduces some regularisation (Chambolle and Pock, 2011; Geman and Geman, 1984; Kolmogorov and Zabih, 2001).

To address the factionalism while dealing with luminance change, Zabih and Woodfill (1994) proposed a local non-parametric method named *census transform* which calculates the matching cost as:

$$\rho_{\text{CT}} = \left|\Gamma\left(\mathcal{I}_l, \mathbf{p}_l\right), \Gamma\left(\mathcal{I}_r, \mathbf{p}_r\right)\right|_{\text{H}}, \tag{4.9}$$



(a) (b)

Figure 4.3: (a) A 5×5 gray patch around pixel (u, v) transformed by census transform. (b) A example shows two face images (upper row) under extreme luminance change but their census transformed images (lower row) are still quite similar. This figure is reproduced from the paper (Froba and Ernst, 2004).

where

$$\Gamma(\mathcal{I}, \mathbf{p}) = \bigotimes_{\substack{\mathbf{p}' \in \mathbf{w_p} \\ \mathbf{p}' \neq \mathbf{p}}} \gamma(\mathcal{I}(\mathbf{p}), \mathcal{I}(\mathbf{p}')) \quad \text{and} \quad \gamma(a, b) := \begin{cases} 0 & \text{if } a \leq b \\ 1 & \text{if } a > b \end{cases}.$$

The concatenation operator $\bigotimes$ simply concatenates the output of γ so it is obvious that Γ will end up with a $(|\mathbf{w_p}| - 1)$-bit string. For example, a 5×5 patch will be transformed to a 24-bit string. Figure 4.3 shows the process of census transform. The cost ρ_{CT} is then computed as the $|\cdot, \cdot|_{\text{H}}$ Hamming distance between the two bit strings.

Census transform has become popular for correspondence problems because of its simplicity and non-parametric property (Froba and Ernst, 2004; Hirschmüller, 2011; Humenberger et al., 2010; Ranftl et al., 2012; Spangenberg et al., 2013). It also turns out to be the most robust measurement under different radiometric conditions according to a study by Hirschmüller and Scharstein (2009). Recently, Hafner et al. (2013) conducted a study to explain, from a mathematical point of view, why the census transform is the most robust measurement for correspondence problems. However the computation bottleneck is that the complexity grows with increasing size of the patch. *Ad-hoc* hardware implementations can speed up the process by many orders of magnitude (Humenberger et al., 2009; Pantilie and Nedevschi, 2012; Weber et al., 2009).

Slanted surface matching

Block matching approaches assume that the scene consists of only fronto-parallel surfaces so that the pixels within the support window have a constant disparity value. This is not true for slant surfaces that can actually have different disparities within the window. Ogale and Aloimonos (2005) have taken this fact into account and proposed a correspondence algorithm dealing with horizontal and vertical slants by creating a first-order (affine) piecewise continuous functions. In addition, the state of the art best performance correspondence algorithms deal slanted shape by fitting a 3D plane at each pixel as the projected support region for computing the matching cost (Bleyer et al., 2011; Heise et al., 2013).

4.3.2 Cost aggregation

Local approaches using the pixel-to-pixel similarity (i.e., AD or TAD) usually require a further process to aggregate neighbourhood costs in order to suppress noise and amplify the local properties. In general the aggregation is performed by convolving a specific kernel over the cost volume. Due to the fact that the aggregation entirely relies on the cost matching result, it inherits the local nature of traditional window-based algorithms in which it is vulnerable to the textureless regions. This is why a larger aggregation region is usually prefered and an efficient algorithm for large window filtering is demanded.

Typical filtering

The simplest cost aggregation can be done by applying a *box filter*:

$$\mathcal{C}_{\mathrm{BF}}(\mathbf{p}, d) = \frac{1}{|\mathbf{w_p}|} \sum_{\mathbf{p}' \in \mathbf{w_p}} \mathcal{C}(\mathbf{p}', d), \tag{4.10}$$

to each slice d in the cost volume. We can also extend the 2D filter to 3D as shown in Figure 4.2b, which favours suppression of noise respectively in fronto-parallel (2D) and slanted (3D) surfaces. Another typical filter for the cost aggregation is a *Gaussian filter*:

$$\mathcal{C}_{\mathrm{GF}}(\mathbf{p}, d) = \frac{1}{2\pi\sigma^2} \left[\sum_{\mathbf{p}' \in \mathbf{w_p}} \exp\left(-\frac{|\mathbf{p}' - \mathbf{p}|^2}{2\sigma^2}\right) \cdot \mathcal{C}(\mathbf{p}', d) \right], \tag{4.11}$$

where σ is the standard deviation of the Gaussian kernel. According to the *central limit theorem*, the Gaussian kernel can also be achieved by convolving the box kernel onto itself several times.

These kind of symmetric kernels can be efficiently implemented by separately convolving a 1D horizontal kernel following by another 1D vertical kernel, or vice versa. Despite the efficiency, the consequence of applying such low-pass filtering is to smooth everything including the costs around object boundaries, where the discontinuity of disparity usually happens. To solve this problem, different attempts include *ad-hoc* windows anchored at different points such as shiftable window schemes (Bobick and Intille, 1999) and adaptive window schemes (Boykov et al., 1998; Kanade and Okutomi, 1994; Kang et al., 2001; Yoon and Kweon, 2005, 2006).

Edge-preserving filtering

Unlike the typical filtering methods, to preserve the fiducial of matching costs around object boundaries during the aggregation process, the convolution kernel can instead be specially designed, such as a *joint bilateral filter* (JBF):

$$\mathcal{C}_{\text{JBF}}(\mathbf{p}, d) = \frac{1}{\mathrm{K}_{\mathbf{w_p}}} \left[\sum_{\mathbf{p}' \in \mathbf{w_p}} \exp(-\frac{|\mathbf{p}' - \mathbf{p}|^2}{\sigma_s^2}) \exp(-\frac{|\mathcal{I}(\mathbf{p}') - \mathcal{I}(\mathbf{p})|^2}{\sigma_r^2}) \cdot \mathcal{C}(\mathbf{p}', d) \right], \tag{4.12}$$

where $\mathrm{K}_{\mathbf{w_p}}$ is a normalisation term which has to be calculated for every window $\mathbf{w_p}$. The parameters σ_s and σ_r adjust the spatial similarity and the range (intensity/-colour) similarity respectively (Petschnigg et al., 2004). Equation 4.12 approaches a Gaussian filter when pixels colours are homogeneous in the window $\mathbf{w_p}$ and tends to a suppressed Gaussian filter along the edge when the window $\mathbf{w_p}$ is across an object boundary. In addition, the JBF degrades to the original bilateral filter when the filtered image and the reference image are identical (Tomasi and Manduchi, 1998).

The JBF is expensive to compute because its kernel is non-linear. An intuitive implementation of it has a complexity $O(s^2 N \cdot |\mathbb{D}|)$. The computation is prohibitive when the window size s is large. Efforts have been focused on speeding up the kernel computation to reach $O(N \cdot |\mathbb{D}|)$ (i.e., $O(1)$ on s) using *integral histogram* but this sacrifices the filtering quality (Porikli, 2008; Ju and Kang, 2009; Yang et al., 2009). The state of the art *guided filter* exploits a guidance image to not only have a $O(N \cdot |\mathbb{D}|)$ implementation but also result in an even better filtering quality (He et al., 2010). The idea is basically to assume that the filtered image

is a linear transformation of the guidance image in a support window $\mathbf{w_p}$. Such a linear assumption implicitly ensures that the filtered image has an edge only if the guidance image has also an edge, and thus can strongly preserve the edge structure. The convolution kernel form of the image guided filter can be derived as:

$$\mathcal{C}_{\mathrm{IGF}}(\mathbf{p}, d) = \frac{1}{|\mathbf{w_p}|^2} \left[\sum_{\mathbf{p}' \in \mathbf{w_p}} \left(1 + \frac{\left(\mathcal{I}(\mathbf{p}) - \overline{\mathcal{I}}_{\mathbf{w_p}}\right)\left(\mathcal{I}(\mathbf{p}') - \overline{\mathcal{I}}_{\mathbf{w_p}}\right)}{\sigma^2_{\mathbf{w_p}} + \epsilon} \right) \cdot \mathcal{C}(\mathbf{p}', d) \right], \quad (4.13)$$

where $\overline{\mathcal{I}}_{\mathbf{w_p}}$ and $\sigma_{\mathbf{w_p}}$ are respectively the mean and the standard deviation of intensities in the support window $\mathbf{w_p}$ in the guided image $\mathcal{I}$. The ϵ is a regularisation term for constraining the parameters of the linear model. Details for the derivation can be found in (He et al., 2010).

Figure 4.4 shows the results of the aforementioned filters for image denoising. One can see that the box and Gaussian filter over smooth the edge while the joint bilateral and guided filter can successfully preserve it. Similar results can be expected when applying different filtering schemes to the cost aggregation. Edge-preserving local filtering has been found useful in several computer vision applications and has shown impressive performance (De-Maeztu et al., 2011; Hosni et al., 2013; Rhemann et al., 2011). The advantage of using these cutting-edge filtering algorithms for the cost aggregation is that they are very efficient and the reconstruction results are even competitive with global optimisation approaches.

Other aggregation methods

There are also works that do not not use filtering for cost aggregation. Hirschmüller (2005) performs the cost aggregation along multi-direction lines. Recently, an non-local aggregation scheme has also been devised based on a minimum spanning tree (Yang, 2012).

4.3.3 Disparity computation/optimisation

After matching cost computation and cost aggregation, the disparity map $\mathcal{D}$ can be optimised by the computed cost values. For local stereo correspondence approaches, the common disparity computation is simply done by the WTA method:

$$\mathcal{D}_{\mathrm{L}}(\mathbf{p}) = \arg\min_{d \in \mathbb{D}} \mathcal{C}(\mathbf{p}, d), \quad (4.14)$$

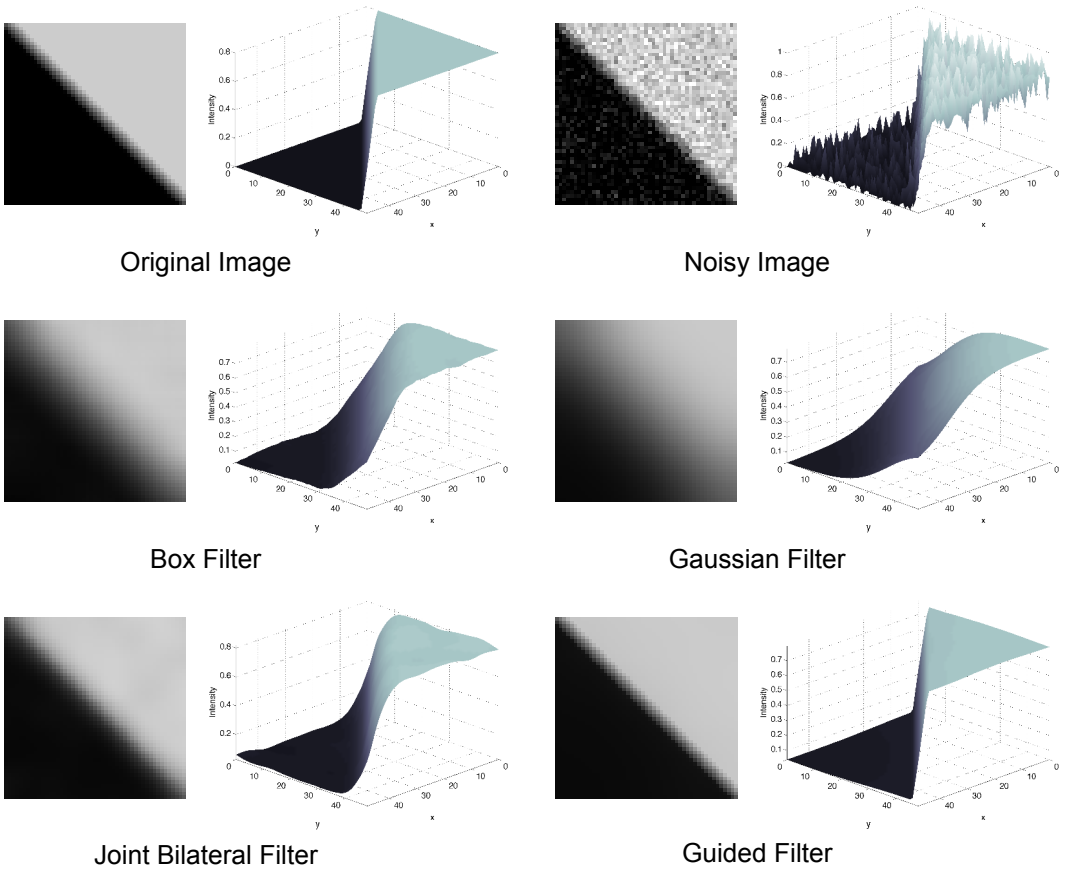


Figure 4.4: Comparison for different filters applied on a 80×80 noisy edge patch. These filters have the same kernel size 20×20. For the bottom two edge-preserving filters, the standard deviations of spatial and range similarity are set as $\{\sigma_s, \sigma_r\} = \{3.0, 0.4\}$ for the JBF, and the regulariser $\epsilon = 0.01$ for the guided filter.

for all pixel $\mathbf{p} \in \mathbf{\Omega}_{\mathcal{I}}$. This however extracts the disparity independently and ignore the neighbourhood relationship. On the other hand, global stereo correspondence approaches use the matching costs to optimise the entire disparity map.

To optimise the disparity map, as we have discussed in Section 2.4 that it is in general not possible to solve such an inverse problem directly. We have to add some smoothness assumption to the unknown solution to form the cost function:

$$\mathcal{D}_{\mathrm{G}} = \arg\min_{\mathcal{D}} \Big\{ E_{\mathrm{reg}}(\mathcal{D}) + \lambda E_{\mathrm{data}}(\mathcal{D}) \Big\}, \tag{4.15}$$

where the E_{reg} and E_{data} are known as the regulariser (smoothness) term and the data (fidelity) term respectively, and the parameter λ balances their effects. The en-

ergy functional is devised such that low-energy states reflect the physical properties of the problem so that the minimiser provides the best solution to that problem.

In a continuous setting, Equation 4.15 is formulated as a variational model and has solutions in continuous space $\mathcal{D} : \mathbf{\Omega}_{\mathcal{I}} \rightarrow \mathbb{R}_{+}$:

$$\mathcal{D}_{\mathrm{G}} = \arg\min_{\mathcal{D}} \int_{\mathbf{\Omega}_{\mathcal{I}}} \left\|\nabla \mathcal{D}\left(\mathbf{p}\right)\right\|_{s}^{s} + \lambda \left\|\rho(\mathcal{I}_{l}, \mathcal{I}_{r}, \mathbf{p}, \mathcal{D}\left(\mathbf{p}\right))\right\|_{t}^{t} d\mathbf{p}, \tag{4.16}$$

where the function ρ can be an arbitrary photometric measurement discussed in Section 4.3.1 and $s, t \geq 1$ determine what variational models are used. For example, it is a Tikhonov regularisation when $s = 2, t = 2$ (Tikhonov, 1943), a Rudin-Osher-Fatemi (ROF) TV-L^2 model when $s = 1, t = 2$ (Rudin et al., 1992) and a TV-L^1 model when $s = 1, t = 1$ (Chan and Esedoglu, 2005). Back to the early age, Marr and Poggio (1976) at first defined a model to extract stereo disparity from image pairs as a continuous variational model and associated the model with psychophysics and neurophysiology for human vision. The Tikhonov variational model was later populated by Horn and Schunck (1981) that was originally for optical flow and subsequently generalised by Poggio et al. (1985) for a variety of computer vision applications including stereo. While Tikhonov's model can be solved for the exact minimiser by Euler-Lagrange equations, models using total variation (TV) such as ROF or TV-L^1 have difficulties in differentiating the L^1 norm at origin. In that case, a primal-dual algorithm can be considered (Chambolle and Pock, 2011; Handa et al., 2011).

In the discrete setting, Equation 4.15 is formulated as a undirected graphical model, which can be Markov Random Fields (MRF) or Conditional Random Fields (CRF), and has solutions in a discrete label set $\mathcal{D} : \mathbf{\Omega}_{\mathcal{I}} \rightarrow \mathcal{L}$:

$$\mathcal{D}_{\mathrm{G}} = \sum_{\{\mathbf{p},\mathbf{p}'\} \in \mathcal{N}} \mathcal{E}_{\mathbf{p},\mathbf{p}'}(\mathcal{D}) + \sum_{\mathbf{p} \in \mathbf{\Omega}_{\mathcal{I}}} \mathcal{V}\left(\mathcal{I}_{l}, \mathcal{I}_{r}, \mathbf{p}, \mathcal{D}\right), \tag{4.17}$$

where $\mathcal{N}$ is the neighbourhood function and typically consists of adjacent pixels. The pairwise function $\mathcal{E}_{\mathbf{p},\mathbf{p}'}$ plays as the regulariser in continuous case in which each $\{\mathbf{p}, \mathbf{p}'\}$ can have its own distinct penalty (Boykov et al., 2001). It also turns out that Equation 4.17 is a Ising model when $\mathcal{L}$ is a binary label set ($|\mathcal{L}| = 2$) and a Potts model when $\mathcal{L}$ is a multiple label set ($|\mathcal{L}| > 2$). To solve Equation 4.17, discrete approaches using dynamic programming (Baker and Binford, 1981; Ohta and Kanade, 1985), graph cuts (Greig et al., 1989) and belief propagation (Yedidia

et al., 2003) have been proposed.

More about variational methods

As we have discussed in Section 2.4.2, the optimisation for Equation 4.16 is an ill-posed non-parametric model problem. From the Bayesian inference's point of view, we need a well-established model to fit the underlying disparity function with respect to the likelihood (data term) and the prior (regulariser term). Early works in computer vision used the Tikhonov regularisation to approach this as a variational problem, in which both the E_{reg} and E_{data} are embedded with a quadratic function (Horn and Schunck, 1981; Poggio et al., 1985). The statistical meaning of the Tikhonov regularisation is that both the likelihood and the prior are assumed to be Gaussian distributions, and it is well-known that a Gaussian kernel favours everywhere smooth solutions but penalises sharp discontinuity. Thus using the Tikhonov regularisation will lead to poor results at object boundaries. Energy functions using Tikhonov regularisation with robust regularisation that do not have this problem are called *discontinuity-preserving* and are later introduced (Grimson and Pavlidis, 1985; Terzopoulos, 1986; Lee and Pavlidis, 1988).

Based on a conjecture that the prior knowledge for the gradient of natural images is a L^1 norm instead of the L^2 norm, Rudin et al. (1992) attempted to use the TV norm as the regulariser. Such a model is known as a ROF model or TV-L^2 model. It turns out that the TV norm regulariser can preserved the edge for image denoising. This can be explained by robust statistics where TV norm is essentially the L^1 norm of the image derivatives (see Section 5.5.5 and Figure 5.4). As a regulariser, it removes small noise but is also able to preserve large discontinuity (outlier) during the variational optimisation (Huber, 1981; Zhang, 1997; Black et al., 1998). Later, Huang and Mumford (1999) studied the statistics of natural images and tried to fit mathematical models to these statistics. They showed that the statistics of intensity derivatives can be represented by the generalised Laplace distribution which is exactly a L^1 norm with a tractable convex configuration. This observation further complemented the conjecture of Rudin et al. (1992). A similar study for the statistics of depths in natural images followed, yielding a similar result (Huang et al., 2000).

The TV-L^1 model was later found to have the surprise property that using L^1 norm for the data term can remove strong outliers such as occlusions and salt-and-pepper noise while preserving details and contrast (Nikolova, 2004; Chan and Esedoglu, 2005). Slesareva et al. (2005) conducted the first work using TV-L^1 model for

the stereo correspondence problem. In image denosing, if the photometric function ρ in Equation 4.16 is simply the AD measurement, we can interpret the optimisation using the L^1 data term as to minimise the sum of AD in a local region:

$$\arg\min_{\mathcal{I}(\mathbf{p})} \sum_{\mathbf{p}' \in \mathbf{w_p}} \left| \mathcal{I}(\mathbf{p}) - \mathcal{I}(\mathbf{p}') \right|, \tag{4.18}$$

which is equivalent to solve $\mathcal{I}(\mathbf{p})$ for

$$\sum_{\mathbf{p}' \in \mathbf{w_p}} \frac{d\left|\mathcal{I}(\mathbf{p}) - \mathcal{I}(\mathbf{p}')\right|}{d\mathcal{I}(\mathbf{p})} = \sum_{\mathbf{p}' \in \mathbf{w_p}} \frac{\mathcal{I}(\mathbf{p}) - \mathcal{I}(\mathbf{p}')}{\left|\mathcal{I}(\mathbf{p}) - \mathcal{I}(\mathbf{p}')\right|} = 0.$$

It is obvious that $\mathcal{I}(\mathbf{p})$ must be the median intensity in order to make the summation zero. Using the L^1 norm data term is therefore similar to applying a median filter, and it is well-known that the median filter is especially effective in edge preservation and dealing with strong noise such as speckle noise and salt-and-pepper noise (Arce, 2005; Ma et al., 2013).

One drawback of using TV as a regulariser is that the tendency of L^1 norm favours piecewise fill-in constant which causes the notorious *staircasing effect*. In stereo correspondence, for instance, a slanted surface will be filled in with piecewise constant disparities, which implies fronto-parallel structures, but this is entirely incorrect. To alleviate the effect, Werlberger et al. (2009) proposed replacing the L^1 norm with a Huber norm for regularisation. Essentially a Huber norm behaves as a L^2 norm around the origin and a L^1 norm in the rest of space:

$$\|\cdot\|_\varepsilon = \begin{cases} \frac{\|\cdot\|_2^2}{2\varepsilon} & \text{if } \|\cdot\|_2 \leq \varepsilon \\ \|\cdot\|_1 - \frac{\varepsilon}{2} & \text{otherwise,} \end{cases} \tag{4.19}$$

where the ε is a small constant delimiting the two norms, as shown in Figure 4.5. The homogeneous areas (with small derivatives) can thus be regularised by the L^2 norm and the discontinuous area (with large derivatives) is regularised by the L^1 norm. Figure 4.6 shows an example of using Huber norm to alleviate the staircasing effect.

The aforementioned TV regularisers, using either the L^1 or Huber norm are eventually a first-order regularisation scheme. Specifically, first-order TV is only able to regularise piecewise constant (fronto-parallel) structures. To theoretically

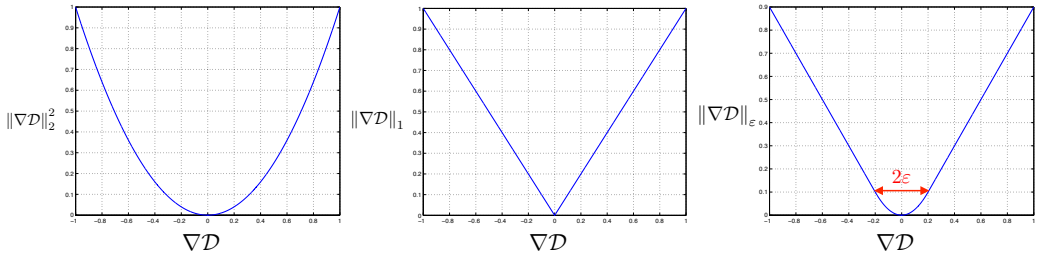
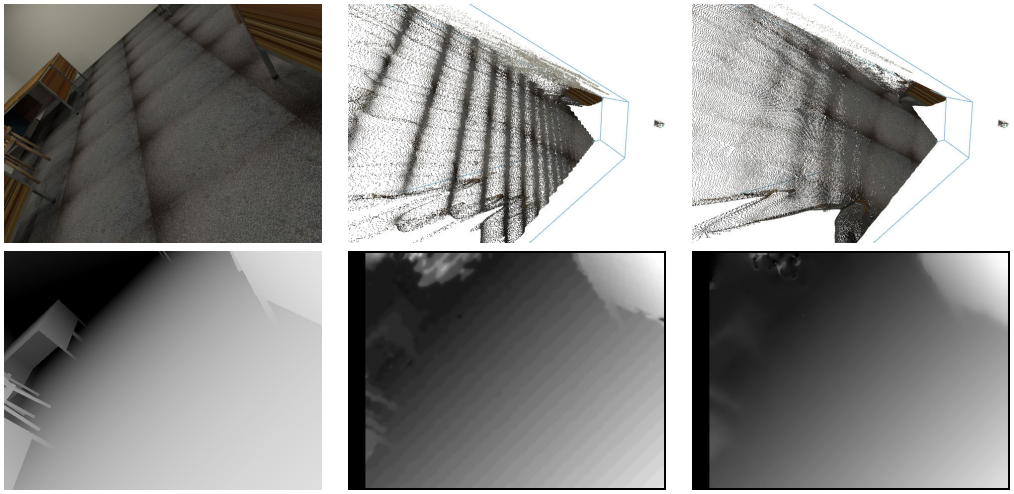


Figure 4.5: The Huber norm (right) is essentially a L^2 norm (left) around the origin and a L^1 norm (middle) in the rest of space. In this example, the delimitation for the Huber norm is $\varepsilon = 0.2$.



(a) (b) (c)

Figure 4.6: Staircasing effect. (a) A scene with a slant surface and its ground truth disparity map. (b) Reconstructed point cloud and disparity map using L^1 norm regulariser. Note that the obvious staircasing effect is shown as piecewise constant depth (fronto-parallel) structures. (c) Reconstructed point cloud and disparity map using Huber norm regulariser. Note that the staircasing effect is much alleviated.

regularise on piecewise affine (slanted) structures, using a higher order regularisation algorithm is required. Having realised this, Bredies et al. (2010) proposed a total generalised variation (TGV) framework which composes of piecewise polynomials of arbitrary order of TV for regularisation. Ranftl et al. (2012, 2013) further showed that second-order TGV regularisation can reconstruct slanted structures naturally.

More about discrete methods

Greig et al. (1989) at first proposed graph cuts for exactly solving Equation 4.17 as the Ising model for black-white image restoration. This idea is later extended

to problems with a greater number of labels (gray) image, exactly solving by using some rather restricted energy functions (Ferrari et al., 1995). In 2D space, while the Ising model can be solved for the exact minimum, it has been proven NP-hard for the Potts model (Boykov et al., 2001; Veksler, 1999). Most of works for solving multi-label Potts model thus seek for an approximate minimum (Boykov et al., 2001; Kolmogorov and Zabih, 2001; Veksler, 1999). Kolmogorov and Zabin (2004) further showed that what kind of energy functions for the Potts model can be solved by the graph cuts. Despite the NP-hard issue, Ishikawa (2003) showed that in some special settings, such as if the pairwise interactions (regulariser term) are convex in terms of a linearly ordered label set, one can compute the exact solution for a multi-label problem. In addition, belief propagation was also proposed for solving the Potts model (Felzenszwalb and Huttenlocher, 2006; Sun et al., 2003; Yedidia et al., 2003) but comparison has shown that the graph cuts and belief propagation have similar performance in the stereo correspondence problem (Tappen and Freeman, 2003).

Another stream in discrete methods for solving Equation 4.17 is based on dynamical programming. Although the 2D multi-label problem is NP-hard, a dynamical programming algorithm can find the global optimum for independent 1D scanlines in polynomial time. In the very beginning the idea was proposed by Baker and Binford (1981) who optimised the disparity based on edge matching following by an optimisation using the dynamical programming. More recent approaches instead have focused on the intensity-based scanline optimisation problem Belhumeur (1996); Birchfield and Tomasi (1998); Bobick and Intille (1999). Hirschmüller (2005, 2008) proposed a *semi-global matching* (SGM) algorithm which calculates an overall cost along many 1D scanlines to cover all directions using dynamic programming yielding no streaking artifacts. A real-time implementation of SGM using the GPU was also addressed (Ernst and Hirschmüller, 2008).

Continuous vs. discrete methods

The main difference between the continuous methods and the discrete methods is that the solution space is continuous or discrete. Until now, it is probably still an open debate about which methodology is the clear winner. An apparent advantage of using variational approaches is that the optimised disparity map can directly reach sub-pixel accuracy, whereas discrete methods require further post-processing. In addition, graph cuts methods suffer from the so-called metrication error (blocky artifacts) and dynamic programming methods have the streaking problem. In contrast, variational approaches can instead model the geometric property in the regulari-

sation scheme directly. While the variational approaches usually require iterative optimisation algorithms, which is in general much slower than discrete methods, most of the computation are for intensity derivatives, photometric, etc., which are basically pixel-wise operations and can be efficiently parallelised and implemented by using modern GPGPU techniques to reach real-time performance.

4.3.4 Stereo correspondence in endoscopes

Reconstruction of the stereo-endoscopic view for surgical navigation has been an active area of research for over a decade. Reconstructing the depth of stereo-endoscopic scenes is an important step in providing accurate guidance in robotic-assisted MIS. With the da Vinci surgical system (Figure 4.1a) the presence of a stereoscopic laparoscope means that computational stereo is a practical and feasible approach to *in vivo* reconstruction (Devernay et al., 2001). Mirota et al. (2011) indicated a variety of image-guided surgery applications demand scene reconstruction. However, surgical scenes are challenging for 3D reconstruction algorithms because of texture-poor appearance, occlusions, specular reflection and discontinuities due to instruments. Mountney et al. (2006, 2010) first showed that sparse 3D features can be used as a map for endoscopic camera tracking. Much of the prior work has focused on beating heart surgery (Devernay et al., 2001; Lau et al., 2004; Richa et al., 2011; Stoyanov et al., 2010) where the reconstructed heart surface could be used for motion stabilisation or registration to a preoperative model.

To achieve smooth and robust stereo reconstruction, Lau et al. (2004) proposed a parametric surface description to overcome texture homogeneity. A similar idea using a thin-plate spline was also proposed (Richa et al., 2011). Alternatively, Stoyanov et al. (2010) devised a region growing algorithm starting from sparse features toward a semi-dense reconstruction, and the reconstruction was further used for estimating dense scene flow (Stoyanov, 2012). Rohl et al. (2012) proposed a hybrid CPU-GPU framework to fuse temporal reconstruction into a global model.

4.3.5 Benchmarking datasets

Figure 4.7a shows a few popular benchmarking datasets. Scharstein and Szeliski (2002) first published the Middlebury Stereo Datasets[1] for benchmarking stereo correspondence algorithms and over the years, new challenging stereo images have been emerging (Scharstein and Szeliski, 2003; Scharstein and Pal, 2007; Hirschmuller and Scharstein, 2007). The Middlebury stereo datasets provide a variety of objects

[1]Middlebury: http://vision.middlebury.edu/stereo/data

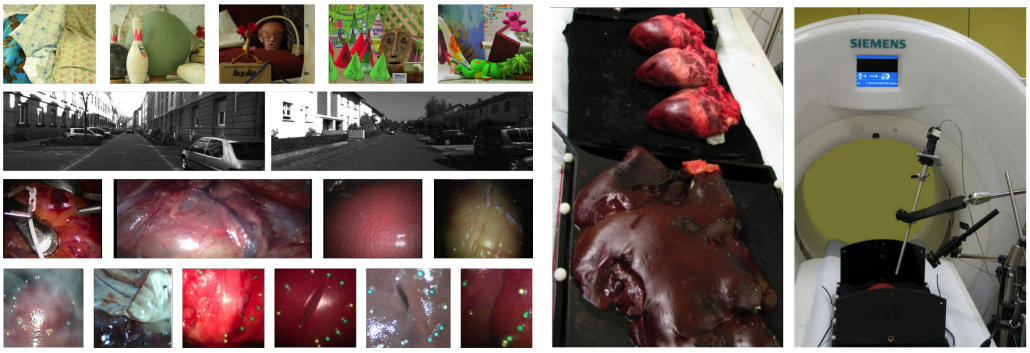
and the corresponding disparity ground truths for stereo correspondence validation, but it is mainly for indoor scenes. In contrast, Geiger et al. (2012) proposed a KITTI datasets[2] which were designed particularly for autonomous car research, and hence the stereo images are general street views. The Hamlyn Centre laparoscopic/endoscopic video datasets[3] provide a number of stereo laparoscopic/endoscopic videos recorded in different surgical procedures, but only two cardiac phantom datasets are provided with the 3D ground truth. Recently the Open-CAS datasets[4] were published to provide a number of stereo endoscopic images from different surgical situations, such as with smoke, blood, cuts, and most importantly all the images have the corresponding ground truth models (Maier-Hein et al., 2014).

While computer vision research can be simply carried out with commercial webcams, working with medical data requires access to hospitals, patient consent and expensive hardware for validation (CT/MRI). In addition, *in vitro* organs are very difficult to maintain under room temperature and hence the setting requires careful arrangement. Figure 4.7b shows an example for the setup using in the Open-CAS datasets. In addition, the content of medical images is in general very different to indoor and outdoor scenes. The video data involves significant deformation due to respiration and instrument tissue interaction, and as a stereo image pair it has severe specular highlights, instrument occlusions, textureless appearance and unstructured surfaces. Overall medical images are quite challenging in stereo matching.

4.4 Proposed approach

Stereo reconstruction has been studied for decades but still remains a challenge in endoscopic imaging. Current approaches can easily fail to reconstruct an accurate and smooth 3D model due to textureless tissue appearance of the real surgical scene and occlusion by instruments. We build on recent advances in computer vision and the use of variational techniques to efficiently and effectively reconstruct stereo endoscopic scenes using stereo image pairs. This is achieved by constructing a cost volume with a reliable ZNCC data term and performing convex optimisation to solve a Huber-ZNCC model. The proposed algorithm can be effectively parallelised on the GPU for real-time performance. Compared with the state of the art, the proposed approach yields more accurate reconstruction in empirical studies.

[2]KITTI: http://www.cvlibs.net/datasets/kitti
[3]Hamlyn: http://hamlyn.doc.ic.ac.uk/vision
[4]Open-CAS: http://opencas.webarchiv.kit.edu



(a) (b)

Figure 4.7: (a) From the first row: Middlebury, KITTI, Hamlyn and Open-CAS stereo image datasets. (b) Setup for building the Open-CAS datasets. The left image exhibits the Porcine organs used in laparoscopic trainer. The right image shows the organ images taken in an interventional radiology suite on the stretcher of a SIEMENS CT scanner.

4.4.1 Cost volume construction

As we have discussed in 4.3.1, the data fidelity is essential since the later convex optimisation significantly relies on it. We illustrate the effects of different measures with a synthetic stereo pair generated by a textured cone model as shown in Figure 4.8a. Raw reconstruction is achieved using the WTA scheme. The simplest AD (Equation 4.3) gives a very noisy raw reconstruction as shown in Figure 4.8b. To reduce the noise, we can apply edge-preserving local filtering for the cost aggregation (Rhemann et al., 2011; Rohl et al., 2012), and Figure 4.8c shows the result after JBF (Equation 4.12) is applied to Figure 4.8b. However, the empirical studies have shown that if the original measurement is error-prone, the subsequent cost aggregation in the cost volume will not help to improve the overall accuracy. This commonly happens in textureless regions, half-occluded areas and where the illumination changes. In contrast, ZNCC (Equation 4.8) implicitly performs the aggregation in the window patch, so correlation is calculated over a pixel neighbourhood already. This results in a measurement more tolerant to different camera gain or luminance changes and can also provide better fidelity in textureless regions. Figure 4.8d shows the raw reconstruction using ZNCC which is similar to the result of Figure 4.8c, and we can also observe the avoidable fronto-parallel effect due to the discrete cost volume and the simple WTA disparity computation.

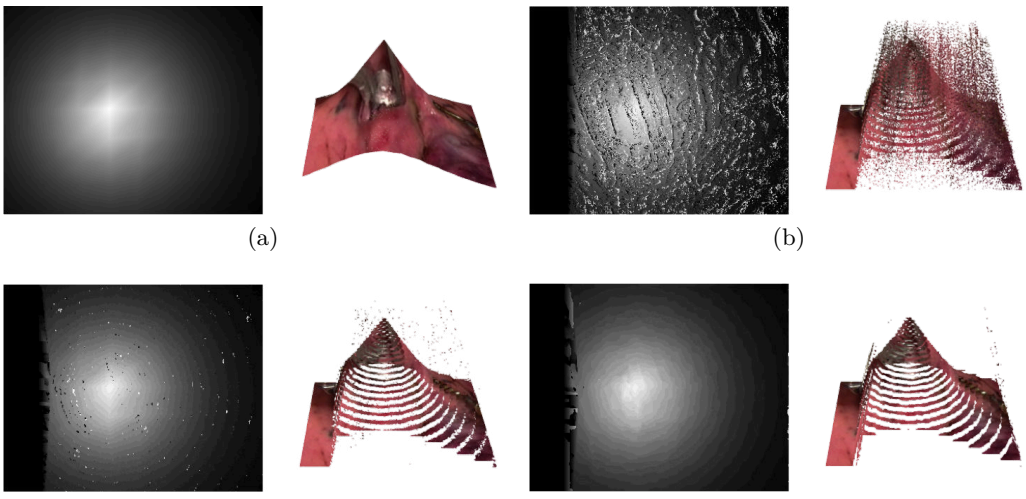


Figure 4.8: Disparity computation results using the WTA scheme with different cost volumes. In each synthetic image pair, the disparity map and its 3D point cloud are shown in the left and right respectively. (a) The synthetic image pair is generated by projecting a textured cone model (ground truth). (b) The result of using AD cost volume. (c) The result after applying the JBF to (b). (d) The result of using ZNCC cost volume.

4.4.2 Disparity optimisation

After the cost volume is constructed and the disparity computation is done by using the WTA scheme, the disparity function $\mathcal{D}_{\mathrm{G}}$ is further optimised by solving a Huber-ZNCC model which is similar to a Huber-L^1 model proposed by Werlberger et al. (2009). But instead of using the L^1 data term, we replace it with the ZNCC cost volume:

$$\mathcal{D}_{\mathrm{G}} = \arg\min_{\mathcal{D}} \int_{\boldsymbol{\Omega}_{\mathcal{I}}} \|\mathcal{F}(\mathbf{p})\nabla\mathcal{D}(\mathbf{p})\|_{\varepsilon} + \lambda C_{\mathrm{ZNCC}}(\mathbf{p}, \mathcal{D}(\mathbf{p}))\, d\mathbf{p}. \tag{4.20}$$

As ZNCC can cause over smooth results around the edges, to enhance the discontinuity preservation, the anisotropic diffusion tensor $\mathcal{F}$ for weighting the regulariser is introduced:

$$\mathcal{F}(\mathbf{p}) = \exp(-\alpha|\nabla\mathcal{I}(\mathbf{p})|^{\beta})nn^{\top} + n^{\perp}n^{\perp\top}, \tag{4.21}$$

where α and β adjust the edge weighting, and n is the normalised vector of image

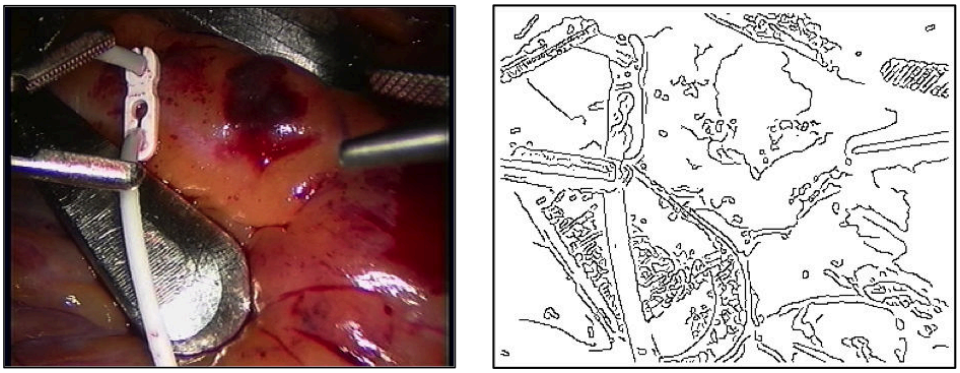


Figure 4.9: Image edges usually accompany with discontinuous depths.

gradient $n = \frac{\nabla \mathcal{I}(\mathbf{p})}{|\nabla \mathcal{I}(\mathbf{p})|}$ and $n^{\perp}$ the perpendicular vector. The idea is that the discontinuity of disparity usually happens around image edges as shown in Figure 4.9. So the strictly decreasing function $\exp(-\alpha|\nabla \mathcal{I}(\mathbf{p})|^{\beta})$ controls the penalty by the magnitude of the image gradient and the symmetric matrix $nn^{\top}$ and $n^{\perp}{n^{\perp}}^{\top}$ further make the derivative of disparity anisotropic for regularisation.

Since Equation 4.20 is non-convex in its data term, the ZNCC cost volume, conventional methods resort to local linearisation and adopt a coarse-to-fine scheme (Zach et al., 2007; Werlberger et al., 2009) to approach the global minimum (Section 5.5.6). However, this generally requires a good initial state. In addition, small structures are lost in the coarsest layer and cannot be recovered in the finer layers. By contrast, having a cost volume helps us to avoid such a warping scheme and preserve the details.

Following a large displacement optical flow algorithm (Steinbrücker et al., 2009), we decouple the regulariser term and the data term in Equation 4.20 by introducing an auxiliary function $\mathcal{A} : \mathbf{\Omega}_I \to \mathbb{D}$ to form a new energy functional:

$$\mathcal{D}_{\mathrm{G}} = \arg\min_{\mathcal{D}} \int_{\mathbf{\Omega}_{\mathcal{I}}} \|\mathcal{F}(\mathbf{p})\nabla\mathcal{D}(\mathbf{p})\|_{\varepsilon} + Q(\mathcal{D}(\mathbf{p}), \mathcal{A}(\mathbf{p})) + \lambda C_{\mathrm{ZNCC}}(\mathbf{p}, \mathcal{D}(\mathbf{p}))\, d\mathbf{p}, \tag{4.22}$$

where

$$Q(\mathcal{D}, \mathcal{A}) = \frac{\theta}{2}(\mathcal{D} - \mathcal{A})^2 .$$

The first part $\|\mathcal{F}\nabla\mathcal{D}\|_{\varepsilon} + Q(\mathcal{D}, \mathcal{A})$ is in fact a Huber-ROF model similar to the original ROF (TV-L^2) model (Rudin et al., 1992) in which the global minimiser for

$\mathcal{D}$ is guaranteed to be found. We thus solve the Huber-ROF model by using the primal-dual algorithm which provides a linear convergence rate $O(1/N)$ (Chambolle and Pock, 2011). The exact minimum of the later part $Q(\mathcal{D}, \mathcal{A}) + \lambda C_{\text{ZNCC}}$ can be found by simply performing an exhaustive search in the cost volume for $\mathcal{A}$. The entire optimisation for Equation 4.22 can therefore be alternately solved and the parameter $\theta \to \infty$ ensures $\mathcal{D} \simeq \mathcal{A}$ when converging.

Note that the Equation 4.20 is actually a first-order TGV model which is only able to reconstruct fronto-parallel structure (Ranftl et al., 2013). However we have observed that instead of applying a rather expensive second-order TGV to reconstruct the slanted structure, using the Huber norm for the regulariser term is a good approximation that avoids the staircasing effect. The same strategy is also used in the state of the art stereo reconstruction algorithm of Heise et al. (2013).

4.5 Non-parametric optimisation

To solve Equation 4.20, it is convenient to have a discrete setting for the problem. We constrain and unfold the 2D image space $\mathbf{\Omega}_{\mathcal{I}}$ to a finite dimensional scalar vector $\mathbb{X} = \mathbb{R}^{MN}$ in a regular Cartesian grid of size $M \times N$:

$$\{(i,j) \,|\, 1 \leq i \leq M, 1 \leq j \leq N\}.$$

The inner product of two scalar field $\mathbf{a}, \mathbf{b} \in \mathbb{X}$ is simply:

$$\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbb{X}} = \sum_{i,j} \mathbf{a}_{i,j} \mathbf{b}_{i,j}, \quad \mathbf{a}, \mathbf{b} \in \mathbb{X}.$$

We define the gradient operator $\nabla : \mathbb{X} \to \mathbb{Y}$ where $\mathbb{Y} = \mathbb{R}^{2MN}$. With Neumann boundary conditions, the gradient of a scalar field $\mathbf{u}$ is computed by forward finite difference:

$$(\nabla \mathbf{a})_{i,j} = \begin{pmatrix} (\nabla \mathbf{a})^1_{i,j} \\ (\nabla \mathbf{a})^2_{i,j} \end{pmatrix}, \tag{4.23}$$

where

$$(\nabla \mathbf{a})^1_{i,j} = \begin{cases} \mathbf{a}_{i+1,j} - \mathbf{a}_{i,j} & \text{if } i < M \\ 0 & \text{if } i = M \end{cases}, \quad (\nabla \mathbf{a})^2_{i,j} = \begin{cases} \mathbf{a}_{i,j+1} - \mathbf{a}_{i,j} & \text{if } j < N \\ 0 & \text{if } j = N. \end{cases}$$

The inner product of two vector field $\mathbf{c}, \mathbf{d} \in \mathbb{Y}$ is:

$$\langle \mathbf{c}, \mathbf{d} \rangle_{\mathbb{Y}} = \sum_{i,j} \mathbf{c}^1_{i,j} \mathbf{d}^1_{i,j} + \mathbf{c}^2_{i,j} \mathbf{d}^2_{i,j}, \quad \mathbf{c} = \begin{pmatrix} \mathbf{c}^1 \\ \mathbf{c}^2 \end{pmatrix}, \mathbf{d} = \begin{pmatrix} \mathbf{d}^1 \\ \mathbf{d}^2 \end{pmatrix} \in \mathbb{Y}.$$

We further define the divergence operator $\nabla\cdot : \mathbb{Y} \to \mathbb{X}$ which maps a vector field back to a scalar field. The divergence operator is in fact the transpose of the negative gradient operation:

$$\nabla\cdot \equiv -\nabla^\top. \tag{4.24}$$

Thus the divergence of a vector field $\mathbf{c}$ can be computed explicitly with the boundary condition corresponding to the gradient operator:

$$(\nabla \cdot \mathbf{c})_{i,j} = \begin{cases} \mathbf{c}^1_{i,j} - \mathbf{c}^1_{i-1,j} & \text{if } 1 < i < M \\ \mathbf{c}^1_{i,j} & \text{if } i = 1 \\ \mathbf{c}^1_{i-1,j} & \text{if } i = M \end{cases} + \begin{cases} \mathbf{c}^2_{i,j} - \mathbf{c}^2_{i,j-1} & \text{if } 1 < j < N \\ \mathbf{c}^2_{i,j} & \text{if } j = 1 \\ \mathbf{c}^2_{i,j-1} & \text{if } j = N. \end{cases}$$

From Equation 4.24, we can also derive the identity:

$$\langle \nabla \mathbf{a}, \mathbf{c} \rangle_{\mathbb{Y}} = -\langle \mathbf{a}, \nabla \cdot \mathbf{c} \rangle_{\mathbb{X}}. \tag{4.25}$$

4.5.1 Solving Huber-ROF model for $\mathcal{D}$

The core concept of the primal-dual algorithm is to solve a saddle-point problem for a functional in primal and dual space alternately. The primal variable is simply the solution we are seeking for and the dual variable is introduced by performing the Legendre-Fenchel transformation on the primal variable. The reason why we would like to work in the dual space is that a Legendre-Fenchel conjugate is always convex. The proof is given by Handa et al. (2011). Legendre-Fenchel transformation

is particularly useful when we are dealing with a non-smooth convex function such as a L^1 norm which is not differentiable at the origin.

Given the vectors of primal variables $\mathbf{x}, \mathbf{a} \in \mathbb{X}$ and a dual variable $\mathbf{y} \in \mathbb{Y}$, we reform the minimisation problem of Equation 4.22 to be a saddle-point problem:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ \varphi\left(\mathbf{x}, \mathbf{y}\right) + \frac{\theta}{2}\left(\mathbf{x} - \mathbf{a}\right)^2 \right\}, \tag{4.26}$$

where the function φ is the Lengendre-Fenchel transformation of the Huber norm (Equation 4.19):

$$\varphi\left(\mathbf{x}, \mathbf{y}\right) = -\left\langle \mathbf{x}, \nabla \cdot \mathbf{W}\mathbf{y} \right\rangle_{\mathbb{X}} - \delta_{\mathbf{s}}(\mathbf{y}) - \frac{\varepsilon}{2}\left\|\mathbf{y}\right\|_2^2, \tag{4.27}$$

where the indicator function $\delta_{\mathbf{s}}$ and the convex set $\mathbf{s}$ are:

$$\delta_{\mathbf{s}}\left(\mathbf{y}\right) = \begin{cases} \frac{\varepsilon}{2} & \text{if } \mathbf{y} \in \mathbf{s} \\ \infty & \text{if } \mathbf{y} \notin \mathbf{s} \end{cases}, \quad \text{and} \quad \mathbf{s} = \left\{ \mathbf{y} \in \mathbb{Y} \,:\, \varepsilon < \left\|\mathbf{y}\right\|_\infty \leq 1 \right\}. \tag{4.28}$$

Note that the primal vector $\mathbf{x}$ and $\mathbf{a}$ are the unfolded disparity function $\mathcal{D}$ and auxiliary function $\mathcal{A}$ respectively. The $2MN \times 2MN$ diagonal element-wise weighting matrix $\mathbf{W}$ is form by using Equation 4.21. Using the identity in Equation 4.25, the gradient flow of Equation 4.26 with respect to $\mathbf{x}$ is:

$$\frac{\partial E\left(\mathbf{x}, \mathbf{a}, \mathbf{y}\right)}{\partial \mathbf{x}} = \frac{\partial \varphi\left(\mathbf{x}, \mathbf{y}\right)}{\partial \mathbf{x}} + \frac{1}{2\theta}\frac{\partial \left(\mathbf{x} - \mathbf{a}\right)^2}{\partial \mathbf{x}} = -\nabla \cdot \mathbf{W}\mathbf{y} + \theta\left(\mathbf{x} - \mathbf{a}\right), \tag{4.29}$$

and with respect to $\mathbf{y}$ is:

$$\frac{\partial E\left(\mathbf{x}, \mathbf{a}, \mathbf{y}\right)}{\partial \mathbf{y}} = \frac{\partial \varphi\left(\mathbf{x}, \mathbf{y}\right)}{\partial \mathbf{y}} - \frac{\partial \delta_{\mathbf{s}}\left(\mathbf{y}\right)}{\partial \mathbf{y}} - \frac{\varepsilon}{2}\frac{\partial \left\|\mathbf{y}\right\|_2^2}{\partial \mathbf{y}} = \mathbf{W}^\top \nabla \mathbf{x} - \varepsilon \mathbf{y}. \tag{4.30}$$

We can therefore solve Equation 4.26 iteratively. Specifically, at iteration n, we first apply gradient ascent with a step size σ to update the dual variable $\mathbf{y}$:

$$\frac{\mathbf{y}^{n+1}-\mathbf{y}^{n}}{\sigma}=\frac{\partial E\left(\mathbf{x},\mathbf{a},\mathbf{y}\right)}{\partial\mathbf{y}} \quad \underset{\text{Equation 4.30}}{\Rightarrow} \quad \mathbf{y}^{n+1}=\frac{\sigma\mathbf{W}^{\top}\nabla\mathbf{x}^{n}+\mathbf{y}^{n}}{1+\varepsilon\sigma} \quad \underset{\text{Equation 4.28}}{\Rightarrow} \quad \mathbf{y}^{n+1}=\frac{\mathbf{y}^{n+1}}{\max\left(1,\|\mathbf{y}^{n+1}\|_{2}\right)}, \tag{4.31}$$

following by applying gradient descent with a step size τ to update the primal variable $\mathbf{x}$:

$$\frac{\mathbf{x}^{n+1}-\mathbf{x}^{n}}{\tau}=-\frac{\partial E\left(\mathbf{x},\mathbf{a},\mathbf{y}\right)}{\partial\mathbf{x}} \quad \underset{\text{Equation 4.29}}{\Rightarrow} \quad \mathbf{x}^{n+1}=\frac{\mathbf{x}^{n}+\tau\nabla\cdot\mathbf{W}\mathbf{y}^{n}+\tau\theta\mathbf{a}^{n}}{1+\tau\theta}. \tag{4.32}$$

In practice, $\mathbf{W}=\mathbf{W}^{\top}$ (symmetricity) and we perform the gradient ∇ and divergence $\nabla\cdot$ operator together with the weighting matrix $\mathbf{W}$ in-place for each pixel. The updates of primal and dual variables are thus pixel-wise computations and can be efficiently parallelised for running on the GPU. We use preconditioning to decide the step size τ and σ, which can significantly reduce the iteration time (Pock and Chambolle, 2011). Figure 4.10 shows the convergence curves with and without preconditioning for the Huber-ROF optimisation for the synthetic image pair using the ZNCC cost-volume (Figure 4.8d).

4.5.2 Exhaustive search for $\mathcal{A}$

Having updated the $\mathbf{x}^{n+1}$, we simply perform an exhaustive search to update $\mathbf{a}$:

$$\mathbf{a}^{n+1}=\arg\min_{\mathbf{a}}\left\{\frac{\theta}{2}\left(\mathbf{x}^{n+1}-\mathbf{a}\right)^{2}+\lambda\mathcal{C}_{\text{ZNCC}}\left(\mathbf{a}\right)\right\}. \tag{4.33}$$

This can also be parallelised and run on the GPU efficiently.

The optimisation steps for solving primal variable $\mathbf{x}$ and auxiliary variable $\mathbf{a}$ proceed until convergence. The parameter θ is adjusted by a *smoothstep* function over the iterations. The proposed algorithm for stereo reconstruction is summarised in Algorithm 1.

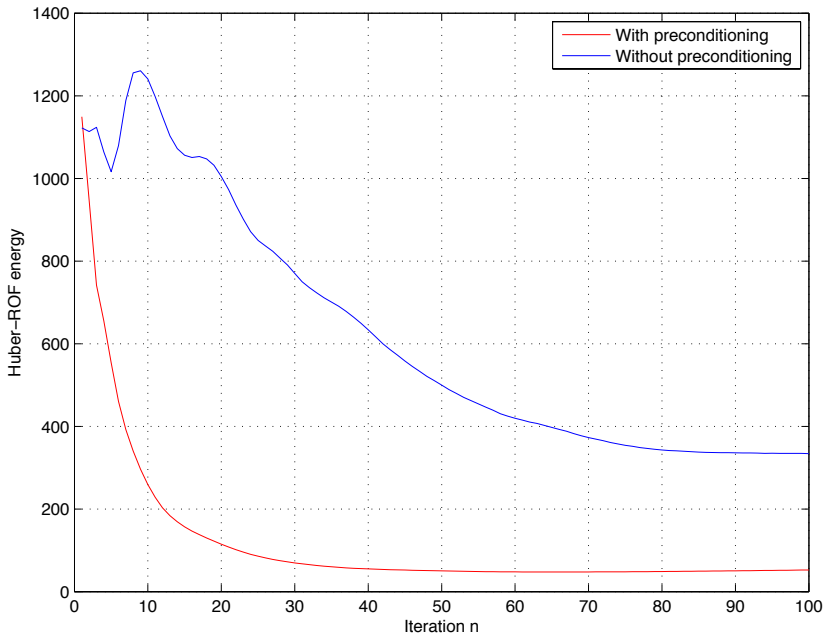


Figure 4.10: The Huber-ROF optimisation with and without preconditioning for the synthetic image pair using ZNCC cost-volume.

Algorithm 1: The proposed stereo construction algorithm

Data: Rectified stereo image pair $\mathcal{I}_l$ and $\mathcal{I}_r$
Result: The global optimised disparity map $\mathcal{D}_{\mathrm{G}}$

Construct $\mathcal{C}_{\mathrm{ZNCC}}$ (Equation 4.8);
Initialise $\mathbf{y}^0 = \mathbf{0}$, $\mathbf{x}^0 = \mathbf{a}^0 \leftarrow \arg\min_{\mathcal{D}} \mathcal{C}_{\mathrm{ZNCC}}(\mathcal{D}), n = 0$;
while $n \neq n_{max}$ **do**
    Update dual variable $\mathbf{y}^{n+1}$ (Equation 4.31) ;
    Update primal variable $\mathbf{x}^{n+1}$ (Equation 4.32) ;
    Update auxiliary variable $\mathbf{a}^{n+1}$ (Equation 4.33) ;
    $\theta = 2\lambda(3(\frac{n}{n_{max}})^2 - 2(\frac{n}{n_{max}})^3)$;
    $n = n + 1$;
end
$\mathcal{D}_{\mathrm{G}} \leftarrow \mathbf{x}^{n_{max}}$;

4.6 Empirical studies

We conducted all experiments on a workstation equipped with 3.1 GHz quad cores CPU and one Nvidia GeForce GTX 670 graphics card with 2 GB global memory. To maximally exploit the power of parallel computation, all the calculations in-

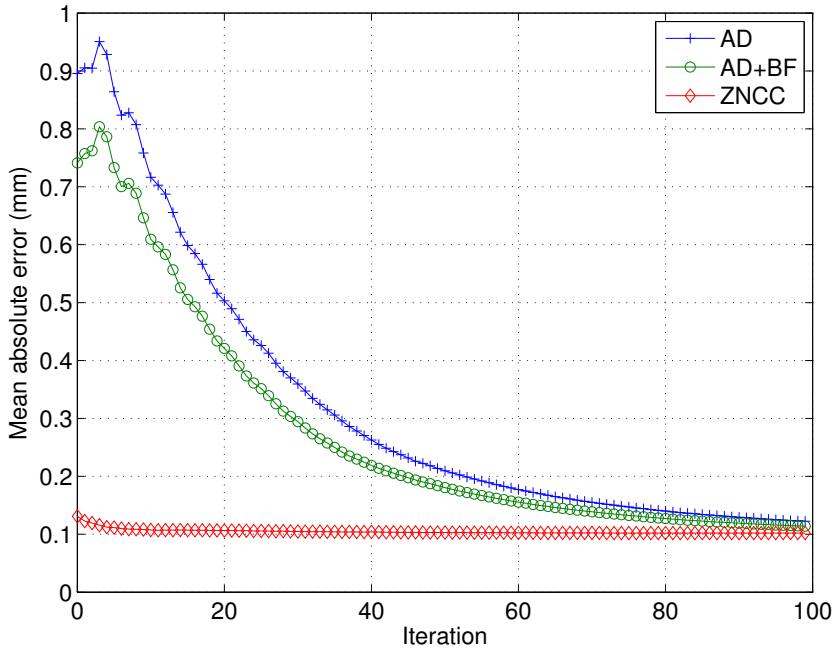


Figure 4.11: Convergence curve of the convex optimisation on different cost volumes built for the synthetic image pair with zero noise $\sigma_n = 0$.

cluding the cost volume construction, WTA, the Huber-ROF optimisation and the exhaustive line search are implemented from scratch using CUDA.

The cost volume construction using ZNCC has time complexity $O(s^2 N \cdot |\mathbb{D}|)$ and the global disparity optimisation has linear convergence rate $O(1/N)$. Currently the proposed reconstruction approach is able to run at 20 fps with the resolution $|\mathbf{\Omega}_I| = 360 \times 288$ and $|\mathbb{D}| = 32$. Specifically, the construction of the cost volume $\mathcal{C}_{ZNCC}$ using 5×5 window takes around 15 ms and the global optimisation for 100 iteration takes around 35 ms.

We first conduct a noise study to evaluate the robustness for different measures. The proposed approach is then quantitatively evaluated using cardiac phantom datasets and a number of real organ image pairs with known ground truth. Images in real robot-assisted laparoscopic prostatectomy are reconstructed for qualitative evaluation.

4.6.1 Synthetic noise study

To investigate the robustness of different data terms, we intentionally add AWGN to the stereo image pair of the cone model (Figure 4.8a). In this experiment,

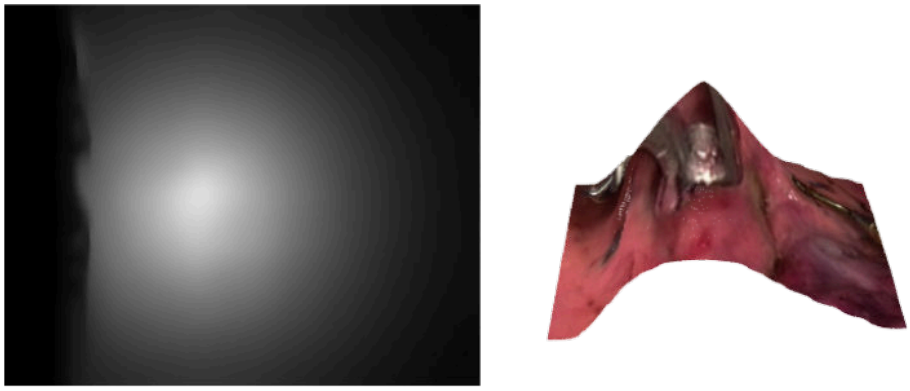


Figure 4.12: Reconstruction result on the synthetic image pair with zero noise $\sigma_n = 0$ using the proposed approach.

the ZNCC cost volume is constructed by using disparity range $\mathcal{D} = [50, 80]$ and window size 3×3, and the Huber-ROF parameters are set as $\{\epsilon, \alpha, \beta, \lambda, n_{max}\} = \{0.001, 10.0, 1.0, 5.0, 200\}$. Figure 4.11 shows the convergence curve for $\sigma_n = 0$ and the resulting disparity and 3D model are shown in Figure 4.12. The mean absolute error (MAE) of the reconstruction compared with the ground truth with different noise variance are reported in Table 4.1. The results show that there is not much difference between different measures when the image is clean. However, when the noise level becomes large, the measure using the simplest pixel-to-pixel AD degrades significantly. In contrast, AD+JBF and ZNCC, which perform local cost aggregation, remain accurate in the presence of noise. ZNCC has the best performance in all cases. One can also observe that the proposed disparity optimisation brings the discrete cone model shown in Figure 4.8d to a continuous one (i.e., sub-pixel accuracy) shown in Figure 4.12.

Table 4.1: Under different degrees of noise σ_n, the reconstruction MAE (mm) compared with the ground truth after the convex optimisation using different data terms for the stereo pair of the cone model.

	$\sigma_n = 0$	$\sigma_n = 0.01$	$\sigma_n = 0.015$	$\sigma_n = 0.02$
AD	0.121	0.623	0.877	2.035
AD+JBF	0.121	0.189	0.798	1.521
ZNCC	**0.102**	**0.185**	**0.661**	**1.487**

Table 4.2: Statistics of different algorithms with respect to MAE, RMSE and the percentage of reconstructed points compared with the ground truth.

		Proposed Approach	SPFP	FCVF	DRTI
Cardiac1	MAE(mm)	**1.24 ± 0.89**	2.36 ± 0.92	4.87 ± 0.87	1.45
	RMSE(mm)	**1.85 ± 0.82**	3.876 ± 0.87	8.24 ± 0.92	N/A
	Density(%)	**100**	92	100	N/A
Cardiac2	MAE(mm)	**1.47 ± 1.23**	3.20 ± 1.15	5.37 ± 1.53	1.53
	RMSE(mm)	**2.66 ± 1.47**	4.85 ± 1.82	7.73 ± 1.56	N/A
	Density(%)	**100**	90	100	N/A

4.6.2 Hamlyn phantom study

The proposed algorithm is quantitatively evaluated by two cardiac datasets collected from the Hamlyn datasets, which have an associated registered CT model as ground truth. It should be noted that the ground truth is generated by a 3D to 2D point-based registration process, in which small registration errors are inevitably introduced.

Before doing the reconstruction, the stereo image pair are rectified by the provided camera calibration. We further remove the black background by setting an intensity threshold, since such a background does not occur in real surgical images and also it may cause bias when comparing different algorithms. The disparity images are cropped by 15 pixels at the image borders when doing the statistics. In this experiment the disparity range is set as $\mathcal{D} = [30, 60]$ and window size 10×10. The Huber-ROF parameters are set as $\{\epsilon, \alpha, \beta, \lambda, n_{max}\} = \{0.001, 10.0, 1.0, 0.1, 200\}$.

In Table 4.2, the MAE and RMSE to the ground truth point are reported for different real-time dense algorithms using a single stereo pair. The corresponding standard deviation among all frames is also reported. The reconstruction results for a single stereo frame are shown in Figure 4.13. Structure propagation using sparse feature points (SPFP) (Stoyanov et al., 2010) is a real-time quasi-dense method and fast cost-volume filtering (FCVF) (Rhemann et al., 2011) is a local edge-preserving filtering method. A recent real-time dense reconstruction using temporal information (DRTI) algorithm (Rohl et al., 2012) that produces highly accurate reconstruction is also compared for the best results of MAE which are quoted in their paper. It is evident that our algorithm outperforms the others.

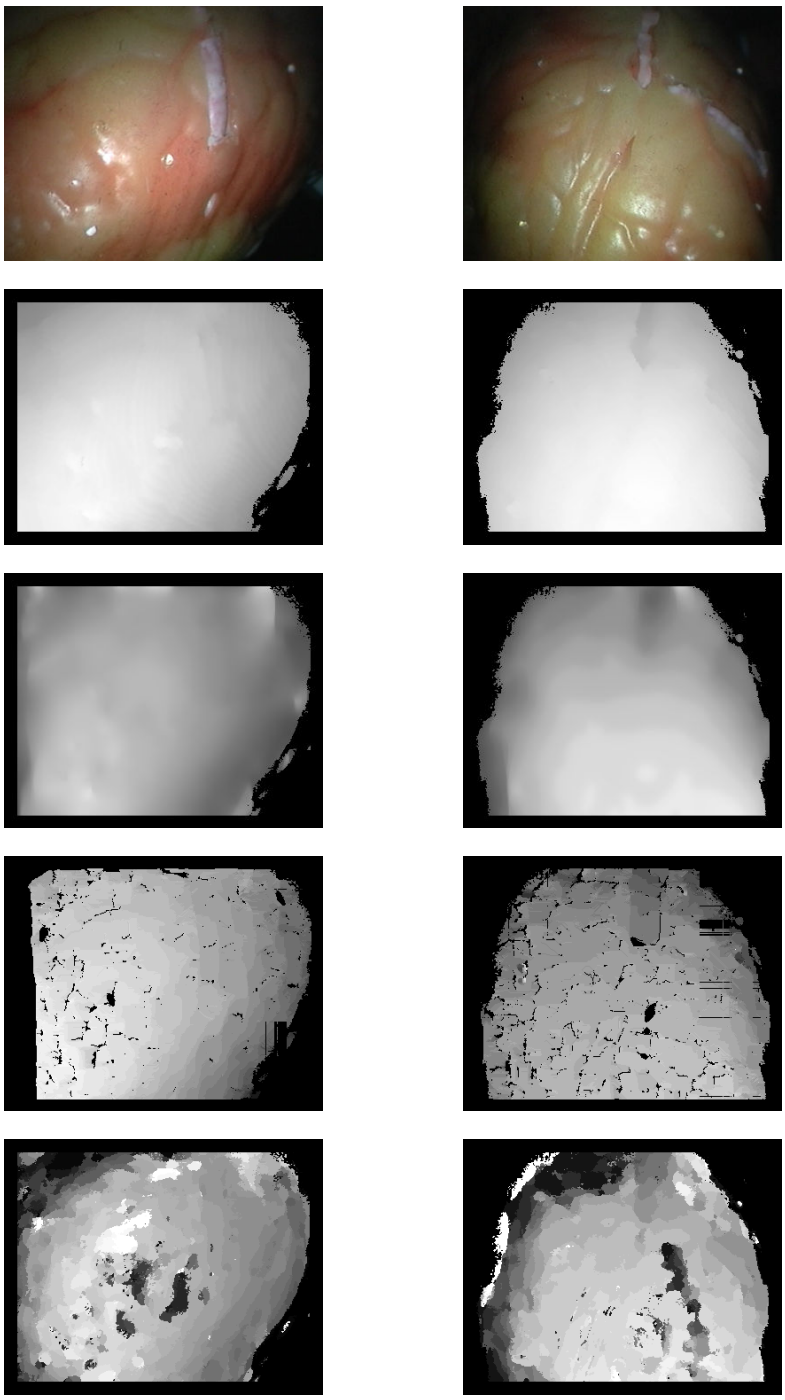


Figure 4.13: The Hamlyn cardiac phantom datasets. Images from top to bottom: Ground truth, the reconstruction result of the proposed approach, SPFP (Stoyanov et al., 2010) and FCVF (Rhemann et al., 2011).

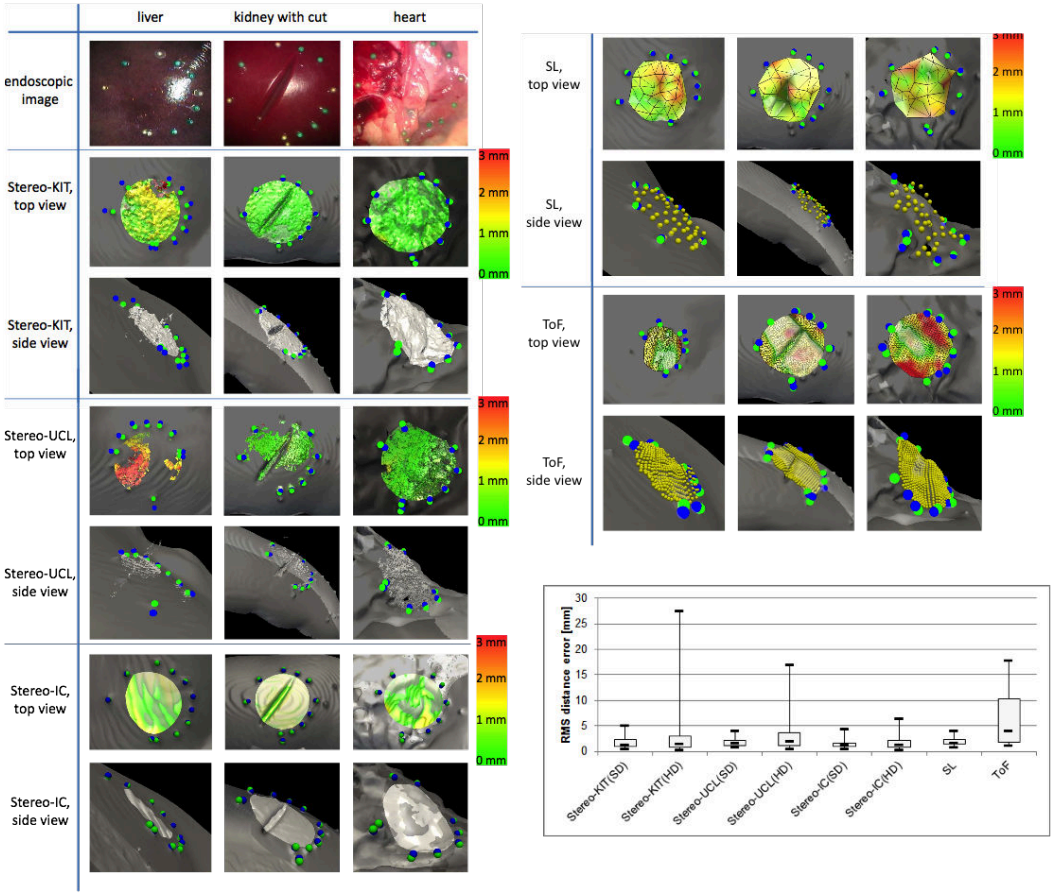


Figure 4.14: Comparisons with different algorithms for the Open-CAS dataset. **Left and top-right**: Reconstruction results obtained by different methods for different shapes including a porcine liver with homogeneous shape and texture, a porcine kidney with a cut and a porcine heart with complex shape and texture. The colours on the reconstructed surface represent the distance to the closest point on the registered reference surface. The large green/blue spheres represented the reference markers and reconstructed markers respectively. **Bottom-right**: Box plot (median, first and third quartiles, minimum and maximum) of the surface reconstruction RMSE for all shapes without blood and smoke for all reconstruction techniques. Stereo-IC has the best performance in both SD and HD image datasets.

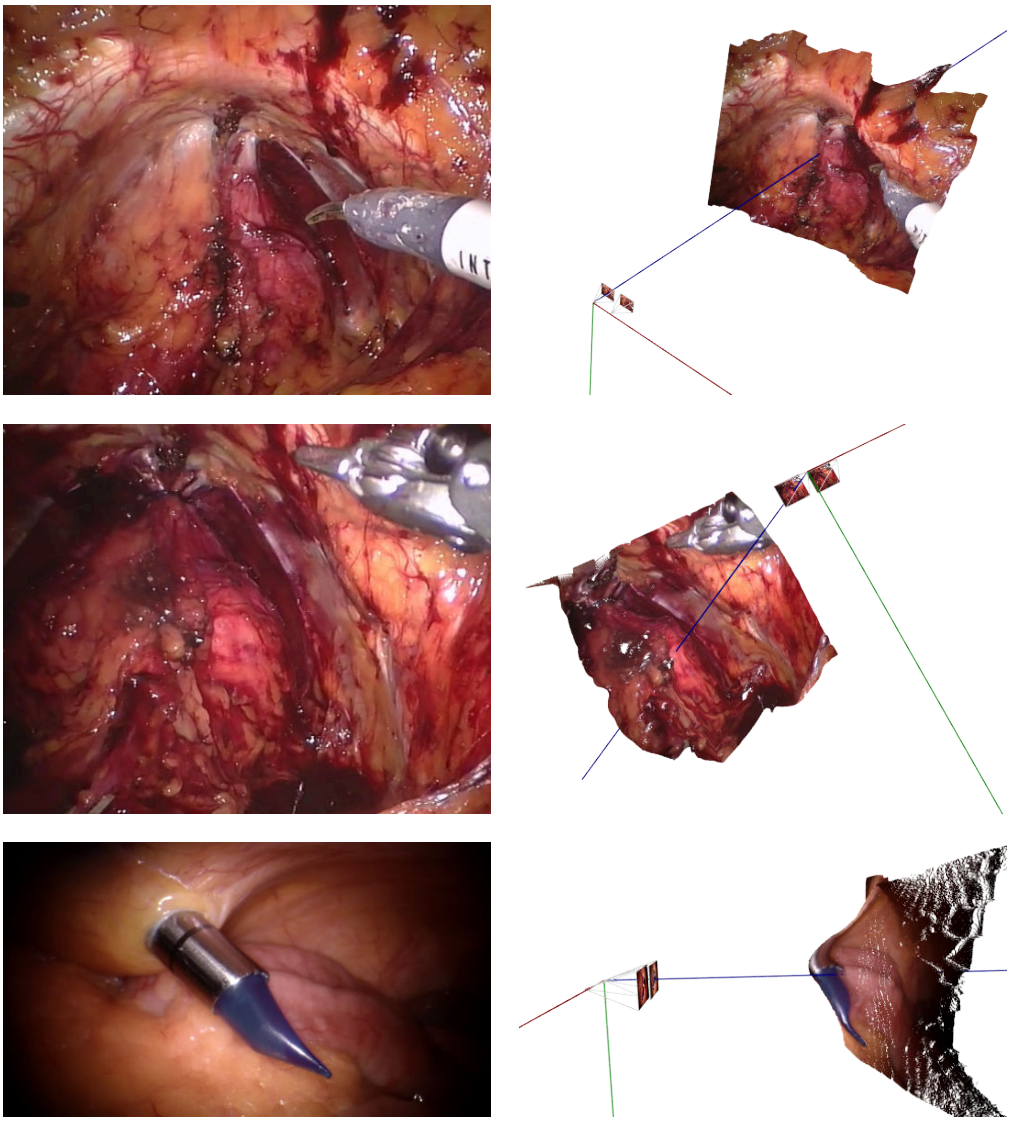


Figure 4.15: Qualitative evaluation of the reconstruction results using the proposed approach. The images are obtained from stereo endoscopic camera in real robot-assisted surgery.

4.6.3 Open-CAS *in vitro* study

Beside the phantom studies, the proposed approach has been involved in a recent joint international initiative work aiming of validating different state-of-the-art passive and active reconstruction methods in a comparative manner (Maier-Hein et al., 2014). The focus is on single-shot techniques such as stereoscopy, SL and ToF which do not require movement of the endoscope. These techniques were tested on the same objects under identical conditions with CT data serving as reference for their reconstruction accuracy, point density, surface coverage and robustness.

Figure 4.14 shows the qualitative and quantitative results of reconstruction accuracy. Here the label Stereo-KIT is proposed by (Rohl et al., 2012), Stereo-UCL by (Stoyanov et al., 2010) and Stereo-IC is our method. SL stands for structured light and ToF for time-of-fly camera. In general among the passive methods, our method has the best reconstruction accuracy and highest surface coverage range. In terms of the active methods, SL has a competitive reconstruction accuracy but very low surface coverage, and ToF suffer from device engineering problems being a prototype of an endoscope. This work has shown that at the moment the passive optical stereoscopic reconstruction algorithm is still the best solution for endoscopic scene reconstruction. See Maier-Hein et al. (2014) for more details.

4.6.4 Prostatectomy *in vivo* study

To our knowledge, there is not yet an *in vivo* dataset providing ground truth for validation. To qualitatively evaluate the performance of the proposed approach on *in vivo* images, endoscopic stereo images from real robot-assisted laparoscopic prostatectomy are reconstructed as shown in Figure 4.15. We have found that the *in vivo* images are very difficult to reconstruct since the blood and soft tissue with homogenous colours cover most of scenes. Severe specular reflection caused by the light source on the endoscopic camera also generates a considerable number of mismatching pixels. We therefore use a larger ZNCC window size 15×15 and set the disparity range $\mathcal{D} = [20, 50]$ and the Huber-ROF parameters $\{\epsilon, \alpha, \beta, \lambda, n_{max}\} = \{0.001, 10.0, 1.0, 0.1, 200\}$. As a result, the general geometries can seem well captured but building *in vivo* ground truth datasets for comparison is still needed, and this is considered as our future work. For better 3D perception of the reconstruction result please watch the accompanying video[5].

[5] http://youtu.be/jqfsv-G7of0

4.7 Conclusions

In this chapter, we have proposed an efficient and effective dense stereo reconstruction method using convex optimisation with a ZNCC cost-volume. Empirical studies have shown that the proposed method outperforms the current state of the art for endoscopic images and can also run in real-time on the GPU. This is a significant advancement towards improved vision-based tracking of the endoscope and is an important step towards providing image guidance to endoscopic procedures. In the next chapter, we will show a dense stereo camera tracking technique which can enable model fusion by using a sequence of video images. This will further improve the reconstructed model and provide a more advanced means of tackling the occlusion at instrument-tissue boundaries.

5 Real-Time Dense Stereo Camera Tracking

Contents

5.1 Introduction

In Chapter 4, we have seen that how a stereo reconstruction algorithm can provide instant rich 3D geometry for an observed endoscopic scene. The reconstructed dense model is useful in several aspects. This chapter focuses on how we can employ the reconstructed model to perform robust visual odometry in order to realise augmented reality in image-guided surgery. Visual tracking in endoscopic scenes is known to be a difficult task due to the lack of textures, tissue deformation and specular reflection. With the availability of a reconstructed model, we propose a real-time visual odometry framework to robustly track the 6-DoF stereo laparoscope pose using the quadrifocal relationship based on an early work by Comport et al. (2010) who used the quadrifocal constraint (Hartley and Zisserman, 2004) for stereo camera tracking. To the best of our knowledge, this work is the first trial of camera tracking using a dense direct approach in endoscopic scenes.

The main is that the instant motion of a stereo camera creates four views which are constrained by quadrifocal geometry. The current pair can be warped back to a reference stereo pair by minimising a photometric error function parameterised by the camera pose and constrained by the quadrifocal geometry. Using a robust estimator can further remove the outliers caused by occlusion, deformation and specular highlights during the optimisation. Since the optimisation uses all pixel data in the images, it results in a very robust pose estimation even for a textureless scene. Furthermore, the proposed tracking algorithm can be efficiently parallelised and run on the GPU. We evaluated the method using a synthetic sequence with a known ground truth model and we also demonstrate the accuracy and robustness of the approach using phantom and real examples of endoscopic augmented reality. This chapter is mainly the work published in IPCAI 2014 (Chang et al., 2014).

5.2 Problem statement

Following the same notations for the image functions defined in Section 4.2, to represent stereo data at an instant timestamp, it is convenient to stack the left and right functions to form a vector such as $\boldsymbol{\mathcal{I}} = (\mathcal{I}_l, \mathcal{I}_r)^\top$ for the image functions. The stereo disparity functions can be represented in a similar way, i.e., $\boldsymbol{\mathcal{D}} = (\mathcal{D}_l, \mathcal{D}_r)^\top$ which also implicitly defines the correspondence point set $\boldsymbol{\mathcal{P}} = (\mathbf{p}_l, \mathbf{p}_r)^\top$.

Given a reference pair $\boldsymbol{\mathcal{I}}^*$ and its disparity map pair $\boldsymbol{\mathcal{D}}^*$, we track the camera by continuously registering the current frame pair $\boldsymbol{\mathcal{I}}$ with the reference pair using quadrifocal warping $w(\boldsymbol{\mathcal{P}}^*, \mathbf{T}_{rl}, \mathbf{K}_l, \mathbf{K}_r; \mathbf{T}(\mathbf{x}))$. Such a generative model is parame-

terised by the camera motion $\mathring{\mathbf{x}}$. We assume that the stereo laparoscope is calibrated in advance and the intrinsic matrices $\mathbf{K}_l$, $\mathbf{K}_r$ and the extrinsic matrix $\mathbf{T}_{rl}$ are constant. For keyframe-based tracking, the task is to estimate the current camera pose $\mathring{\mathbf{x}}$ for the reference frame to the consecutive current frames in real-time and reassign new keyframes whenever necessary.

5.3 Background

Visual odometry is the process of determining the position and orientation of a camera moving in 3D space using only the associated image data. In minimally invasive surgery (MIS), visual odometry is an element of surgical vision that enables endoscope/laparoscope tracking without additional hardware such as optical or electromagnetic trackers (Stoyanov, 2012). Such tracking is crucial for image-guided surgery because the accuracy of camera tracking dominates the stability of applications such as registration of a preoperative model to the surgical site (Chang et al., 2012) or building a mosaic for dynamic view expansion (Totz et al., 2011). By using a visual odometry approach it is possible to overcome the hand-eye calibration and to reduce error propagation while simplifying clinical translation.

5.3.1 Endoscopic camera tracking

Camera tracking based on photometrics in endoscopic scenes is difficult because of the homogeneous appearance of certain tissues, tissue deformation and severe specularities caused by the strong illumination intensity. Previous approaches have adopted a sparse feature-based simultaneous localisation and mapping (SLAM) approach to stereo laparoscope tracking (Mountney et al., 2006; Grasa et al., 2011). In such systems, salient features build a long-term map in order to globally correct for camera drift. They are, however, severely affected by large highlights and lack of scene rigidity.

5.3.2 Dense methods for endoscopic scenes

Recent dense approaches have shown promising results where the camera tracking benefits from the entire image data resulting in a very robust motion estimation even without post pose optimisation in a texture-poor or occluded scene (Newcombe et al., 2011b,a; Comport et al., 2010; Gao et al., 2013). To the best of our knowledge, there has been no previous proposal for dense stereo camera tracking in endoscopic scenes. In Chapter 3, we have shown that the dense tracking is extremely accurate (tracked

camera position with 0.33 mm average error) in a simulated environment where a perfect textured 3D model is provided.

5.4 Proposed approach

We propose a dense approach for real-time stereo laparoscope tracking. Our method uses a combination of stereo reconstruction (Chang et al., 2013), which is effective at recovering snapshots of the surgical site geometry (Maier-Hein et al., 2013), and quadrifocal tracking (Comport et al., 2010; Hartley and Zisserman, 2004). Benefiting from recent GPU technology and parallelisable optimisation algorithms, the proposed dense visual odometry can reach real-time performance. We validate the proposed approach with a ground truth study using a photo realistic rendering of a surgical scene. We also demonstrate the robustness of the tracking on a real phantom video as well as *in vivo* clinical MIS sequences.

5.4.1 Dense stereo reconstruction

The task of stereo reconstruction is to provide the disparity map $\mathcal{D}$ in order to establish the point correspondence $\boldsymbol{\mathcal{P}}$ across the stereo pair. To obtain the disparity map $\mathcal{D}$, we use the approach proposed in Chapter 4, which solves a variational function for left and right disparity map. To restate the method, a disparity map $\mathring{\mathcal{D}}$ is optimised by:

$$\begin{aligned} \mathring{\mathcal{D}} &= \arg\min_{\mathcal{D}} E_{\mathcal{D}}, \quad \text{where} \\ E_{\mathcal{D}} &= \sum_{\mathbf{p}\in\boldsymbol{\Omega}_I} \|\mathcal{F}(\mathbf{p})\nabla\mathcal{D}(\mathbf{p})\|_{\varepsilon} + \lambda C_{\text{ZNCC}}(\mathbf{p}, \mathcal{D}(\mathbf{p})). \end{aligned} \tag{5.1}$$

The details and the solver of Equation 5.1 are referred to Section 4.4.2. The two disparity maps in the pair $\boldsymbol{\mathcal{D}}^*$ are obtained by performing the stereo reconstruction on left and right image respectively.

5.4.2 Dense stereo camera tracking

Given the reference disparity map pair $\boldsymbol{\mathcal{D}}^*$, the registration warping with respect to the camera motion $\mathbf{x}$ can be obtained by optimising the photometric energy function:

$$\mathring{\mathbf{x}} = \arg\min_{\mathbf{x}} E_{\mathbf{x}}, \quad \text{where}$$
$$E_{\mathbf{x}} = \frac{1}{2} \sum_{\boldsymbol{\mathcal{P}}^* \in \boldsymbol{\mathcal{R}}^*} \Big(\boldsymbol{\mathcal{I}}(w(\boldsymbol{\mathcal{P}}^*; \mathbf{T}(\mathbf{x})\hat{\mathbf{T}})) - \boldsymbol{\mathcal{I}}^*(\boldsymbol{\mathcal{P}}^*) \Big)^2. \tag{5.2}$$

All the corresponding pixels from the reference frame pair form the set $\boldsymbol{\mathcal{R}}^* = \{\{\mathbf{p}_l^*, \mathbf{p}_r^*\}_1, \{\mathbf{p}_l^*, \mathbf{p}_r^*\}_2, \ldots, \{\mathbf{p}_l^*, \mathbf{p}_r^*\}_n\}$ which mutually includes the left and right matching pair with a total of n correspondences used for tracking. Note that the $\mathbb{SE}(3)$ transformation matrix $\mathbf{T}(\mathbf{x})$ is applied to the current pose $\hat{\mathbf{T}}$ at the centralised pose $\mathbf{T}_c$ of the stereo-rig baseline as shown in Figure 5.1. This establishes a canonical coordinate for the stereo geometry, in which the left and right camera poses can be obtained via

$$\mathbf{T}_c = \exp^{\log(\mathbf{T}_{rl})/2}, \quad \mathbf{T}_l = \mathbf{T}_c^{-1} \quad \text{and} \quad \mathbf{T}_r = \mathbf{T}_c \mathbf{T}_{rl}^{-1}. \tag{5.3}$$

Quadrifocal geometry

To maximally exploit the stereo image data for tracking, we adopt the quadrifocal geometry as a constraint for associating points across the four views. However, instead of tackling the rather complicated quadrifocal tensor (Hartley and Zisserman, 2004), two trifocal tensors are decoupled from the four-view in order to bring the quadrifocal constraint into the joint optimisation (Comport et al., 2010). Figure 5.1 shows an example of the trifocal geometry for the left setup. Note that here we elaborate only the left trifocal setup for conciseness.

A trifocal tensor $\mathcal{T} = [\mathcal{T}_1(\mathbf{x}), \mathcal{T}_2(\mathbf{x}), \mathcal{T}_3(\mathbf{x})]$ is a $3 \times 3 \times 3$ matrix. Each slice in the tensor is defined by $\mathcal{T}_j = \mathbf{a}_j \mathbf{b}_4^\top(\mathbf{x}) - \mathbf{a}_4 \mathbf{b}_j^\top(\mathbf{x})$ where $\mathbf{a}_j$ are the columns of $\mathbf{T}_{rl}$ and $\mathbf{b}_j(\mathbf{x})$ are the columns of the motion matrix $\mathbf{T}(\mathbf{x})$. We use the point-line-point configuration in which the line $\mathbf{l}_r = [-1, -1, u+v]^\top$ that corresponds to each of the three tensor slices forms the columns of a homography matrix:

$$\mathcal{H}(\mathbf{x}) = [\mathcal{H}_1(\mathbf{x}), \mathcal{H}_2(\mathbf{x}), \mathcal{H}_3(\mathbf{x})] \quad \text{and} \quad \mathcal{H}_j(\mathbf{x}) = \mathcal{T}_j^\top(\mathbf{x}) \mathbf{K}_r^{-1} \mathbf{l}_r.$$

The corresponding point $\mathbf{p}_l$ in the current image can be simply obtained by the homography transformation of the reference point $\mathbf{p}_l^*$. We can now define the warping function in Equation 5.2 for each correspondence as:

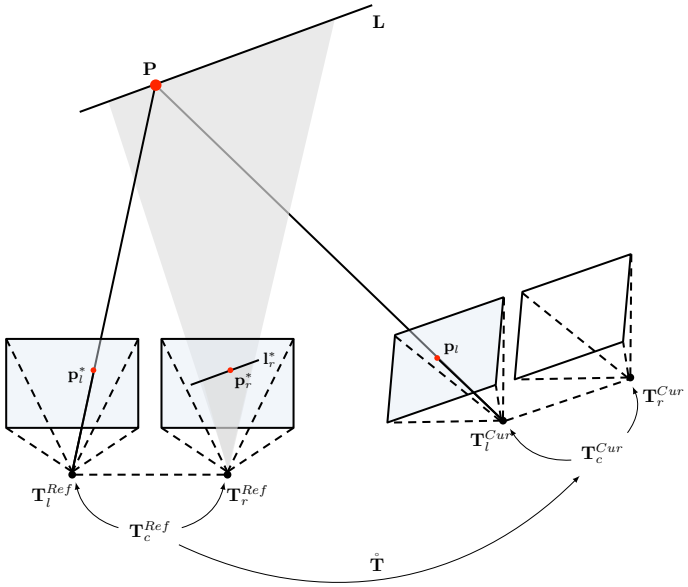


Figure 5.1: Point-line-point trifocal geometry: The point $\mathbf{p}_l^*$ in the left reference frame is transformed to the point $\mathbf{p}_l$ in the left current frame using the homography formed by back-projecting the corresponding line $\mathbf{l}_r^*$, which defines an incidence relation $\mathbf{p}_l^* \leftrightarrow \mathbf{l}_r^* \leftrightarrow \mathbf{p}_l$.

$$\mathbf{p}_l = w(\mathbf{p}_l^*; \mathbf{x}) = \pi\Big(\mathbf{K}_l \mathcal{H}(\mathbf{x}) \mathbf{K}_l^{-1} \dot{\mathbf{p}}_l^*\Big), \tag{5.4}$$

where $\dot{\mathbf{p}}_l^* = [\mathbf{p}_l^*, 1]^\top$ and π is the dehomogenisation function projecting a point to its image coordinate.

5.5 Parametric optimisation

We adopt efficient second-order minimisation (ESM) algorithm together with robust estimation and multi-resolution scheme to optimise Equation 5.2. ESM is mainly the combination of a forward compositional algorithm (FCA) and an inverse compositional algorithm (ICA), which can avoid local minima and takes fewer iterations to converge (Malis, 2004; Baker and Matthews, 2004). The optimisation of quadrifocal warping can be framed using ESM due to the fact that the warping is simply a homography transformation in which the warped current images and the reference images have a linear relationship. Dense tracking by warping a 2.5D surface projection image has no such property, since the projection is non-linear. In such a case we can only adopt the first-order forward compositional algorithm (Newcombe et al., 2011b; Baker et al., 2004). The optimisation scheme is elaborated and derived in

details in this section.

5.5.1 Photometric cost function

The $E_{\mathbf{x}}$ in Equation 5.2 is a parametric cost function which takes the motion parameter $\mathbf{x}$ to map a cost value. In the context of tracking, the cost function is designed that the photometric cost is minimum when the consecutive video frames are best matching between the current and reference frame pair with the relative motion parameters $\mathbf{x}$.

For each pixel correspondence in the reference frame pair $\forall \boldsymbol{\mathcal{P}}^* \in \boldsymbol{\mathcal{R}}^*$, we stack their photometric cost to form a vector $f(\mathbf{x}) \in \mathbb{R}^n$ where $n = |\boldsymbol{\mathcal{R}}^*|$. Equation 5.2 can then be reformed as

$$E_{\mathbf{x}} = \frac{1}{2}\|f(\mathbf{x})\|^2 = \frac{1}{2} f(\mathbf{x})^\top f(\mathbf{x}). \tag{5.5}$$

The property of $f(\mathbf{x})$ is of particular interesting since the cost function has to have a local minimum in order to make the optimisation with respect to $\mathbf{x}$ meaningful.

One may verify this by warping a reference frame pair in a local quadrifocal region and observing the corresponding cost values. Figure 5.2 shows $E_{\mathbf{x}}$ around the identity quadrifocal warping of the reference frame pair, i.e., $f(\mathbf{x}) = \boldsymbol{\mathcal{I}}^*(w(\boldsymbol{\mathcal{P}}^*; \mathbf{T}(\mathbf{x}))) - \mathcal{I}^*(\boldsymbol{\mathcal{P}}^*)$ where $\mathbf{x} \to \mathbf{0}$ for two synthetic images. Using the synthetic scenes allows us to have perfect point correspondences $\boldsymbol{\mathcal{P}}^*$ so that we can fundamentally study the designed cost function. It is apparent that at the identity pose, there is a global minimum, i.e., $E_{\mathbf{x}} = 0$. Within a certain warping range the photometric cost function is generally smooth. Note that there are also a number of local minima.

To find the minimiser $\mathring{\mathbf{x}} = \arg\min_{\mathbf{x}} E_{\mathbf{x}}$, an intuitive approach is exhaustive search. However in the context of 6-DoF camera tracking, the search space is growing rapidly with respect to the discretisation sample size s, i.e. $O(s^6)$, and consequently reaches a unaffordable computation cost. Therefore an efficient and effective algorithm that converges to $\mathring{\mathbf{x}}$ is desirable.

5.5.2 Optimisation with second-order linear approximation

Having observed that the cost function $E_{\mathbf{x}}$ is potentially smooth, we can approach the true motion $\mathring{\mathbf{x}}$ by an incremental update scheme in which the derivatives of $E_{\mathbf{x}}$ with respect to $\mathbf{x}$ at a current estimate is computed and the algorithm moves $\mathbf{x}$ in

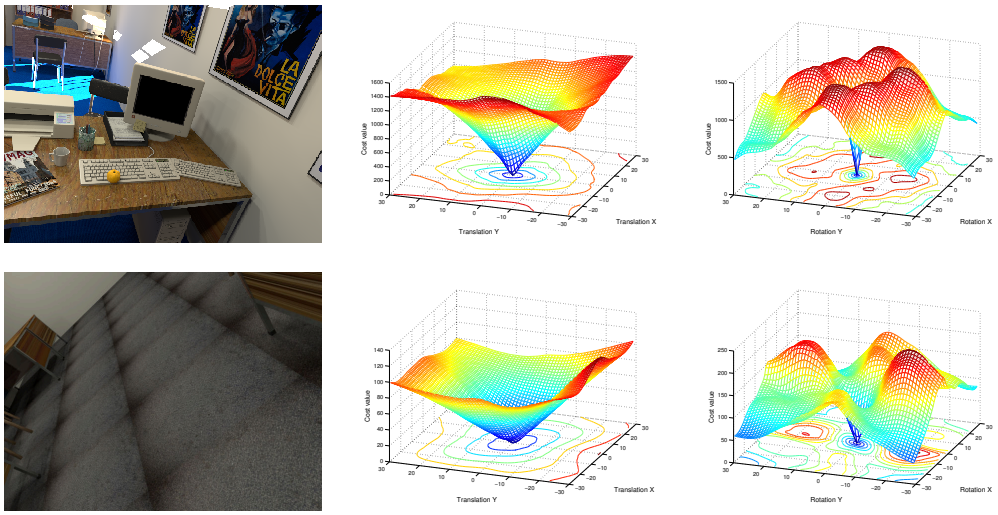


Figure 5.2: Photometric cost function space under quadrifocal warping. Note that in each example the cost function is smooth around the identity pose and there are a number of local minima existing.

downhill linear steps as proposed by Lucas and Kanade (1981). Specifically, in each iteration the algorithm estimates an update $\hat{\mathbf{x}}$ and the corresponding transformation $\hat{\mathbf{T}} \leftarrow \mathbf{T}(\hat{\mathbf{x}})\hat{\mathbf{T}}$ for the quadrifocal warping, until it converges. It is assumed that the true motion parameter $\mathring{\mathbf{x}}$ exists so that $\exists\mathring{\mathbf{x}} : \mathbf{T}(\mathring{\mathbf{x}})\hat{\mathbf{T}} = \mathring{\mathbf{T}}$ (see Figure 5.1). The gradient descent scheme makes the important assumption that the starting estimate is within the convex basin of the global minimum $\mathring{\mathbf{x}}$ otherwise we may get stuck in a local minimum.

In order to apply gradient-descent-like algorithms for the optimisation, the basic photometric cost function $f(\mathbf{x})$ is locally linearised by a Taylor series expansion at $\mathbf{x} = \mathbf{0}$:

$$\hat{f}(\mathbf{x}) = f(\mathbf{0}) + \nabla f(\mathbf{0})\mathbf{x} + \mathcal{O}(\|\mathbf{x}\|^2), \tag{5.6}$$

where the Jacobian $\nabla f(\cdot) \in \mathbb{R}^{n\times 6}$ and $\mathcal{O}(\|\mathbf{x}\|^2)$ is the the second-order remainder as the approximation error. By replacing the $f(\mathbf{x})$ in Equation 5.5 with its approximation in Equation 5.6 and differentiating the photometric cost function for extremum, we arrive at

$$\nabla E_{\mathbf{x}} \approx \nabla \hat{E}_{\mathbf{x}} = \nabla \hat{f}(\mathbf{x})^{\top} \hat{f}(\mathbf{x}) = \mathbf{0}. \tag{5.7}$$

By expanding $\hat{f}(\mathbf{x})$ and evaluating for the approximate local update $\hat{\mathbf{x}}$

$$\nabla \hat{f}(\hat{\mathbf{x}})^\top (f(\mathbf{0}) + \nabla f(\mathbf{0})\hat{\mathbf{x}}) = \mathbf{0}, \tag{5.8}$$

which shows the relation of an over-determined system $\nabla f(\mathbf{0})\hat{\mathbf{x}} = -f(\mathbf{0})$. We can thus seek for a least-square solution for $\hat{\mathbf{x}}$ with the *normal equation* form:

$$\mathbf{J}^\top \mathbf{J}\hat{\mathbf{x}} = -\mathbf{J}^\top f(\mathbf{0}), \tag{5.9}$$

where $\mathbf{J} = \nabla f(\mathbf{0})$. The least-square solution of $\hat{\mathbf{x}}$ can now be solved by Gauss-Newton or Levenberg-Marquardt (Hartley and Zisserman, 2004, Appendix 6). The approximate Hessian matrix $\mathbf{J}^\top \mathbf{J}$ is in practice more desirable than a true Hessian matrix used in a Newton's method because calculating the second-order derivatives can be expensive. Incrementally updating $\hat{\mathbf{x}}$ by solving the normal equation can achieve quadratic convergence when the vector $f(\mathbf{x})$ is linear on $\mathbf{x}$ (i.e. the cost function is locally quadratic), otherwise the convergence rate is only linear (Malis, 2004). FCA and ICA are essential for solving the Equation 5.9 (Baker and Matthews, 2004).

The photometric least-square minimisation subject to the quadrifocal constraint can be solved even more efficiently. Instead of first-order Taylor expansion, we can approximate the cost function in second-order:

$$\hat{f}(\mathbf{x}) = f(\mathbf{0}) + \nabla f(\mathbf{0})\mathbf{x} + \frac{1}{2}\mathbf{x}^\top \nabla^2 f(\mathbf{0})\mathbf{x} + \mathcal{O}(\|\mathbf{x}\|^3). \tag{5.10}$$

By observing the linear approximation of the Jacobian of the cost function $\nabla f(\mathbf{x})$:

$$\nabla \hat{f}(\mathbf{x}) = \nabla f(\mathbf{0}) + \mathbf{x}^\top \nabla^2 f(\mathbf{0})\mathbf{x} + \mathcal{O}(\|\mathbf{x}\|^2), \tag{5.11}$$

truncating the remainders and replacing $\mathbf{x}^\top \nabla^2 f(\mathbf{0})\mathbf{x}$ in Equation 5.10 with 5.11, we can have a second-order linear approximation but with only first-order Taylor expansion terms:

$$\hat{f}(\mathbf{x}) = f(\mathbf{0}) + \frac{1}{2}(\nabla f(\mathbf{0}) + \nabla f(\mathbf{x}))\mathbf{x}. \tag{5.12}$$

Following Equation 5.7, 5.8 and 5.9, the update $\hat{\mathbf{x}}$ can now be solved with $\mathbf{J} = \frac{1}{2}(\nabla f(\mathbf{0}) + \nabla f(\mathbf{x}))$:

$$\hat{\mathbf{x}} = -(\mathbf{J}^\top \mathbf{J})^{-1} \mathbf{J}^\top f(\mathbf{0}). \tag{5.13}$$

Such a solver using second-order approximation is termed efficient second-order minimisation (ESM).

5.5.3 The importance of using smooth motion parameterisation

Whether the incremental update scheme is effective or not is significantly affected by the linearisation of the cost function $f(\mathbf{x})$ at the identity $\mathbf{x} = \mathbf{0}$. The parameterisations $\mathbf{x}$ of $f(\mathbf{x})$ play a crucial role in making the local neighbourhood around the minimiser $\mathring{\mathbf{x}}$ convex, linear and smooth (see Figure 5.2). This is particularly important for higher dimensional motion such as a 6-DoF 3D rigid-body transformation. However many common parameterisations break the linear assumption. Gimbal lock, for example, is a classic problem when using Euler angles to parameterise 3D rotation[1].

Referring to A, Lie group has a rich theory together with its Lie algebra being locally manifold. The transformation in Lie group is smooth and in Lie algebra is analytically differentiable. Awareness of the importance of using smooth motion parameterisation is increasing in robotics as well as the computer vision community (Drummond and Cipolla, 1999; Malis, 2004; Lin et al., 2009; Ma, 2004). The motion can be smooth due to the fact that each intermediate transform is also within the group. For example, a three-dimensional Special Orthogonal group $\mathbb{SO}(3)$ is a subgroup of Special Euclidean $\mathbb{SE}(3)$. If the initial and final poses are within a Lie group, intermediate transformations along the path will also belong to the group's subgroups. The tangent space of a Lie group, i.e., the Lie algebra, also corresponds to shortest geodesics connecting the two transforms in manifold.

Figure 5.3 shows the difference between the conventional linear interpolation in Euclidean space (i.e., $\mathbf{p}_2(t) = \mathbf{p}_1 + t(\mathbf{T} - \mathbf{I})\mathbf{p}_1$) and the Lie algebraic interpolation (i.e., $\mathbf{p}_2(t) = \exp(t\mathbf{x})\mathbf{p}_1$) for a $\mathbb{SIM}(2)$ two-dimensional similarity transformation case. Note that the conventional linear interpolation will make the transformation non-rigid since it interpolates on only points but the interpolated transformation by using Lie algebra will remain rigid as the transformation is always in the subgroups.

[1] Gimbal lock appears when two gimbals rotate around the same axis, the rotation system loses one degree of freedom.

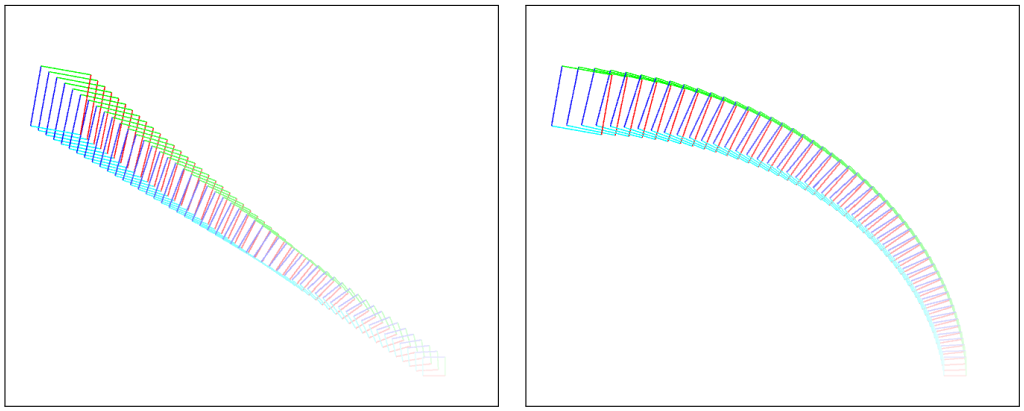


Figure 5.3: A 2D rectangle moves from the bottom-right to the top-left undergoing translation, rotation and scaling transformation in $\mathbb{SIM}(2)$ group. **Left**: Conventional linear interpolation. **Right**: Lie algebraic interpolation. Note that the conventional way simply interpolates the initial and final positions in Euclidean space whereas the Lie algebraic way interpolates rotation, translation and scaling in their local manifold space.

5.5.4 Jacobian derivation

To solve for $\hat{\mathbf{x}}$ in Equation 5.13 using ESM, we have to derive the overall Jacobian:

$$\mathcal{J} = \frac{1}{2}(\mathbf{J} + \mathbf{J}^*), \tag{5.14}$$

where $\mathbf{J} = \nabla f(\mathbf{0})$ and $\mathbf{J}^* = \nabla f(\mathbf{x})$ are the current and reference Jacobian respectively. By applying the chain rule, each Jacobian matrix is derived from the original cost function (Equation 5.2) with respect to the motion of centralised canonical coordinate $\mathbf{x}$:

$$\mathbf{J} = \frac{\partial \boldsymbol{\mathcal{I}}(w(\boldsymbol{\mathcal{P}}^*; \mathbf{T}(\mathbf{x})\hat{\mathbf{T}}))}{\partial \mathbf{x}} = \mathbf{J}_{\boldsymbol{\mathcal{I}}}\mathbf{J}_w\mathbf{J}_{\mathbf{T}}\mathbf{J}_c \quad \text{and} \quad \mathbf{J}^* = \frac{\partial \boldsymbol{\mathcal{I}}^*(\boldsymbol{\mathcal{P}}^*)}{\partial \mathbf{x}} = \mathbf{J}_{\boldsymbol{\mathcal{I}}^*}\mathbf{J}_w\mathbf{J}_{\mathbf{T}}\mathbf{J}_c, \tag{5.15}$$

where

$$\mathbf{J}_{\boldsymbol{\mathcal{I}}} = \frac{\partial \boldsymbol{\mathcal{I}}(\boldsymbol{\mathcal{P}})}{\partial \boldsymbol{\mathcal{P}}}\bigg|_{\boldsymbol{\mathcal{P}}=w(\boldsymbol{\mathcal{P}}^*,\mathbf{T}(\mathbf{0})\hat{\mathbf{T}})}, \quad \mathbf{J}_{\boldsymbol{\mathcal{I}}^*} = \frac{\partial \boldsymbol{\mathcal{I}}^*(\boldsymbol{\mathcal{P}}^*)}{\partial \boldsymbol{\mathcal{P}}^*}, \quad \mathbf{J}_w = \frac{\partial w(\boldsymbol{\mathcal{P}}^*, \mathbf{T}(\mathbf{x})\hat{\mathbf{T}})}{\partial \mathbf{T}(\mathbf{x})}\bigg|_{\mathbf{x}=\{\mathbf{x}_l,\mathbf{x}_r\}=\mathbf{0}},$$

$$\mathbf{J}_{\mathbf{T}} = \left.\frac{\partial \mathbf{T}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\{\mathbf{x}_l,\mathbf{x}_r\}=\mathbf{0}} \quad \text{and} \quad \mathbf{J}_c = \left.\frac{\partial \mathbf{T}_c(\mathbf{x}_l,\mathbf{x}_r)}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{0}}.$$

It turns out that only $\mathbf{J}_{\boldsymbol{\mathcal{I}}}$ has to be updated in each iteration since it is evaluated at the current pose $\mathbf{T}(\mathbf{0})\hat{\mathbf{T}}$, and the rest of Jacobian remains constant.

Derivation of $\mathbf{J}_{\mathcal{I}}$ and $\mathbf{J}_{\mathcal{I}^*}$

The Jacobian of current image $\mathbf{J}_{\boldsymbol{\mathcal{I}}}$ with respect to the point is a 2×4 matrix:

$$\mathbf{J}_{\boldsymbol{\mathcal{I}}} = \begin{pmatrix} \mathbf{J}_{\mathcal{I}_l} & \mathbf{0}_{1\times 2} \\ \mathbf{0}_{1\times 2} & \mathbf{J}_{\mathcal{I}_r} \end{pmatrix}, \tag{5.16}$$

where

$$\mathbf{J}_{\mathcal{I}_l} = \begin{pmatrix} \frac{\partial \mathbf{I}_l(\mathbf{p}_l)}{\partial u_l} & \frac{\partial \mathbf{I}_l(\mathbf{p}_l)}{\partial v_l} \end{pmatrix} \quad \text{and} \quad \mathbf{J}_{\mathcal{I}_r} = \begin{pmatrix} \frac{\partial \mathbf{I}_r(\mathbf{p}_r)}{\partial u_r} & \frac{\partial \mathbf{I}_r(\mathbf{p}_r)}{\partial v_r} \end{pmatrix}.$$

The image gradient is calculated by centred finite difference. The sub-pixels and missing points caused by the warping w is bilinearly interpolated. The Jacobian of reference image $\mathbf{J}_{\boldsymbol{\mathcal{I}}^*}$ can be derived similarly but with constant point in the reference image coordinate.

Derivation of $\mathbf{J}_w$

The $\mathbf{J}_w$ is a 4×32 matrix composed of the derivatives of a projection function and two left (Equation 5.4) and right trifocal function with respect to the $\mathbf{T}(\mathbf{x}) \in \mathbb{SE}(3)$ at the identity $\mathbf{x}_l = \mathbf{x}_r = \mathbf{0}$:

$$\mathbf{J}_w = \begin{pmatrix} \mathbf{J}_{w_l} & \mathbf{0}_{2\times 16} \\ \mathbf{0}_{2\times 16} & \mathbf{J}_{w_r} \end{pmatrix}, \tag{5.17}$$

where

$$\mathbf{J}_{w_{\{l,r\}}} = \left.\frac{\partial \pi(\mathcal{Z}_{\{l,r\}})}{\partial \mathcal{Z}_{\{l,r\}}} \frac{\partial \mathcal{Z}_{\{l,r\}}}{\partial \mathbf{T}(\mathbf{x})}\right|_{\mathbf{x}=\{\mathbf{x}_l,\mathbf{x}_r\}=\mathbf{0}},$$

$$\mathcal{Z}_l = \mathbf{K}_l \mathcal{H}_l(\mathbf{x}_l)\mathbf{K}_l^{-1}\dot{\mathbf{p}}_l^* \quad \text{and} \quad \mathcal{Z}_r = \mathbf{K}_r \mathcal{H}_r(\mathbf{x}_r)\mathbf{K}_r^{-1}\dot{\mathbf{p}}_r^*.$$

Writing $\mathcal{Z} = [x, y, z]^\top$, the projection function $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ is defined as $\pi(\mathcal{Z}) = \frac{1}{z}[x, y]^\top$ so its derivative with respect to $\mathcal{Z}$ is

$$\frac{\partial \pi(\mathcal{Z})}{\partial \mathcal{Z}} = \frac{1}{z}\begin{pmatrix} 1 & 0 & -\frac{x}{z} \\ 0 & 1 & -\frac{y}{z} \end{pmatrix}.$$

The derivative of the trifocal function with respect to the $\mathbf{T}(\mathbf{0})$ is calculated by unrolling the 4×4 matrix and deriving the gradient at each element:

$$\frac{\partial \mathcal{Z}}{\partial \mathbf{T}(\mathbf{0})} =$$

$$\begin{pmatrix} \frac{\mathrm{d}x}{\mathrm{d}\mathbf{R}_{11}} & \frac{\mathrm{d}x}{\mathrm{d}\mathbf{R}_{21}} & \frac{\mathrm{d}x}{\mathrm{d}\mathbf{R}_{31}} & 0 & \frac{\mathrm{d}x}{\mathrm{d}\mathbf{R}_{12}} & \frac{\mathrm{d}x}{\mathrm{d}\mathbf{R}_{22}} & \frac{\mathrm{d}x}{\mathrm{d}\mathbf{R}_{32}} & 0 & \frac{\mathrm{d}x}{\mathrm{d}\mathbf{R}_{13}} & \frac{\mathrm{d}x}{\mathrm{d}\mathbf{R}_{23}} & \frac{\mathrm{d}x}{\mathrm{d}\mathbf{R}_{33}} & 0 & \frac{\mathrm{d}x}{\mathrm{d}\mathbf{t}_1} & \frac{\mathrm{d}x}{\mathrm{d}\mathbf{t}_2} & \frac{\mathrm{d}x}{\mathrm{d}\mathbf{t}_3} & 0 \\ \frac{\mathrm{d}y}{\mathrm{d}\mathbf{R}_{11}} & \frac{\mathrm{d}y}{\mathrm{d}\mathbf{R}_{21}} & \frac{\mathrm{d}y}{\mathrm{d}\mathbf{R}_{31}} & 0 & \frac{\mathrm{d}y}{\mathrm{d}\mathbf{R}_{12}} & \frac{\mathrm{d}y}{\mathrm{d}\mathbf{R}_{22}} & \frac{\mathrm{d}y}{\mathrm{d}\mathbf{R}_{32}} & 0 & \frac{\mathrm{d}y}{\mathrm{d}\mathbf{R}_{13}} & \frac{\mathrm{d}y}{\mathrm{d}\mathbf{R}_{23}} & \frac{\mathrm{d}y}{\mathrm{d}\mathbf{R}_{33}} & 0 & \frac{\mathrm{d}y}{\mathrm{d}\mathbf{t}_1} & \frac{\mathrm{d}y}{\mathrm{d}\mathbf{t}_2} & \frac{\mathrm{d}y}{\mathrm{d}\mathbf{t}_3} & 0 \\ \frac{\mathrm{d}z}{\mathrm{d}\mathbf{R}_{11}} & \frac{\mathrm{d}z}{\mathrm{d}\mathbf{R}_{21}} & \frac{\mathrm{d}z}{\mathrm{d}\mathbf{R}_{31}} & 0 & \frac{\mathrm{d}z}{\mathrm{d}\mathbf{R}_{12}} & \frac{\mathrm{d}z}{\mathrm{d}\mathbf{R}_{22}} & \frac{\mathrm{d}z}{\mathrm{d}\mathbf{R}_{32}} & 0 & \frac{\mathrm{d}z}{\mathrm{d}\mathbf{R}_{13}} & \frac{\mathrm{d}z}{\mathrm{d}\mathbf{R}_{23}} & \frac{\mathrm{d}z}{\mathrm{d}\mathbf{R}_{33}} & 0 & \frac{\mathrm{d}z}{\mathrm{d}\mathbf{t}_1} & \frac{\mathrm{d}z}{\mathrm{d}\mathbf{t}_2} & \frac{\mathrm{d}z}{\mathrm{d}\mathbf{t}_3} & 0 \end{pmatrix}.$$

Derivation of $\mathbf{J_T}$

The $\mathbf{J_T}$ is a 32×12 matrix in which its top-left and bottom-right block is identical and is composed of the unrolled generator matrices of $\mathbb{SE}(3)$. Specifically the derivative of $\mathbf{T}(\mathbf{x}) \in \mathbb{SE}(3)$ with respect to each dimension of $\mathbf{x} \in \mathfrak{se}(3)$ is

$$\frac{\partial \mathbf{T}(\mathbf{x})}{\partial \mathbf{x}_i} = \frac{\partial}{\partial \mathbf{x}_i}\exp(g(\mathbf{x}))\bigg|_{\mathbf{x}=0} = \mathbf{G}_i \quad \text{and} \quad i = 1, \cdots, 6. \tag{5.18}$$

This can be proved as the 4×4 matrix $\mathbf{G}_i$ are the generators of $\mathbb{SE}(3)$ and the generation function $g : \mathbb{R}^6 \to \mathbb{R}^{4\times4}$ is defined as

$$g(\mathbf{x}) = \sum_{i=1}^{6} \mathbf{x}_i \mathbf{G}_i.$$

See Appendix A for the $\mathbf{G}_i$. The $\mathbf{J_T}$ is therefore

$$\mathbf{J}_{\mathbf{T}} = \begin{pmatrix} \mathbf{g}_1\mathbf{g}_2\mathbf{g}_3\mathbf{g}_4\mathbf{g}_5\mathbf{g}_6 & \mathbf{0}_{16\times 6} \\ \mathbf{0}_{16\times 6} & \mathbf{g}_1\mathbf{g}_2\mathbf{g}_3\mathbf{g}_4\mathbf{g}_5\mathbf{g}_6 \end{pmatrix}, \tag{5.19}$$

where each $\mathbf{g}_i$ vector is the row-major unrolled $\mathbf{G}_i$.

Derivation of $\mathbf{J}_c$

The $\mathbf{J}_c$ is a 12×6 matrix and this last Jacobian transforms the derivatives from the left and right coordinate system to the centralised canonical coordinate system. This establishes a bilinear constraint to the stereo pair of two trifocal tensors to form a quadrifocal tensor. To derive $\mathbf{J}_c$, we have to use the equality of the $\mathbb{SE}(3)$ adjoint map $\mathbf{Adj}_{\mathbf{T}} : \mathbb{R}^6 \to \mathbb{R}^6$, which is a 6×6 matrix:

$$\mathbf{Adj}_{\mathbf{T}} = \begin{pmatrix} \mathbf{R} & [\mathbf{t}]_\times \mathbf{R} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}.$$

The adjoint map is derived from the equality:

$$g(\mathbf{Adj}_{\mathbf{T}} \cdot \mathbf{x}) = \mathbf{T} \cdot g(\mathbf{x}) \cdot \mathbf{T}^{-1} \quad \Rightarrow \quad \exp(g(\mathbf{Adj}_{\mathbf{T}} \cdot \mathbf{x})) = \mathbf{T} \cdot \exp(g(\mathbf{x})) \cdot \mathbf{T}^{-1}, \tag{5.20}$$

in which the proof can be found in Rossman (2002). From the transformation of centralised coordinate (Equation 5.3) we have

$$\begin{aligned} \mathbf{T}_l = \mathbf{T}_l\mathbf{T}_c\mathbf{T}_l^{-1} \quad &\Rightarrow \quad \exp(g(\mathbf{x}_l)) = \mathbf{T}_l \cdot \exp(g(\mathbf{x})) \cdot \mathbf{T}_l^{-1} \quad \text{and} \\ \mathbf{T}_r = \mathbf{T}_r\mathbf{T}_c\mathbf{T}_r^{-1} \quad &\Rightarrow \quad \exp(g(\mathbf{x}_r)) = \mathbf{T}_r \cdot \exp(g(\mathbf{x})) \cdot \mathbf{T}_r^{-1}. \end{aligned}$$

Using the equality of adjoint map in Equation 5.20, we have

$$\begin{aligned} \exp(g(\mathbf{Adj}_{\mathbf{T}_l} \cdot \mathbf{x})) &= \mathbf{T}_l \cdot \exp(g(\mathbf{x})) \cdot \mathbf{T}_l^{-1} \quad \text{and} \\ \exp(g(\mathbf{Adj}_{\mathbf{T}_r} \cdot \mathbf{x})) &= \mathbf{T}_r \cdot \exp(g(\mathbf{x})) \cdot \mathbf{T}_r^{-1}. \end{aligned}$$

Together with Equation 5.18, the Jacobian $\mathbf{J}_c$ is therefore

$$\mathbf{J}_c = \begin{pmatrix} \mathbf{Adj}_{\mathbf{T}_l} \\ \mathbf{Adj}_{\mathbf{T}_r} \end{pmatrix}. \tag{5.21}$$

5.5.5 Robust estimation

The original energy function in Eq. 5.5 is the standard least-square method which assumes the residuals have a zero-mean Gaussian distribution. However, the residual distribution is usually not Gaussian, especially when there are outliers appearing in the scene. For example, occluding objects which do not belong to the original reconstructed model, lighting changes or specularities will generate a considerable number of outliers. In endoscopic scenes, non-rigid objects and self-similar structures can even generate more distraction. These outliers can corrupt the fitting data and will eventually make the tracking failed.

A cost function that can differentiate the contributions of inlier and outlier measurements is desirable. Specifically, the function should by its nature be able to reward inliers and suppress outliers. However, how likely the incoming measurement is an inlier or outlier requires a prior statistical study to be performed in advance. For example, the statistics obtained from the study of natural gradient images shows that the underlying gradient is more likely to have a Laplacian distribution (Huang and Mumford, 1999), so one can model the natural images with a Laplacian function to remove the outliers. This methodology is however costly and requires that we study the targeted application scenes first.

Robust statistics provides another way to make the cost function resilient to outliers. Back in the early days of computer vision Gruen (1985) as a pioneer proposed an adaptive statistical weighting scheme for the least square formulations. Stewart (1999) later provides an excellent review for various robust cost functions that are used in computer vision for parameter estimation. Recent remarkable works including Kerl et al. (2013) and Ren and Reid (2012) used t-distribution and Geman-McClure for modelling photometric residual with RGB-D data.

To show how robust statistics works, denoting $r = f(\mathbf{x})$, the original least-square norm in Equation 5.5 can be written as $E_{\mathbf{x}} = \rho(r) = \frac{1}{2}r^2$. We can instead replace the cost function ρ with other robust norms called *M-estimators* (Zhang, 1997). Table 5.1 shows a few candidates that can be chosen as the alternatives. Since most of M-estimator functions are not strictly convex, instead of solving the ρ function directly we can instead solve an iteratively reweighted least squares (IRLS) problem, which can be derived as

Table 5.1: Robust functions and their influence and weight functions.

Type	$\rho(r)$	$\psi(r)$	$w(r)$
L^2	$\frac{1}{2}r^2$	r	1
L^1	$\lvert r\rvert$	$\text{sign}(r)$	$\frac{1}{\lvert r\rvert}$
Geman-McClure	$\frac{r^2/2}{\sigma+r^2}$	$\frac{r}{(\sigma+r^2)^2}$	$\frac{1}{(\sigma+r^2)^2}$
Tukey $\begin{cases}\lvert r\rvert \leq c \\ \text{else}\end{cases}$	$\begin{cases}\frac{c^2}{6}\left(1-(1-(\frac{r}{c})^2)^3\right) \\ \frac{c^2}{6}\end{cases}$	$\begin{cases}r(1-(\frac{r}{c})^2)^2 \\ 0\end{cases}$	$\begin{cases}(1-(\frac{r}{c})^2)^2 \\ 0\end{cases}$

$$\because \quad \frac{\partial E_{\mathbf{x}}}{\partial \mathbf{x}} = \rho'(r)\frac{\partial r}{\partial \mathbf{x}} \quad \overset{\rho'=\psi}{\Longrightarrow} \quad \frac{\psi(r)}{r} r \frac{\partial r}{\partial \mathbf{x}} \quad \overset{\frac{\psi}{r}=w}{\Longrightarrow} \quad w(r) r \frac{\partial r}{\partial \mathbf{x}}$$

$$\therefore \quad E_{\mathbf{x}} = \frac{1}{2} w(r_{k-1}) r_k^2. \tag{5.22}$$

The function ψ and w are termed the *influence function* and *weight function* respectively (Zhang, 1997; Stewart, 1999). The subscript k indicates the iteration number. The weight $w(r_{k-1})$ should be recomputed after each iteration in order to be assigned to r_k^2 in the next iteration. Figure 5.4 shows the different cost functions ρ and the corresponding influence and weight functions. From the chart we can see that the quadratic L^2 norm has a linear influence function ($\psi(r) = r$) and a constant weight function ($w(r) = 1$). This is why a L^2 norm cannot distinguish inliers or outliers as the weight function is constant everywhere. In contrast, L^1, Geman-McClure and Tukey can suppress the outliers as their weight functions assign smaller weights to the large-error measurements. It is worth noting that the Tukey function in fact completely ignores outliers as the weight function is zero outside the inlier zone.

Using Equation 5.22 together with the derivation of the normal equation in Equation 5.13 and the overall Jacobian in Equation 5.14, the robust ESM optimisation with the IRLS scheme in each iteration is

$$\mathbf{x} = -(\mathcal{J}^\top \mathbf{W} \mathcal{J})^{-1} \mathcal{J}^\top \mathbf{W} r, \tag{5.23}$$

where the $\mathbf{W}$ is a $n \times n$ diagonal weighting matrix determined by the Tukey M-

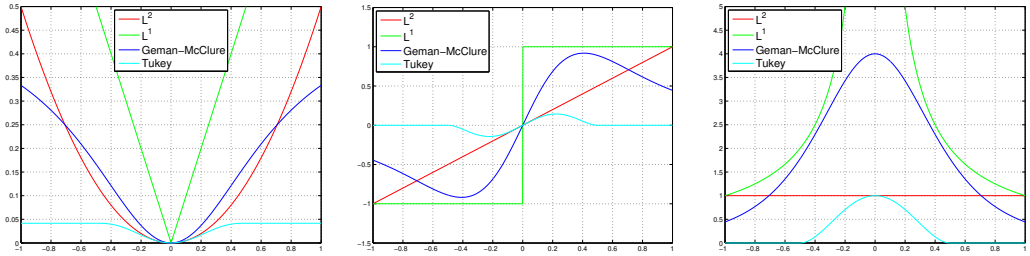


Figure 5.4: Robust functions and their influence and weight functions. **Left**: The cost function ρ. **Middle**: The influence function ψ. **Right**: The weight function w. The Geman-McClure is plot with $\sigma = 0.5$ and the Tukey with $c = 0.5$.

estimator in each iteration.

In practice, the per-pixel weight and Jacobian are calculated just-in-place since all the computation is actually pixel-wise. This makes the GPGPU implementation of the ESM with IRLS for the quadrifocal optimisation very efficient. In practice, we use CUDA shared memory to store the results of the weighted approximate Hessian $\mathcal{J}^\top \mathbf{W} \mathcal{J}$ and the weighted steepest descent $\mathcal{J}^\top \mathbf{W} r$ followed by a *reduction* operation to sum the calculation. The accumulated results are downloaded back to the CPU memory all at once. In this way, we minimised the GPU/CPU memory I/O times. The inverse operation is then performed by direct inverse method on the CPU.

5.5.6 Large displacement motion

Because the Tukey norm is not a strictly convex function, one cannot expect to find the true global minimum. Furthermore the linearisation with respect to the parameters $\mathfrak{se}(3)$ only holds for small camera motions (see Section 5.5.2). To make the method more robust towards rapid camera motions a common coarse-to-fine scheme is adopted.

The coarse-to-fine scheme can be thought as having multi-resolution in the cost function space (Figure 5.2). In the coarser resolution the iteratively incremental optimisation is able to update $\hat{\mathbf{x}}$ with a larger stride, which can potentially avoid possible local minimum that will appear in the finer resolution (Bergen et al., 1992). Figure 5.5 shows an image pyramid where a full resolution 640×480 image is downsampled to 20×15. The hierarchical image pyramid is constructed by Gaussian blurring using the convolution kernel $\mathcal{K} = kk^\top$ where $k = [0.0625, 0.2500, 0.3750, 0.2500, 0.0625]^\top$ followed by a squared sub-sampling (Burt and Adelson, 1983). We can observe that in the coarser images the gradients will be calculated with larger strides then the gradients in the full resolution. Such a larger-

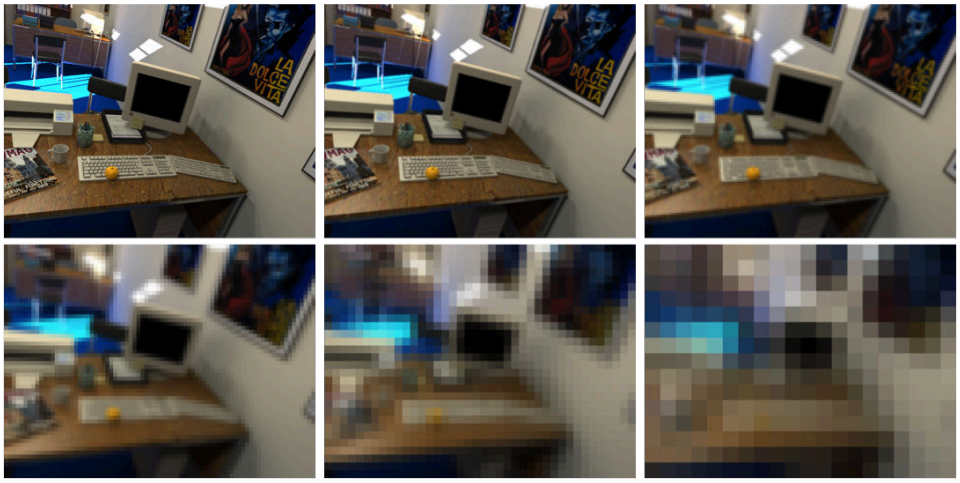


Figure 5.5: Image pyramid constructed by standard downsampling scheme using Gaussian blur and squared sub-sampling from level 0 (640×480) at top-left to level 5 (20×15) at bottom-right.

stride gradient will eventually contribute to the Jacobian chain and as a consequence a large displacement motion $\hat{\mathbf{x}}$ will be updated.

5.5.7 Reference frame selection

The proposed dense stereo visual odometry has the advantage that the reconstruction can be done any time to provide a dense model for the quadrifocal tracking without the need of a bootstrapper. However, reconstructing a model for every frame is unnecessary and in fact frame-to-frame tracking is susceptible to drift. To constrain tracking and prevent drift, we adopt frame-to-model tracking which is essentially the same concept as the keyframe strategy in novel visual SLAM systems (Klein and Murray, 2007; Newcombe et al., 2011b). Therefore, the relatively expensive stereo reconstruction is triggered whenever necessary and this requires a clever system flow design.

The first stereo pair is compulsory to create the reference frame pair. Whether a subsequent stereo frame pair is selected as a new reference frame is based on two criteria: 1) if the overlay between the warping image and the reference image is below a threshold; 2) if the RMSE of Equation 5.2 is larger than a threshold. The first criterion occurs when the scope explores a sufficiently large area of the scene, so that there is not enough of the previous reference model in view. The second criterion can also be associated with the insufficient overlap case but it is additionally useful that when the scene is invaded by other objects or when the scene deformation is large, so we have to immediately reconstruct a new reference model for tracking.

5.6 Empirical studies

We conducted all experiments on a workstation equipped with 3.1 GHz quad cores CPU and one Nvidia GeForce GTX 670 graphics card with 2 GB global memory. The ground truth study is performed on Matlab using only CPU. To maximally exploit the power of parallel computation, for real video sequences, all the calculations including the robust ESM, multi-resolution image pyramid construction are implemented from scratch using CUDA and run on the GPU.

The optimisation for quadrifocal warping using ESM has a quadratic convergence rate. Currently the proposed tracking system is able to run at 25 fps with a stereo pair resolution $|\mathbf{\Omega}_{I_{\{l,r\}}}| = 360 \times 288$. Specifically, using 5-layer image pyramid and the iteration profile shown in Table 5.2, quadrifocal warping optimisation takes around 40 ms.

We conducted ground truth studies which use POV-Ray realistic rendering scenes provided by Handa et al. (2012)[2]. The dataset contains ground truth camera trajectories and depth maps and is designed for the research of visual tracking undergoing rapid and shaky camera motion. On the other hand, the real endoscopic video sequences are acquired from real prostatectomy surgery using da Vinci robotic platform.

5.6.1 Synthetic ground truth study

Two 640×480 synthetic sequences are used for testing the accuracy of the quadrifocal tracking. The first is a synthetic office scene sequence for 50 frames with a very shaky and rapid camera motion. The second one is a synthetic simple slant-floor scene for 20 frames with flying-out camera motion. The contents of these two ground truth sequences are not necessarily challenging but they provide perfect depth models to study the fundamental properties of the proposed quadrifocal tracking algorithm.

For each sequence experiment, the first frame pair is used as the reference for tracking the rest. Figure 5.6 and 5.7 show the tracking result at a current frame pair. Comparing the reference and current images (left view), we can see the displacement is quite large (more than 100 pixels). Such a large displacement requires the incremental update scheme to run more iterations to find the true minimum but this also increases the likelihood of becoming trapped in a local minimum. With the help of the coarse-to-fine scheme, we can see that the warping is mostly done at the coarsest level 5 as shown in Figure 5.6c and 5.7c. This is because the 100-pixel

[2]VaFRIC dataset: http://www.doc.ic.ac.uk/~ahanda/VaFRIC

Table 5.2: The iteration number required by the robust ESM in each pyramid level.

Level	0	1	2	3	4	5
Number of iteration (office scene)	1	2	4	4	5	15
Number of iteration (simple scene)	3	3	4	4	8	14

displacement in the full-resolution image is largely decreased to only $\frac{100}{2^5} \approx 3$-pixel in the coarsest level. As reported in Table 5.2, the average iteration number for ESM convergence in each level can also help to explain the idea. The registration in full-resolution merely provides further refinement so most of the computation is concentrated in the coarser images. This also significantly saves the computation time.

By observing the final warped and error images as shown in Figure 5.6d, 5.7d 5.6e and 5.7e, we can see that there are errors between the warped images and the reference images. The errors are mostly due to object occlusions and lighting conditions. This is where the robust estimation can significant help with. As shown in the Figure 5.6f and 5.7f, the Tukey M-estimator assigns 0-weight to those error-prone outlier pixels. The quadrifocal tracking thus uses only those inlier pixels for registering images. As a consequence, the tracker is extremely robust even with outliers existing. Figure 5.8a and 5.8b show the tracked trajectories compared with the ground truths. For the office and simple scene, the RMSE is only 0.11 mm and 0.02 mm respectively.

5.6.2 Synthetic noise study

In order to study how resilient the dense quadrifocal tracking is with respect to noises and luminance changes, we use POV-Ray for realistic rendering for a bladder and a pelvis phantom model. The luminance is intentionally set as a point light source and materials with strong specularity are used to simulate the real surgical scene where the only light source is at the middle of the endoscopic cameras as shown in Figure 4.1a. Figure 5.9a shows a realistically rendered stereo frames. Following the same methodology as adopted in Handa et al. (2012) and Gao et al. (2013), we perform quadrifocal tracking of the endoscopic camera on the da Vinci robotic platform when viewing a real phantom model as shown in Figure 5.10a to generate a real camera trajectory. We then use this camera trajectory to render the ground truth sequence. The ground truth trajectory is shown in Figure 5.9c.

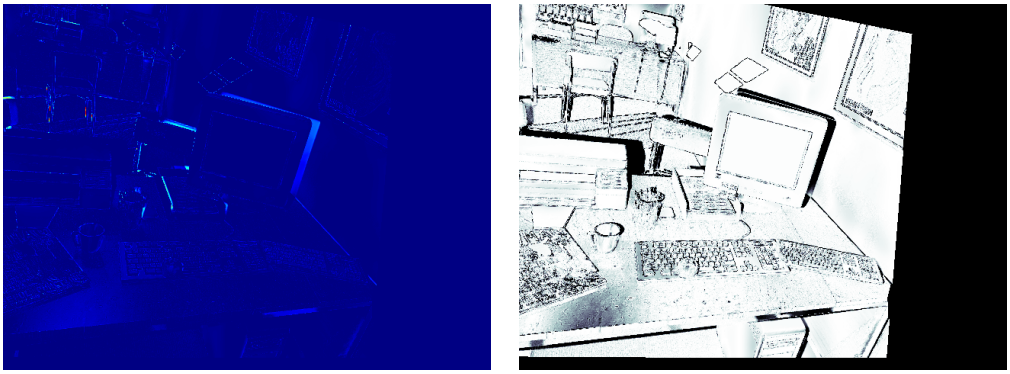
The first frame pair is used as the reference for tracking the rest. The method-

(a) Reference image

(b) Current image

(c) Warped current image at level 5

(d) Warped current image at level 0



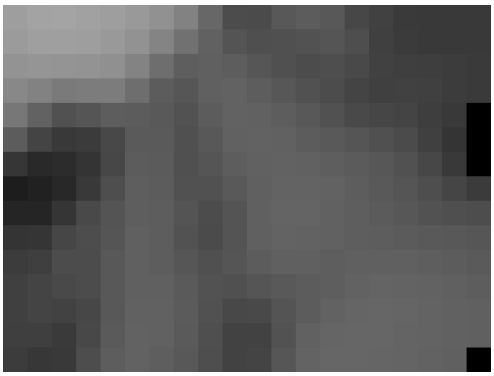
(e) Error image

(f) Tukey weight image

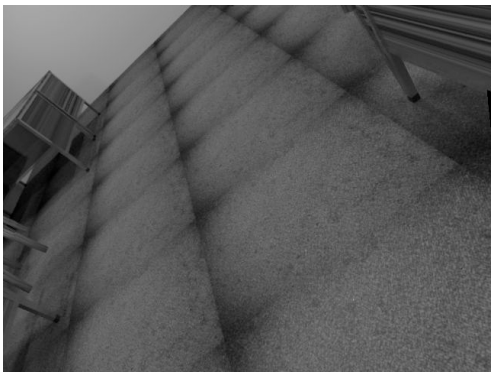
Figure 5.6: Quadrifocal tracking in a synthetic office scene. Note that the artifacts around the monitor, the glasses and the printer in (d) are the occluded areas in the original image (b).

(a) Reference image

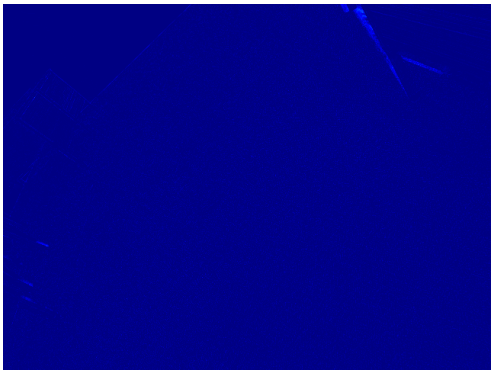
(b) Current image



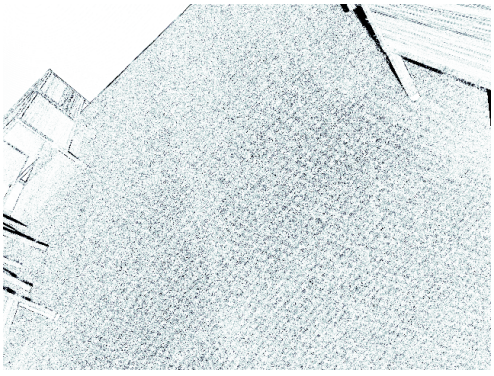
(c) Warped current image at level 5



(d) Warped current image at level 0

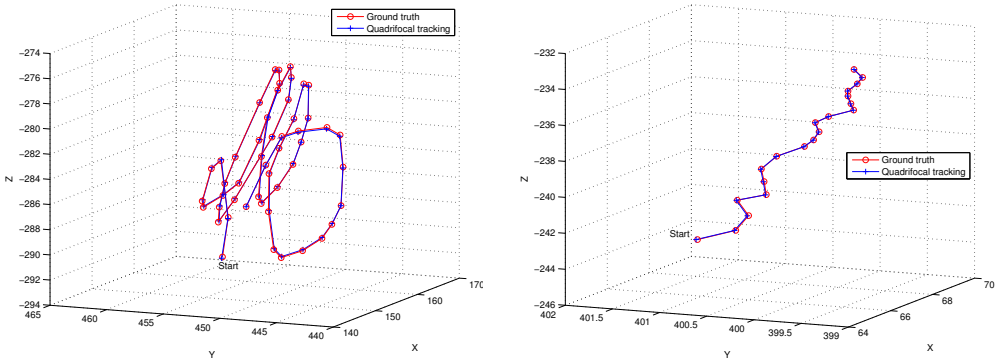


(e) Error image



(f) Tukey weight image

Figure 5.7: Quadrifocal tracking in a synthetic simple scene.



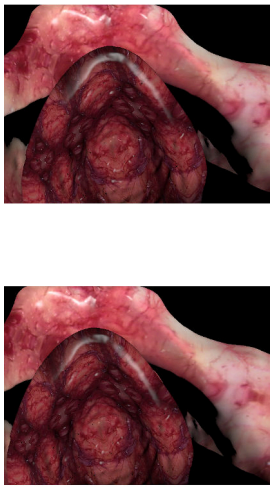
(a) Office scene (b) Simple scene

Figure 5.8: (a) Tracked camera trajectory in the office scene for 50 frames. Compared with the ground truth the RMSE is 0.11 mm. (b) Tracked camera trajectory in the simple scene for 20 frames. Compared with the ground truth the RMSE is 0.02 mm.

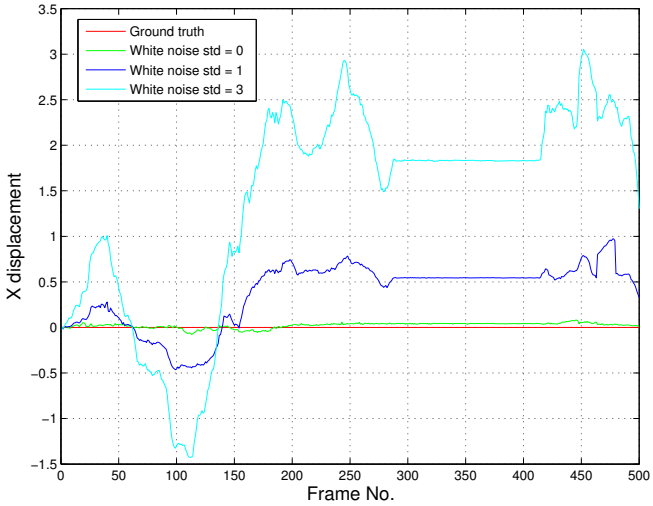
ology for the noise study is to add AWGN to the reference model with different standard deviation and observe how this will affect the tracking. Figure 5.9b shows the tracking errors along the x-axis under different levels of white noise. It reveals several important results. Firstly, as the green curve shows, tracking with a perfect model gives almost no drift even under large luminance changes. But in practice a perfect reconstruction is never achievable. The blue curve is closer to the real situation where we have a decent, but not perfect reconstruction. Due to the imperfect model, the camera drifts about 0.5 mm after tracking for 100 frames. The cyan curve shows that with a very bad reconstructed model, the tracking can still work but with a significant drift.

5.6.3 Real endoscopy study

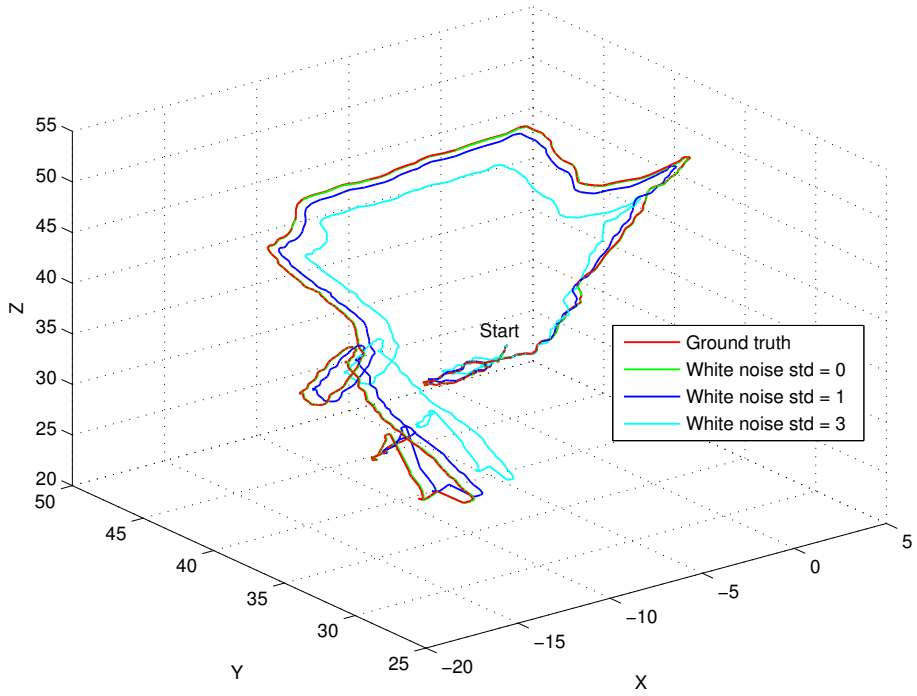
The real video sequences are acquired from da Vinci robot's stereo endoscopy with size 720×576 and downsampled to 360×288. To validate the proposed approach on real data, we use a phantom and a clinical endoscopy sequence to conduct a qualitative evaluation. The phantom is an anatomical pelvis and prostate model from Educational and Scientific Products Limited with added surrounding tissue features made from coloured silicone and outer areas filled with polyeurathane expanding foam to avoid unrealistic sharp edges as shown in Figure 5.10a. Figure 5.10c shows the reconstructed disparity map of the Figure 5.10b where the depth discontinuity around the instrument is preserved. With this well-reconstructed model, when the instrument starts to move, the robust estimator assigns low weight for the tracked



(a)

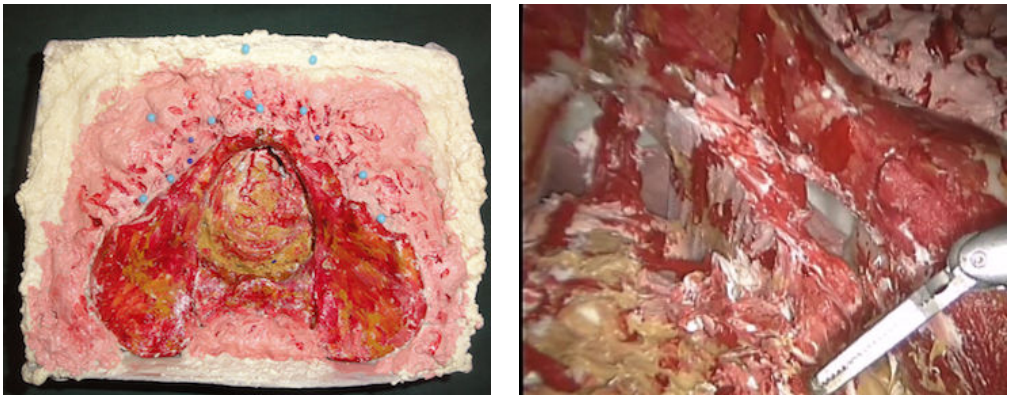


(b)

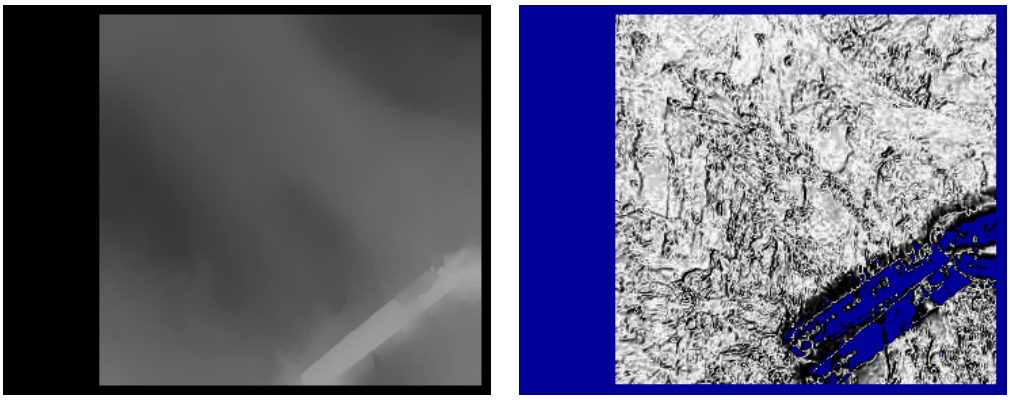


(c)

Figure 5.9: Synthetic noise study. (a) The realistically rendered stereo frames with a pelvis and bladder models. (b) The displacement of the tracked x-translation away from the ground truth. (c) The 3D trajectories.



(a) (b)



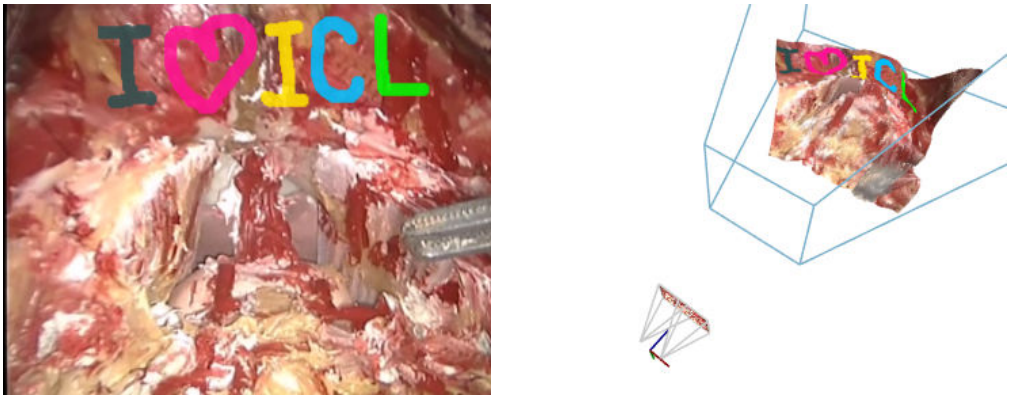
(c) (d)

Figure 5.10: (a) The painted plastic phantom. (b) A viewport from the da Vinci robot's endoscopy. (c) The reconstructed disparity map used for quadrifocal tracking. (d) The Tukey M-estimator weighting image where the blue pixels are rejected and gray pixels from black to white corresponds to the weight value from 0.1 to 1.0.

pixels which do not belong to the model or even completely rejects them, as shown in Figure 5.10d.

The proposed dense approach can be applied to a variety of applications. We demonstrate augmented reality (AR) using the reconstructed dense model. As shown in Figure 5.11, we can draw text on the dense model and maintain their position on the surface. Note that this is not possible for sparse feature approach in which there is no a dense geometry to be drawn on. This method could be useful as it allows surgeons to tag AR annotation in the endoscopic scenes. A promising application for this is the Firefly, which labels cancerous tissue at a given point in time. This method could then maintain alignment as the endoscope is moved.

Another useful function of the robust tracking using a dense model is to detect occlusions. As shown in Figure 5.12c and Figure 5.12e, the dense reference model



(a) (b)

Figure 5.11: Text augmented reality. (a) Drawing text on the 2D image. (b) The text is actually drawn on the reconstructed 3D model.

provides a strong prior to reject the occluding instrument which is judged directly by Tukey's weight shown in Figure 5.12d. When a new reference model is added, the occlusion can be also detected by comparing the depths between the tagged markers and the new model. Note that in Figure 5.12d, those specularities are also rejected for the quadrifocal tracking. For computer assisted image-guided surgery, we can also augment the preoperative models into the endoscopic scene as shown in Figure 5.12f.

The tracking quality can be observed in the supplementary video in which the AR is shown incredibly stable within extremely challenging scenes containing deformation, specularity and instrument occlusion[3].

5.7 Conclusions

In this chapter, we proposed a dense visual odometry method for tracking the motion of the stereo laparoscope in MIS by using the quadrifocal constraint with the stereo reconstructed model using the method proposed in Chapter 4. The dense approach has been shown to achieve promising results for synthetic, phantom and clinical data even in sequences with deformation, heavily specular highlights and instruments occluding the surgical site. Promising applications of the proposed technique include image-guided surgery with AR overlay onto the laparoscopic images.

In our future work we will focus on building a fully dense SLAM system with keyframes refined by pose graph optimisation to accurately maintain a global map

[3] http://youtu.be/pwaf3dd43W0

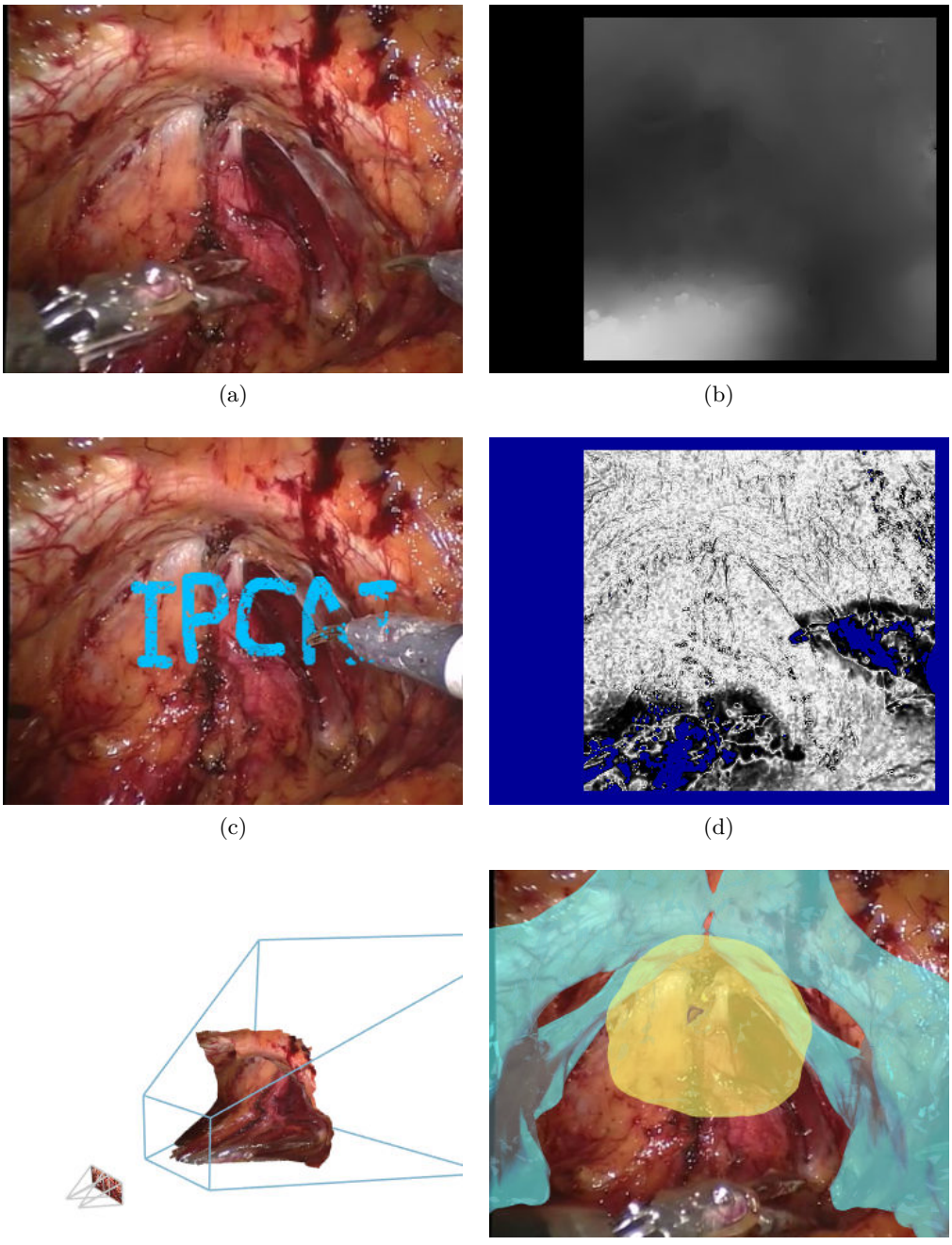


Figure 5.12: AR in real robotic prostatectomy surgery. (a) Real endoscopic image. (b) The reconstructed disparity image of (a). (c) An example for occlusion detection. (d) The Tukey weights of (c) showing that the pixels from the invading instrument together with the specularities are mostly rejected. (e) The reconstructed model used for the quadrifocal tracking and AR. (f) Preoperative models augmented into the *in vivo* endoscopic scene.

while efficiently selecting known keyframes for tracking.

6 Conclusions and future research

Contents

6.1 Conclusions

This thesis has presented a novel pure visual-based stereoscopic camera tracking system for RALP. The system consists of dense methods including dense stereo reconstruction and dense stereo visual odometry components. Several validation experiments have been conducted in each component and the results have shown that the proposed system is extremely robust and can be successfully applied to endoscopic scenes where a feature-based tracking system may prove infeasible.

The final system using dense methods was inspired by the preliminary research in Chapter 3 where we found the state of the art feature-based visual SLAM system, PTAM, failed in tracking the endoscopic camera in the RALP scenes. Without a number of decent tracked camera poses, the original trial of multi-view 3D preoperative model to 2D images registration using colour-consistency ultimately failed.

We also realised that the preoperative models, due to the scan and segmentation errors, will not have exactly the same shape as the tissue geometry observed through the endoscope. This made the colour-consistency registration converge at the wrong pose, which will readily cause the following dense tracking to drift. Nevertheless, through synthetic studies we have shown that the dense tracking by itself is a promising direction to overcome these problems. We therefore abandoned the idea of 3D to 2D image registration. In Chapter 4, instead, we directly reconstruct the 3D geometry of the surgical site using the stereoscopic camera. The proposed reconstruction algorithm has turned out to be more accurate than other previous methods for our endoscopic images.

Having shown that we are able to reconstruct a decent 3D model using the stereoscopic camera, in Chapter 5 we looked back the original camera ego-motion estimation problem again and replace PTAM with the proposed stereo tracking algorithm using the quadrifocal constraint. In this way, we maximally exploit all the pixel data in the stereo images and the resulting camera tracking system is extremely robust even within very challenging scenes such as those that appear in RALP. A number of AR applications were also demonstrated using the proposed tracking system. The work in this thesis has resulted in a novel camera ego-motion estimation system that can be a robust front-end to enable AR applications in IGS. The results shown in the videos (Appendix B) demonstrate that robust tracking can be achieved in difficult circumstances and this is a significant contribution to the field of MIS.

6.2 Discussions

There are many implicit insights thorough this thesis in terms of the mathematical principle of the problems and the implementation of a practical system. They are disclosed in this section to offer hints for future research directions in AR for IGS and MIS.

6.2.1 Revisiting colour consistency

In Chapter 3, the 3D to 2D image registration for aligning a pre-operative model with endoscopic images eventually failed. However in the literature, computer vision algorithms using colour consistency has been proved very successful (Lucas and Kanade, 1981; Horn and Schunck, 1981). In fact with Lambertian lighting assumption (Basri and Jacobs, 2003), colour consistency is theoretically sound and has been used for solving several computer vision problems such as stereo (Lucas and Kanade, 1981; Scharstein and Szeliski, 2002), multi-view stereo (Furukawa and Ponce, 2010; Kolev et al., 2010) and optical flow (Horn and Schunck, 1981; Newcombe and Davison, 2010). Even some recent successful tracking systems are built on the colour consistency assumption (Klein and Murray, 2007; Newcombe and Davison, 2010; Newcombe et al., 2011b; Comport et al., 2010).

When looking into those successful cases using colour consistency, one must be careful that there are other assumptions that must be established as well. Beside Lambertian reflection, scene rigidity (for camera tracking), surface geometry consistency (for 3D surface registration) and motion group assumption (for 2D or 3D image registration) are significant factors that must be considered when modelling a vision problem. The devil in the details which undermine the use of colour consistency for our 3D to 2D registration problem is mainly because the scene is not rigid and the surface geometry of the preoperative model is not consistent with the scene. Within a non-rigid scene, PTAM as a tracking algorithm has failed as a front end and the subsequent 3D to 2D image registration using the inaccurate tracked poses is ultimately doomed. Even if the tracked poses are perfectly accurate, the inconsistent surface geometry caused by manual segmentation from CT/MRI is yet another flaw. Two negatives will not make a positive in this case.

When one devises a vision-based algorithm dealing with complex scenes, the above mentioned assumptions have to be considered. Essentially, when tackling inverse problems in computer vision, one should be sure to solve the optimisation based on sound mathematical models. Only through this means can the optimum

be the true solution for the vision problem. For example, if image A and B are within homography mapping (8-DoF), solving a set of affine parameters (6-DoF) for image A cannot find the true mapping from A to B at all, no matter how brilliant an optimisation algorithm is used. Unless by luck the parameters happen to be coplanar. On the other hand, over-parameterisation can also cause the optimisation to be inefficient and ineffective, such as in the case of the opposite problem - mapping two affine images using homography. Precisely modelling a vision problem with the right mathematical model can significantly ease the overall optimisation.

6.2.2 Real-time stereo

When dealing with deformable scenes, the advantage of using stereo vision is to snapshot an instant rigid scene so that a 3D model can be reconstructed from the stereo image pair. Stereo reconstruction is a well-understood problem in computer vision where a comprehensive literature reviews can be found in Section 4.3. For endoscopic scenes, the main challenge is texture-poor appearance which means the photometric-based correspondent search easily fails. Artificial textures generated by specular highlights make the mis-matching problem even worse. Using a stereo endoscopic camera, a stereo reconstruction algorithm primarily has to tackle this issue.

The proposed stereo reconstruction algorithm in Chapter 4 is built on a successful variational model which uses a patch-based photometric cost function for the data term and a L^1-L^2-hybrid Huber-norm for the regulariser term. Although the patch-based photometric cost can indeed ease the problem of lack of texture, it brings another undesired front-parallel effect to the reconstruction. The Huber-norm regulariser subsequently alleviates this effect, which makes the ultimate reconstruction smooth with respect to the depth.

This is the first proof of the power of using a mathematically sound model in this thesis, which has shown the state of the art performance for 3D surface reconstruction in endoscopic scenes (Maier-Hein et al., 2014). Recent reconstruction algorithms have tended to focus on finding more sophisticated variational models for dealing with non-textured areas, such as a second-order TGV model which can regularise depth for slant structures (Ranftl et al., 2012, 2013) and a SLAM system using super-pixel representation for stereo matching and Manhattan piecewise planar priors (Concha et al., 2014).

However the optimisation of a variational model requires tuning many hyper-parameters. Specifically, the primal-dual optimisation algorithm used in Sec-

tion 4.4.2 needs to be tuned for the disparity range $\mathcal{D}$, regularisation weight λ, edge weight α and β and ε for the Huber-norm, despite the fact that Pock and Chambolle (2011) have proposed an effective algorithm for determining the step size parameters τ and σ. To ease the parameter tuning problem, we rely on a real-time stereo reconstruction system. The system allows us to set those parameters on the fly while showing the reconstruction result instantly. A real-time stereo reconstruction system in this case turns out to be extremely crucial, not only for the later real-time dense tracking but also for tuning the reconstruction result. It is also possible to use machine learning methods to learn these hyper-parameters if we have a number of training datasets with known reference surfaces (Snoek et al., 2012).

6.2.3 Why quadrifocal tracking?

Given that we are able to perform stereo reconstruction to obtain the dense 3D geometry (point cloud), there are several ways to approach dense tracking. One can use iterative closest point (ICP) to align two point clouds (Newcombe et al., 2011a). Alternatively, one can also adopt 2.5D image registration using only one of the stereo camera (Newcombe et al., 2011b; Baker et al., 2004). There are even works using genetic algorithms such as particle swarms for camera pose estimation (Liebelt and Schertler, 2007; Baik et al., 2013). Why are we particularly interested in using the quadrifocal constraint (Hartley and Zisserman, 2004) which is also used by Comport et al. (2007, 2010) for dense tracking?

First of all, the suggestion of ICP is definitely a bad idea. This is mainly due to scene deformation. The point cloud reconstructed from previous timestamp would not necessarily have the same geometry as the current reconstructed point cloud, since the scene has deformed. On the other hand, the two consecutive reconstruction processes are completely independent so there is no way to associate deforming areas from frame to frame, unless we apply extra deformation detection. In addition, reconstructing a 3D point cloud for every single frame is impractically expensive. The successful case of using ICP for dense tracking is built on the fact that the scene is static and we can obtain point cloud with almost zero computation overhead, thanks to novel RGB-D sensors such as Kinect (Newcombe et al., 2011a).

Genetic algorithms or derivative-free algorithms are also out of our concern because we aim for a real-time tracking system. This implicitly means that we need a model which can be optimised by gradient-based algorithms such as gradient-decent or Gauss-Newton. DTAM can use only the 2.5D Lucas-Kanade forward compositional algorithm because intuitively it is a monocular setup (Newcombe et al.,

2011b; Baker and Matthews, 2004). A subtle fact that it cannot exploit the more efficient ESM algorithm is because of the non-linear transformation of the perspective projection (Baker et al., 2004).

However, a stereo setup essentially comes along with the ESM since the quadrifocal warping consists of two bundled trifocal warpings, and the trifocal warping is nothing but a homography transformation which is a linear operation. Using the quadrifocal constraint not only provides the opportunity of using ESM for optimisation but also enables us to associate deforming areas from frame to frame using robust estimation. This is yet another proof of the significance of a sound mathematical model that can hugely influence the practical result. Eventually we can realise a real-time dense tracking system while taking the outliers including deformation, specularity and occlusions into account.

6.2.4 Are dense methods the panacea?

Dense methods are probably not the panacea. The original intention of this thesis is to devise a practical image-guided system particularly for augmented reality applications. What we have found is that the performance of an augmented reality system depends on the quality of camera tracking. Monocular camera tracking using sparse features is limited to texture abundant and static scenes. Dense methods on the other hand can provide more robust tracking in dynamic scenes but require much more computation effort on redundant data such as pixels in textureless and occlusion areas. The debates between sparse feature based methods versus dense methods still remain (Irani and Anandan, 2000; Tri, 2000; Handa, 1999).

Not surprisingly, a more sophisticated tracking system would come up with a leveraging algorithm resorting to a semi-dense paradigm (Engel et al., 2013). Instead of reconstructing everything in the scene, only pixels with gradient, i.e., edges, are used for building the map and tracking the camera. Such a methodology is promising for visual SLAM in endoscopic scenes. For example, vessels may not be detected as salient points by automatic feature extraction algorithms but can still contribute gradients for constructing the scene structure and drive the camera tracking.

6.2.5 Difficulties arising from the endoscopic scenes for visual tracking

The difficulties presented by endoscopic images should not be underestimated. Although we have focused on only RALP in this thesis, other MIS procedures share common issues. The specularities, deforming tissue, smoke and presence of instru-

ments moving over the scene all pose challenging problems for any vision-based pose estimation algorithm.

We do not tackle those problems explicitly in this thesis, but rely on using a huge number of measurements. Incorporating a robust function in dense tracking is crucial. Having a large number of observed data enables a non-linear least square optimisation algorithm to solve for the optimum with consideration of the distribution of data residual, in which the outliers should be removed in order to achieve robustness. This also applies to sparse feature tracking methods where a larger number of correspondences is also preferable (Strasdat et al., 2010). Once the number is large enough to have statistical meaning, a robust function can always come in and play an important role.

Novel GPGPU hardwares enable efficient parallel computing for many computer vision problems especially for the dense paradigm. In this thesis, we show reconstruction and dense tracking problems that can be solved by highly parallelisable algorithms together with efficient CUDA implementation to reach real-time performance, where computation takes too long using the CPU. Dealing with very difficult computer vision problems such as in endoscopic scenes means a promising computer vision algorithm would be necessarily computationally complicated. The popularity of GPU for solving difficult computer vision problems is not only foreseeable but is also a reality in current research.

While we have shown our real-time AR system overcoming certain deformation, smoke, and sever clutter situations (Appoximate B), it does not necessarily mean the proposed system has entirely solved the difficult problems in endoscopic scenes. Once the disturbance (outliers) becomes too large, the tracking will still fail. Mathematical modelling will continue to play an important role. Precisely modelling those endoscopic problems and incorporating the models into the optimisation will be required to produce an ultimately robust AR system.

6.3 Future directions

There are a few intriguing ideas and thoughts that we have not yet explored given the time frame of this thesis. We believe following these research lines, a more sophisticated IGS system can be realised and eventually be used in the theatre in the near future.

6.3.1 Multi-view stereo reconstruction and pose optimisation

Although epipolar geometry can give a strong matching constraint which makes it feasible to estimate dense surface geometry for every pixel in a stereo pair, for texture-poor areas, where the stereo matching is ill-conditioned, the reconstruction results are likely to be inaccurate. In a variational optimisation framework, we can use a regulariser to smooth the poorly matching areas but in the end this is a "fill-in" depth method which may not recover the true geometry.

Given that now we are able to track the camera in the endoscopic scenes, the next step is intuitively to realise multi-view stereo reconstruction (Newcombe et al., 2011b). The multiple camera views should effectively increase the width of the stereo baseline. It is hope that by fusing more stereo frames, the homogeneous-texture areas can have a better reconstruction. This is promising because the epipolar lines can now pass through more different areas in all tracked frames with different viewing angles. The cost volume can thus have a larger cost variance for a pixel in the homogeneous areas.

Furthermore, the camera drift issue exists in all visual odometry systems because the ego-motion estimation has no true coordinate to refer to. A fusing model can provide such constraint so that we can perform pose graph optimisation to globally correct the drift.

6.3.2 Camera re-localisation

Tracking failure is still possible if the scenes have too much deformation or occlusion, which will make the quadrifocal warping ill-conditioned and eventually the camera will drift away. Currently the proposed system does not handle such cases. The ability to re-localise the camera after tracking failure is thus desirable.

A possible approach is to use a number of stored keyframes so that once the tracking is lost we can recover the pose by aligning the nearest stored keyframe. An efficient implementation would be to only register the coarsest image patches such as 20×15 (Lovegrove and Davison, 2010). A small number of keyframes should be sufficient since the endoscopic camera will only explore a limited area.

6.3.3 Registration of preoperative model

Registration of preoperative images to the patient is a fundamental requirement for any image-guided surgery system. This issue was tackled in the early work detailed in Chapter 3, but these attempts were ultimately unsuccessful. Instead we have

concentrated on a robust reconstruction and tracking system. Assuming we have a good initial alignment the methods proposed here will maintain registration as the endoscopic camera moves.

Manual registration is one option and has been proposed for some image guidance applications Pratt et al. (2012). But ideally we would like to have a robust and automated registration method. The biggest problem is that we do not currently have a decent model of the visible surface in the surgical scene. The use of the pubic arch is perhaps reasonable, but there is tissue over the surface of the bone that we have not accurately modelled from the preoperative scans.

One avenue for research is to try to identify the true operatively viewed surface in the preoperative MRI. Perhaps the layer of fascia above the bone can accurately be identified. Or perhaps we can learn from numerous operations where the viewed surface lies. The more accurate surface we hope to achieve from fusion may help to achieve this. In any case, with highly deforming soft tissue and visually challenging endoscopic scenes it is likely that registration will be an active area of research for some years to come and that dense vision will play an important part in this process. We have achieved an important step in this direction by providing accurate and robust tracking of the endoscope using only dense vision methods.

A Lie groups and their algebra

In linear algebra, a group consists of any set G and operation $\otimes$ that have the properties:

Closure	$\forall a, b \in G,\ a \otimes b \in G.$
Associativity	$\forall a, b, c \in G,\ (a \otimes b) \otimes c = a \otimes (b \otimes c) \in G.$
Identity	$\exists a \in G, \forall b \in G : a \otimes b = b \otimes a = b,$ and $a = 1_G.$
Inverse	$\forall a \in G, \exists b \in G : a \otimes b = b \otimes a = 1_G.$

A Lie group is any group which is also a finite dimensional smooth manifold, where the group operations of multiplication and inversion are smooth maps. Of particular interesting to us is that every Lie group has a correspondent Lie algebra which represents the tangent space in which the algebra expresses the derivative of the group's manifold (Rossman, 2002).

Parameterising a Lie group matrix with its algebra turns out to be a minimum parameterisation. This can significantly help in optimisation especially when we are consider incremental update scheme. Taking the *special orthogonal* Lie group $\mathbb{SO}(3)$ which is a 3×3 rotation matrix as an example, instead of optimising all 9-element in the matrix in Lie group space, its Lie algebra $\mathfrak{so}(3)$ reveals that the optimisation only need to apply to the minimal 3-vector parameterisation.

The mapping between a Lie group and its algebra is performed by the *exponential map* and *matrix logarithm*. For example, a *special Euclidean* Lie group $\mathbb{SE}(3) \in \mathbb{R}^{4\times4}$ matrix $\mathbf{T}$ can be mapped from its 6-vector Lie algebra $\mathbf{x} \in \mathfrak{se}(3)$ as

$$\mathbf{T} = \exp(g(\mathbf{x})), \quad \text{where} \quad g(\mathbf{x}) = \sum_{i=1}^{6} \mathbf{x}_i \mathbf{G}_i.$$

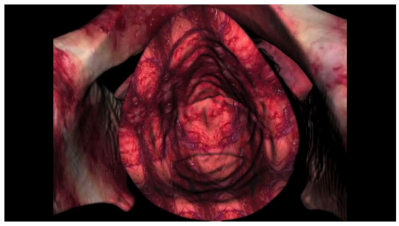
The $\mathbb{SE}(3)$ generator $\mathbf{G}_i$ is

$$\mathbf{G}_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \mathbf{G}_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \mathbf{G}_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{G}_4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \mathbf{G}_5 = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \mathbf{G}_6 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

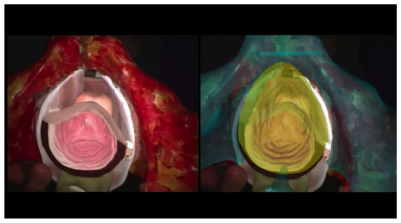
For the matrix logarithm and other Lie groups, please refer to (Rossman, 2002).

B Supplemental videos



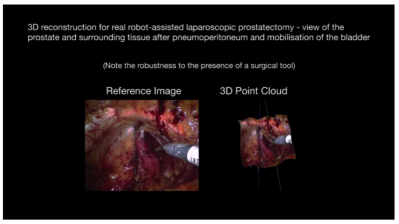
Phantom Simulation Sequence.

http://youtu.be/UyLnC6De1kw



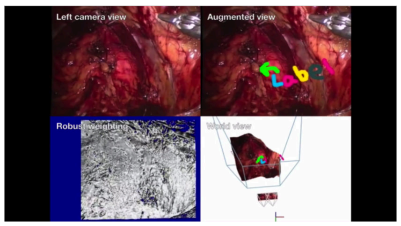
Dense Monocular Camera Tracking
(CMIG 2013)

http://youtu.be/fPWp8XiurCI



Real-Time Dense Stereo Reconstruction for Laparoscopic Prostatectomy
(MICCAI 2013)

http://youtu.be/jqfsv-G7of0



Real-Time Dense Stereo Camera Tracking for Laparoscopic Prostatectomy
(IPCAI 2014)

http://youtu.be/pwaf3dd43W0

Bibliography

Discussion for direct versus features session. In Bill Triggs, Andrew Zisserman, and Richard Szeliski, editors, *Vision Algorithms: Theory and Practice*, volume 1883 of *Lecture Notes in Computer Science*, pages 295–297. Springer Berlin Heidelberg, 2000. 24, 126

Gonzalo R. Arce. *Nonlinear signal processing: A statistical approach*. John Wiley & Sons, 2005. 71

Young Ki Baik, Junghyun Kwon, Hee Seok Lee, and Kyoung Mu Lee. Geometric particle swarm optimization for robust visual ego-motion estimation via particle filtering. *Image and Vision Computing*, 31(8):565–579, 2013. 125

Harlyn H Baker and Thomas O Binford. Depth from edge and intensity based stereo. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 2, pages 631–636, 1981. 69, 73

Simon Baker and Iain Matthews. Lucas-Kanade 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255, 2004. 98, 101, 126

Simon Baker, Raju Patil, Kong Man Cheung, and Iain Matthews. Lucas-Kanade 20 Years On: Part 5. Technical Report CMU-RI-TR-04-64, Robotics Institute, 2004. 98, 125, 126

Ronen Basri and David W Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(2): 218–233, 2003. 123

Peter N Belhumeur. A Bayesian approach to binocular steropsis. *International Journal of Computer Vision (IJCV)*, 19(3):237–260, 1996. 73

James R Bergen, Patrick Anandan, Keith J Hanna, and Rajesh Hingorani. Hierarchical model-based motion estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 237–252, 1992. 32, 109

Stan Birchfield and Carlo Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(4):401–406, 1998. 73

Michael J. Black, Guillermo Sapiro, David H. Marimont, and David Heeger. Robust anisotropic diffusion. *IEEE Transactions on Image Processing*, 7(3):421–432, 1998. 70

Michael Bleyer, Christoph Rhemann, and Carsten Rother. PatchMatch Stereo - Stereo matching with slanted support windows. In *Proceedings of the British Machine Vision Conference*, volume 11, pages 1–11, 2011. 65

Aaron F Bobick and Stephen S Intille. Large occlusion stereo. *International Journal of Computer Vision (IJCV)*, 33(3):181–200, 1999. 66, 73

Fred L Bookstein. *Morphometric tools for landmark data: geometry and biology*. Cambridge University Press, 1997. 47

Yuri Boykov, Olga Veksler, and Ramin Zabih. A variable window approach to early vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(12):1283–1294, 1998. 66

Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1222–1239, 2001. 69, 73

Kristian Bredies, Karl Kunisch, and Thomas Pock. Total generalized variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526, 2010. 72

Thomas Brox and Jitendra Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(3):500–513, 2011. 62

Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983. 109

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision (JMIV)*, 40(1):150–145, 2011. 63, 69, 79

Tony F. Chan and Selim Esedoglu. Aspects of total variation regularized L^1 function approximation. *SIAM Journal on Applied Mathematics*, 65(5):1817–1837, 2005. 69, 70

Ping-Lin Chang, Dongbin Chen, and Philip "Eddie" Edwards. Registration of a 3D preoperative model with 2D endoscopic images using parallel tracking and mapping (PTAM) with colour-consistency. In *Online Proceedings of the Medical Image Analysis and Understanding (MIUA)*, 2011. 43

Ping-Lin Chang, Dongbin Chen, Daniel Cohen, and Philip "Eddie" Edwards. 2D/3D registration of a preoperative model with endoscopic video using colour-consistency. In *Proceedings of the Augmented Environments for Computer-Assisted Interventions (AE-CAI) in Conjunction with MICCAI*, volume 7264, pages 1–12, 2012. 33, 43, 95

Ping-Lin Chang, Danail Stoyanov, Andrew J. Davison, and Philip "Eddie" Edwards. Real-time dense stereo reconstruction using convex optimisation with a cost-volume for image-guided robotic surgery. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 8149, pages 42–49, 2013. 58, 96

Ping-Lin Chang, Ankur Handa, Andrew J. Davison, Danail Stoyanov, and Philip "Eddie" Edwards. Robust real-time visual odometry for stereo endoscopy using dense quadrifocal tracking. In *Proceedings of the Information Processing in Computer-Assisted Interventionas (IPCAI)*, pages 11–20. 2014. 94

Matthew J. Clarkson, Daniel Rueckert, Derek L. G. Hill, and David J. Hawkes. Using photo-consistency to register 2D optical images of the human face to a 3D surface model. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1266–1280, 2001. 42

Andrew I. Comport, Ezio Malis, and Patrick Rives. Accurate quadrifocal tracking for robust 3d visual odometry. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 40–45, 2007. 125

Andrew I. Comport, Ezio Malis, and Patrick Rives. Real-time quadrifocal visual odometry. *The International Journal of Robotics*, 29(2-3):245–266, 2010. 23, 24, 94, 95, 96, 97, 123, 125

Alejo Concha, Wajahat Hussain, Luis Montano, and Javier Civera. Manhattan and piecewise-planar constraints for dense monocular mapping. In *Online Proceedings of Robotics: Science and Systems (RSS)*, 2014. 124

Franklin C. Crow. Summed-area tables for texture mapping. *SIGGRAPH Computer Graphics*, 18(3):207–212, 1984. 62

Andrew J Davison. Real-time simultaneous localisation and mapping with a single camera. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1403–1410, 2003. 23

Leonardo De-Maeztu, Stefano Mattoccia, Arantxa Villanueva, and Rafael Cabeza. Linear stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1708–1715, 2011. 67

Frédéric Devernay, Fabien Mourgues, and Ève Coste-Manière. Towards endoscopic augmented reality for robotically assisted minimally invasive cardiac surgery. In *International Workshop on Medical Imaging and Augmented Reality*, pages 16–20, 2001. 74

Tom Drummond and Roberto Cipolla. Visual tracking and control using lie algebras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 652–657, 1999. 102

Jakob Engel, Thomas SchÃűps, and Daniel Cremers. Semi-dense visual odometry for a monocular camera. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 126

Ines Ernst and Heiko Hirschmüller. Mutual information based semi-global stereo matching on the GPU. In *Proceedings of the Advances in Visual Computing*, pages 228–239, 2008. 73

Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient belief propagation for early vision. *International Journal of Computer Vision (IJCV)*, 70(1):41–54, 2006. 73

Pablo A. Ferrari, Arnoldo Frigessi, and Paula Gonzaga de Sá. Fast approximate map restoration of multicolor images. *Journal of the Royal Statistical Society*, 57 (3):485–500, 1995. 73

J Michael Fitzpatrick, Jay B West, and Calvin R Maurer Jr. Predicting error in rigid-body point-based registration. *IEEE Transactions on Medical Imaging (TMI)*, 17 (5):694–702, 1998. 22

Bernhard Froba and Andreas Ernst. Face detection with the modified census transform. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 91–96, 2004. 64

Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(8):1362–1376, 2010. 123

Qinquan Gao, Ping-Lin Chang, Daniel Rueckert, S Mohammed Ali, Daniel Cohen, Philip Pratt, Erik Mayer, Guang-Zhong Yang, Ara Darzi, and Philip "Eddie" Edwards. Modeling of the bony pelvis from mri using a multi-atlas AE-SDM for registration and tracking in image-guided robotic prostatectomy. *Computerized Medical Imaging and Graphics (CMIG)*, 37(2):183–194, 2013. 43, 95, 112

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 75

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, (6):721–741, 1984. 63

Oscar G Grasa, Javier Civera, and J. M. M. Montiel. EKF monocular slam with relocalization for laparoscopic sequences. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 4816–4821, 2011. 95

D. M. Greig, B. T. Porteous, and A. H. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society*, 51(2):271–279, 1989. 69, 72

W. Eric L. Grimson and Theo Pavlidis. Discontinuity detection for visual surface reconstruction. *Computer Vision, Graphics, and Image Processing*, 30(3):316–330, 1985. 70

Armin Gruen. Adaptive least squares correlation: A powerful image matching technique. *South African Journal of Photogrammetry, Remote Sensing and Cartography*, 14(3):175–187, 1985. 107

David Hafner, Oliver Demetz, and Joachim Weickert. Why is the census transform good for robust optic flow computation? In *Proceedings of the Scale Space and Variational Methods in Computer Vision*, volume 7893, pages 210–221, 2013. 64

Ankur Handa. *Analysing high frame-rate camera tracking*. PhD thesis, Imperial College London, 1999. 24, 126

Ankur Handa, Richard A. Newcombe, Adrien Angeli, and Andrew J. Davison. Applications of Legendre-Fenchel transformation to computer vision problems. Technical Report DTR11-7, Department of Computing at Imperial College London, 2011. 69, 80

Ankur Handa, RichardA. Newcombe, Adrien Angeli, and Andrew J. Davison. Real-time camera tracking: When is high frame-rate best? In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 7578, pages 222–235. 2012. 111, 112

Marsha Jo Hannah. *Computer matching of areas in stereo images.* PhD thesis, 1974. 63

Richard I. Hartley and Andrew Zisserman. *Multiple view geometry in computer vision.* Cambridge University Press, 2004. 94, 96, 97, 101, 125

Kaiming He, Jian Sun, and X Tang. Guided image filtering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–14, 2010. 66, 67

Philipp Heise, Sebastian Klose, Brian Jensen, and Alois Knoll. PM-Huber: PatchMatch with Huber regularization for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2360–2367, 2013. 65, 79

Yong Seok Heo, Kyoung-Mu Lee, and Sang-Uk Lee. Robust stereo matching using adaptive normalized cross-correlation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(4):807–822, 2011. 63

Heiko Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 807–814, 2005. 67, 73

Heiko Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(2):328–341, 2008. 73

Heiko Hirschmüller. Semi-global matching motivation, developments and applications. In *Photogrammetric Week*, volume 11, pages 173–184, 2011. 64

Heiko Hirschmuller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 74

Heiko Hirschmüller and Daniel Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(9):1582–1599, 2009. 64

Berthold K Horn and Brian G Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981. 31, 69, 70, 123

Berthold K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, 4(4):629–642, 1987. 47

Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(2):504–511, 2013. 67

Jinggang Huang and David Mumford. Statistics of natural images and models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 541–547, 1999. 70, 107

Jinggang Huang, Ann B. Lee, and David Bryant Mumford. Statistics of range images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 324–331, 2000. 70

Peter J Huber. *Robust statistics*. Wiley, New York, 1981. 70

Martin Humenberger, Christian Zinner, and Wilfried Kubinger. Performance evaluation of a census-based stereo matching algorithm on embedded and multi-core hardware. In *Proceedings of the International Symposium on Image and signal processing and analysis*, pages 388–393, 2009. 64

Martin Humenberger, Tobias Engelke, and Wilfried Kubinger. A census-based stereo vision algorithm using modified semi-global matching and plane fitting to improve matching quality. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 77–84. IEEE, 2010. 64

M. Irani and P. Anandan. About direct methods. In Bill Triggs, Andrew Zisserman, and Richard Szeliski, editors, *Vision Algorithms: Theory and Practice*, volume 1883, pages 267–277. Springer Berlin Heidelberg, 2000. 24, 126

Hiroshi Ishikawa. Exact optimization for Markov random fields with convex priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25 (10):1333–1336, 2003. 73

Pierre Jannin, J Michael Fitzpatrick, David J Hawkes, Xavier Pennec, Ramin Shahidi, Michael W Vannier, et al. Validation of medical image processing in image-guided therapy. *IEEE Transactions on Medical Imaging (TMI)*, 21(12): 1445–9, 2002. 42

Randy L Jensen, James L Stone, and Robert A Hayne. Introduction of the human Horsley-Clarke stereotactic frame. *Neurosurgery*, 38(3):563–567, 1996. 21

Myung-Ho Ju and Hang-Bong Kang. Constant time stereo matching. In *Proceedings of the IEEE International Machine Vision and Image Processing Conference*, pages 13–17, 2009. 66

Takeo Kanade and Masatoshi Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 16(9):920–932, 1994. 66

Takeo Kanade, Hiroshi Kano, Shigeru Kimura, Atsushi Yoshida, and Kazuo Oda. Development of a video-rate stereo machine. In *Proceedings of the International Robotics and Systems Conference (IROS)*, volume 3, pages 95–100, Aug 1995. 61

Diana C Kang, Miranda J Hardee, Susan F Fesperman, Taryn L Stoffs, and Philipp Dahm. Low quality of evidence for robot-assisted laparoscopic prostatectomy: results of a systematic review of the published literature. *European urology*, 57 (6):930–937, 2010. 20

Sing Bing Kang, Richard Szeliski, and Jinxiang Chai. Handling occlusions in dense multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–103. IEEE, 2001. 66

Christian Kerl, Jurgen Sturm, and Daniel Cremers. Robust odometry estimation for RGB-D cameras. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3748–3754, 2013. 107

Georg Klein and David Murray. Parallel tracking and mapping for small AR workspaces. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 225–234, 2007. 23, 24, 42, 44, 46, 110, 123

Kalin Kolev, Thomas Pock, and Daniel Cremers. Anisotropic minimal surfaces integrating photoconsistency and normal information for multiview stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 538–551, 2010. 123

Vladimir Kolmogorov and Ramin Zabih. Computing visual correspondence with occlusions using graph cuts. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 508–515. IEEE, 2001. 63, 73

Vladimir Kolmogorov and Ramin Zabin. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26(2):147–159, 2004. 73

William W. Lau, Nicholas A. Ramey, Jason J. Corso, Nitish V. Thakor, and Gregory D. Hager. Stereo-based endoscopic tracking of cardiac surface deformation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 3217, pages 494–501, 2004. 74

David Lee and Theodosios Pavlidis. One-dimensional regularization with discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 10(6):822–829, 1988. 70

J. P. Lewis. Fast normalized cross-correlation. *Vision interface*, 10(1):120–123, 1995. 63

Joerg Liebelt and Klaus Schertler. Precise registration of 3d models to images by swarming particles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 125

Dahua Lin, Eric Grimson, and John Fisher. Learning visual flows: A Lie algebraic approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 747–754, 2009. 102

Steven Lovegrove and Andrew J Davison. Real-time spherical mosaicing using whole image alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 73–86, 2010. 128

Bruce D. Lucas. *Generalized image matching by the method of differences*. PhD thesis, Robotics Institute, Carnegie Mellon University, July 1984. 32

Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference*

on Artificial Intelligence (IJCAI), volume 81, pages 674–679, 1981. 31, 32, 100, 123

Yi Ma. *An invitation to 3-D vision: From images to geometric models*, volume 26. springer, 2004. 102

Ziyang Ma, Kaiming He, Yichen Wei, Jian Sun, and Enhua Wu. Constant time weighted median filtering for stereo matching and beyond. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 71

Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information. *Medical Image Analysis (MedIA)*, 3(4):373–386, 1999. 48

L. Maier-Hein, A. Groch, A. Bartoli, S. Bodenstedt, G. Boissonnat, P.-L. Chang, N.T. Clancy, D. S. Elson, S. Haase, E. Heim, J. Hornegger, P. Jannin, H. Kenngott, T. Kilgus, B. Müller-Stich, D. Oladokun, S. Röhl, T. R. Dos Santos, H.P. Schlemmer, A. Seitel, S. Speidel, M. Wagner, and D. Stoyanov. Comparative validation of single-shot optical techniques for laparoscopic 3D surface reconstruction. *IEEE Transactions on Medical Imaging (TMI)*, 2014. 58, 75, 90, 124

Lena Maier-Hein, Peter Mountney, A. Bartoli, H. Elhawary, Daniel S. Elson, Anja Groch, A. Kolb, M. Rodrigues, J. Sorger, Stefanie Speidel, and Danail Stoyanov. Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery. *Medical Image Analysis (MedIA)*, 17(8):974–996, 2013. 96

Ezio Malis. Improving vision-based control using efficient second-order minimization techniques. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, volume 2, pages 1843–1848, 2004. 98, 101, 102

David Marr and Tomaso Poggio. Cooperative computation of stereo disparity. *Science*, 194(4262):283–7, October 1976. 69

Larry Matthies, Takeo Kanade, and Richard Szeliski. Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision (IJCV)*, 3(3):209–238, 1989. 61

Daniel J. Mirota, Masaru Ishii, and Gregory D. Hager. Vision-based navigation in image-guided interventions. *Annual Review of Biomedical Engineering*, 13:297–319, 2011. 58, 74

Peter Mountney, Danail Stoyanov, Andrew J. Davison, and Guang-Zhong Yang. Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9, pages 347–54, 2006. 43, 74, 95

Peter Mountney, Benny Lo, Surapa Thiemjarus, Danail Stoyanov, and Guang Zhong-Yang. A probabilistic framework for tracking deformable soft tissue in minimally invasive surgery. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 34–41. 2007. 43

Peter Mountney, Danail Stoyanov, and Guang-Zhong Yang. Three-dimensional tissue deformation recovery and tracking. *IEEE Signal Processsing Magazine*, 27(4): 14–24, 2010. 58, 74

Fabien Mourgues, F Devemay, and Eve Coste-Maniere. 3D reconstruction of the operating field for image overlay in 3D-endoscopic surgery. In *Proceedings of the IEEE and ACM International Symposium on Augmented Reality*, pages 191–192, 2001. 43

Richard A Newcombe and Andrew J Davison. Live dense reconstruction with a single moving camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1498–1505, 2010. 23, 44, 123

Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. KinectFusion: Real-time Dense Surface Mapping and Tracking. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 127–136, 2011a. 23, 42, 95, 125

Richard A. Newcombe, Steven J. Lovegrove, and Andrew J. Davison. DTAM: Dense tracking and mapping in real-time. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 2320–2327, 2011b. 23, 24, 42, 51, 95, 98, 110, 123, 125, 128

Mila Nikolova. A variational approach to remove outliers and impulse noise. *Journal of Mathematical Imaging and Vision*, 20(1-2):99–120, 2004. 70

Abhijit S. Ogale and Yiannis Aloimonos. Shape and the stereo correspondence problem. *International Journal of Computer Vision*, 65(3):147–162, 2005. 65

Yuichi Ohta and Takeo Kanade. Stereo by intra-and inter-scanline search using dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, (2):139–154, 1985. 69

Cosmin D. Pantilie and Sergiu Nedevschi. Optimizing the census transform on CUDA enabled GPUs. In *Proceedings of the IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 201–207, 2012. 64

Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama. Digital photography with flash and no-flash image pairs. 23(3):664–672, 2004. 66

T. Pock and A. Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1762–1769, 2011. 82, 125

Tomaso Poggio, Vincent Torre, and Christof Koch. Computational vision and regularization theory. *Nature*, 317:26, 1985. 69, 70

Fatih Porikli. Constant time O(1) bilateral filtering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 66

Michael JD Powell. The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 2009. 46

Philip Pratt, Erik Mayer, Justin Vale, Daniel Cohen, Eddie Edwards, Ara Darzi, and Guang-Zhong Yang. An effective visualisation and registration system for image-guided robotic partial nephrectomy. *Journal of Robotic Surgery*, 6(1):23–31, 2012. 129

Victor A Prisacariu and Ian D Reid. PWP3D: Real-time segmentation and tracking of 3D objects. *International Journal of Computer Vision (IJCV)*, 98(3):335–354, 2012. 33, 55

CH Quartucci Forster and Clesio L Tozzi. Towards 3D reconstruction of endoscope images using shape from shading. In *Proceedings of the Brazilian Symposium on Computer Graphics and Image Processing*, pages 90–96, 2000. 43

Rene Ranftl, Stefan Gehrig, Thomas Pock, and Horst Bischof. Pushing the limits of stereo using variational stereo estimation. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, number 1, pages 401–407, 2012. 64, 72, 124

Rene Ranftl, Thomas Pock, and Horst Bischof. Minimizing TGV-based variational models with non-convex data terms. In *Proceedings of the Scale Space and Variational Methods in Computer Vision*, volume 7893, pages 282–293. 2013. 72, 79, 124

Carl Yuheng Ren and Ian Reid. A unified energy minimization framework for model fitting in depth. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops and Demonstrations*, pages 72–82, 2012. 107

Christoph Rhemann, Asmaa Hosni, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3017–3024, 2011. 62, 67, 76, 86, 87

Rogério Richa, Antônio Padilha Lanari Bo, and Philippe Poignet. Towards robust 3D visual tracking for motion compensation in beating heart surgery. *Medical Image Analysis (MedIA)*, 15(3):302–315, 2011. 74

Sebastian Rohl, Sebastian Bodenstedt, Stefan Suwelack, Rudiger Dillmann, Stefanie Speidel, Hannes Kenngott, and Beat P Muller-Stich. Dense GPU-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration. *Medical physics*, 39(3):1632–45, 2012. 74, 76, 86, 90

Wulf Rossman. *Lie groups: an introduction through linear groups*. Oxford, 2002. 106, 131, 132

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. 69, 70, 78

Mathieu Salzmann and Pascal Fua. Deformable surface 3D reconstruction from monocular images. *Synthesis Lectures on Computer Vision*, 2(1):1–113, 2010. 43

Jason Sanders and Edward Kandrot. *CUDA by Example: An Introduction to General-Purpose GPU Programming*. Addison-Wesley Professional, 1st edition, 2010. 39

Daniel Scharstein and Chris Pal. Learning conditional random fields for stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 74

Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision (IJCV)*, 4(1):131–140, 2002. 60, 74, 123

Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–195, 2003. 74

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. 47

Natalia Slesareva, Andrés Bruhn, and Joachim Weickert. Optic flow goes stereo: A variational method for estimating discontinuity-preserving dense disparity maps. In *Proceedings of the Pattern Recognition*, pages 33–40, 2005. 62, 70

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pages 2951–2959, 2012. 125

Robert Spangenberg, Tobias Langner, and Raúl Rojas. Weighted semi-global matching and center-symmetric census transform for robust driver assistance. In *Proceedings of the Computer Analysis of Images and Patterns*, pages 34–41, 2013. 64

Jakob Spoerk, Christelle Gendrin, Christoph Weber, Michael Figl, Supriyanto Ardjo Pawiro, Hugo Furtado, Daniella Fabri, Christoph Bloch, Helmar Bergmann, Eduard Gröller, et al. High-performance GPU-based rendering for real-time, rigid 2D/3D-image registration and motion prediction in radiation oncology. 22(1): 13–20, 2012. 45

Frank Steinbrücker, Thomas Pock, and Daniel Cremers. Advanced data terms for variational optic flow estimation. In *Proceedings of the vision, modeling, and visualization workshop*, pages 155–164, 2009. 62, 78

Charles V Stewart. Robust parameter estimation in computer vision. *SIAM Review*, 41(3):513–537, 1999. 107, 108

Danail Stoyanov. Stereoscopic scene flow for robotic assisted minimally invasive surgery. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 7510, pages 479–86, 2012. 74, 95

Danail Stoyanov, Ara Darzi, and Guang Zhong Yang. Dense 3D depth recovery for soft tissue deformation during robotically assisted laparoscopic surgery. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 41–48. 2004. 43

Danail Stoyanov, M. Scarzanella, Philip Pratt, and Guang-Zhong Yang. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 6361, pages 275–282, 2010. 74, 86, 87, 90

Hauke Strasdat, JMM Montiel, and Andrew J Davison. Real-time monocular SLAM: Why filter? In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2657–2664, 2010. 127

Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(7):787–800, 2003. 73

Marshall F. Tappen and William T. Freeman. Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 900–906, 2003. 73

Demetri Terzopoulos. Regularization of inverse visual problems involving discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, (4):413–424, 1986. 70

Andrey Nikolayevich Tikhonov. On the stability of inverse problems. 39(5): 195âĂŞ198, 1943. 35, 69

Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 839–846, 1998. 66

Johannes Totz, Peter Mountney, Danail Stoyanov, and Guang-Zhong Yang. Dense surface reconstruction for enhanced navigation in MIS. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 14, pages 89–96, 2011. 95

Olga Veksler. *Efficient graph-based energy minimization methods in computer vision.* PhD thesis, Cornell University, 1999. 73

Olga Veksler. Fast variable window for stereo correspondence using integral images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 556–561, 2003. 62

Hanzi Wang, Daniel Mirota, Masaru Ishii, and Gregory D Hager. Robust motion estimation and structure recovery from endoscopic image sequences with an adaptive scale kernel consensus estimator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7, 2008. 43

Steve Webb. *The physics of medical imaging.* Medical science series. Taylor & Francis, 1988. 20

Michael Weber, Martin Humenberger, and Wilfried Kubinger. A very fast census-based stereo matching implementation on a graphics processing unit. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 786–793, 2009. 64

Manuel Werlberger, Werner Trobin, Thomas Pock, Andreas Wedel, Daniel Cremers, and Horst Bischof. Anisotropic Huber-L^1 optical flow. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2009. 71, 77, 78

Chia-Hsiang Wu, Yung-Nien Sun, Yi-Chiao Chen, and Chien-Chen Chang. Endoscopic feature tracking and scale-invariant estimation of soft-tissue structures. *IEICE Transactions on Information and Systems*, 91(2):351–360, 2008. 43

Qingxiong Yang. A non-local cost aggregation method for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number 1, pages 1402–1409, 2012. 67

Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Real-time $O(1)$ bilateral filtering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 557–564, 2009. 66

Jonathan S Yedidia, William T Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. *Exploring Artificial Intelligence in the New Millennium*, 8:236–239, 2003. 69, 73

Kuk-Jin Yoon and In-So Kweon. Locally adaptive support-weight approach for visual correspondence search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 924–931, 2005. 66

Kuk-Jin Yoon and In So Kweon. Adaptive support-weight approach for correspondence search. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(4):650–656, 2006. 66

Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume II, 1994. 63

Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime TV-L^1 optical flow. In *Proceedings of the Pattern Recognition*, pages 214–223, 2007. 78

Zhengyou Zhang. Parameter estimation techniques: A tutorial with application to conic fitting. *Image and Vision Computing*, 15(1):59–76, 1997. 70, 107, 108