Technische Universität Berlin

THE ROBUST 3-D SCENEFLOW

Generalized Video-based 3-D Analysis using
Robust Camera and Scene Geometry Estimations

Von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Verleihung des akademischen Grades

Doktor der Ingenieurwissenschaften
Dr.-Ing.

genehmigte Dissertation

vorgelegt von

Jang Heon Kim

aus Süd Korea

Promotionsausschuss:
Vorsitzender: Prof. Dr.-Ing. Klaus-Robert Müller
Berichter: Prof. Dr.-Ing. Olaf Hellwich
Berichter: Prof. Dr.-Ing. Thomas Sikora

Tag der wissenschaftlichen Aussprache : 9. 9. 2008

Berlin 2009
D 83

Technische Universität Berlin

THE ROBUST 3-D SCENEFLOW

Generalized Video-based 3-D Analysis using
Robust Camera and Scene Geometry Estimations

by

Jang Heon Kim

Master of Science in Electronic Engineering
Yonsei University, Seoul, Korea, 2003

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

Doctor of Engineering
in
Electrical Engineering and Computer Science

Faculty IV. Electrical Engineering and Computer Sciences
Technische Universität Berlin

Berlin 2009
D 83

Technische Universität Berlin
Address: Strasse des 17. Juni 135, 10623 Berlin, Germany
Phone: +49 (0)30 314-0, Fax: +49 (0)30 314-23222
http://www.tu-berlin.de

Institut für Telekommunkationssysteme
Address: Sekr. EN 1, Einsteinufer 17, 10587 Berlin, Germany
Phone: +49 (0)30 314-25093, Fax: +49 (0)30 314-22514
http://www.nue.tu-berlin.de

Declaration

This dissertation is submitted to the Technische Universität Berlin in partial fulfillment for the degree of Doctor of Philosophy. It is an account of work undertaken at the Faculty IV. Electrical Engineering and Computer Sciences between May 2004 and September 2008 under the supervision of Prof. Dr.-Ing. Thomas Sikora.

I already read and understand all regulations for completion of the degree of Doctor of Philosophy. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration. This dissertation is not substantially the same as any I have submitted for a degree or diploma or other qualification at any other University. I further state that no part of my dissertation/thesis has already been, or is being concurrently submitted for any other degree, or other qualification.

September, 2008

Jang Heon Kim

To my Parents
inspiring me and
dedicating their whole lives
for my education.

Contents

List of Tables

List of Figures

Acknowledgments

The author would like to express his gratitude to those whose support was essential for the completion of this thesis.

I am deeply in debt to Professor Dr. Thomas Sikora whose help, stimulating suggestions and encouragement helped me in all the time of research for and writing of this thesis. His knowledge and advice have always helped me building and finishing my researches in Berlin. I am grateful to have this opportunity to work on this thesis under his supervision. I would like to thank my thesis committee members, Professor Dr. Olaf Hellwich and Professor Dr. Klaus-Robert Müller, for providing me with valuable advice and direction to improve the thesis. My thanks extend for their enthusiasm, support and encouragement that have been a continual source of inspiration for this work.

I would also express my sincere gratitude to Dr. Ralf Schäfer, Peter Kauff, Dr. Peter Eisert, Dr. Oliver Schreer and Dr. Aljoscha Smolic at the Image Processing Department of Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut for providing me an opportunity to study in Germany. My special thanks are to Ingo Feldmann, Ralf Tanger and Valeri George for their kind supports.

To Matthias Kunter, who was my colleague and roommate at Technischen Universität Berlin, I am deeply grateful for his cooperation and sharing of knowledge and experience. Also I want to thank him for the comments on my research works. I thank Sebastian Knorr, Mustafa Karaman and Lutz Goldmann, for their knowledgeable help and suggestions in 2-D and 3-D video analysis and thesis preparation. I am sincerely grateful to Ms. Birgit Boldin for kind helps. Dr. Ronald Glasberg, Amjad Samour, Georges Haboub, Carsten Clemens, Martin Haller, Rüdiger Knörig, Gunnar Eisenberg, Shan Jin, Andreas Krutz, Tilman Liebchen, Engin Kurutepe, Woong Hee Kim, Andreas Cobet, Markus Schwab, Jan Weil, Frank Lukowski, Arved Peix, Wiryadi and all other coworkers and friends at Technischen Universität Berlin who gave me advice and made my life in Berlin a more rewarding experience.

My sincere love and thanks to my parents, my brother and sister and my family, I am thankful for their love and devoted support for all these years.

Abstract

Recovering 3-D information from several 2-D images is one of the most important topics in computer vision. There are a lot of applications in different areas such as TV contents, games, medical applications, robot navigation and special effects, etc. Multi-stereo and Structure-from-Motion methods aim to recover the 3-D camera pose and scene structure for a rigid scene from an uncalibrated sequence of 2-D images. The 3-D camera pose can be estimated as the principle projection ray for a camera observing a rigid scene. The line of principal ray defines the collineation of all viewpoints observing the same scene. However, the projective camera geometry is involved in a number of distorted images for a Euclidean geometric object since the projection is a non-metrical form of geometry. This means that the collineation of projective ray is not always satisfied in metrically distorted pixels in the viewpoints, and the distortion is the image form of divergent rays on the 3-D surfaces. The estimation of dense scene geometry is a process to recover the metric geometry and to adjust the global ray projection passing through each 2-D image point to a 3-D scene point on real object surfaces. The generalization of 3-D video analysis depends on the density and robustness of the scene geometry estimation. In this dissertation, the 3-D sceneflow method that analyzes jointly stereo and motion is proposed for retrieving the camera geometry and reconstructing dense scene geometry accurately. The stereo viewpoints provide more robust 3-D information against noises, and the viewpoints of the camera motion increase the spatial density of the 3-D information. This method utilizes a brightness-invariance constraint that a set of spatio-temporal correspondences do not change for a 3-D scene point. Due to physical imperfections in the imaging sensors and bad locations of detected features and false matches, image data contain a lot of outliers. A unified scheme of robust estimation is proposed to eliminate outliers both from feature-based camera geometry estimates and dense scene geometry estimates. With a robust error weight, the error distribution of estimates can be restricted in a smoothly invariant support region, and the anisotropic diffusion regularization efficiently eliminates outliers in the regional structure. Finally, the structure-preserving dense 3-D sceneflow is obtained from stereo-motion sequences for a 3-D natural scene.

Zusammenfassung

Die Gewinnung der 3D-Information aus unterschiedlichen 2D-Bildern gehört zu den wichtigsten Forschungsthemen der Computer Vision. Dieses Feld hat eine Vielzahl von Anwendungen in den verschiedensten Bereichen, z. B. bei der Generierung von TV-Inhalten, für Videospiele, im medizinischen Bereich, bei der Roboternavigation und für Spezialeffekte im Kino. Dabei gibt es bereits eine Menge Forschungsarbeiten zum Thema, welche allgemein zum Ziel haben, die 3D-Koordinaten der Kamera und die Szenenstruktur für statische Inhalte aus unkalibrierten 2D-Bilder bestimmen, dies zum Beispiel mittels des Multi-Stereo-Ansatzes oder durch die Structure-from-Motion-Analyse. Die 3D-Kameraposition kann mittels der Hauptprojektionsachse einer Kameraaufnahme für eine statische Szene geschätzt werden. Diese Achse definiert die Collineation der Blickpunkte, welche die gleiche Szene beobachten. Allerdings verursacht die projektive Kamerageometrie bestimmte Bildverzerrungen für Euklidische Objekte, da sie selbst keine metrische Abbildung darstellt. Dies hat zur Folge, dass die Collineation von Strahlen nicht mehr für metrisch verzerrte Pixel erfüllt ist, wobei dann die Verzerrung selber die Abbildung divergenter Strahlen darstellt. Die Schätzung der dichten Szenengeometrie ist ein Prozess, mit dessen Hilfe einerseits die metrische Geometrie wiedergewonnen wird und andererseits die globale Projektion so adjustiert wird, dass jeder Bildpunkt bei der Projektion in den 3D-Raum auf die echte Objektoberfläche abgebildet wird. Die Verallgemeinerung der 3D-Videoanalyse hängt dabei von der Dichte und Robustheit der Schätzung der 3D-Szenengeometrie ab. In dieser Dissertation wird zur Ermittlung der Kamerageometrie und für die Rekonstruktion der dichten Szenengeometrie die Methode des dreidimensionalen Szenenflusses vorgeschlagen, welche die Stereo- und Bewegungsinformationen gemeinsam analysiert. Hinsichtlich verschiedenartiger Rauschstörungen bieten Stereoanordnungen robustere 3D-Informationen, hingegen erhöhen die Blickpunkte einer sich bewegenden Kamera die räumliche Kameradichte. Die vorgeschlagene Methode geht dabei von der Helligkeitsinvarianzbedingung aus, welche besagt, dass raum-zeitliche Korrespondenzen zwischen Abbildungen eines 3D-Szenenpunktes unveränderlich sind. Aufgrund physikalischer Ungenauigkeiten der optischen Sensoren, schlechter Lokalisierungen von Merkmalspunkten und sich daraus ergebender fehlerhafter Punktzuordnungen enthalten die Bilddaten eine Menge von Ausreißern. Deshalb wird hier ein vereinheitlichter Ansatz zur robusten Schätzung und Eliminierung der Ausreißer bei gleichzeitiger Bestimmung der Kamera-

geometrie sowie der dichten Szenengeometrie verfolgt. Mittels einer robusten Fehlerwichtung, bei der die Fehlerverteilung der Schätzungen auf einen näherungsweise invarianten Bereich begrenzt wird, und durch Verwendung einer auf anisotroper Diffusion basierender Regularisierung werden die Ausreißer in der regionalen Struktur effizient eliminiert. Letztendlich kann der dichten 3D-Szenenfluss aus Stereobildsequenzen mit natürlichen Inhalten gewonnen werden.

Chapter 1

Introduction

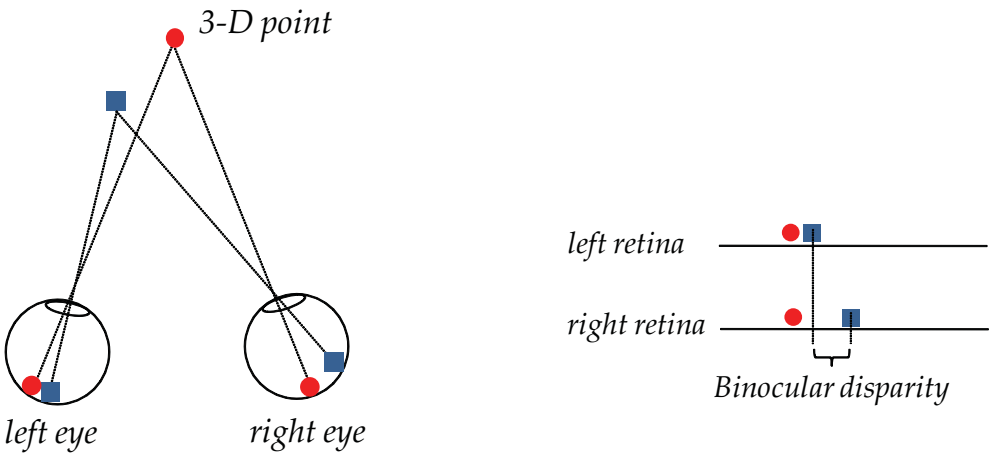
1.1 Depth Cues

One of the major functions of the human visual system is to construct a 3-D representation of the real world surrounding us. The images on our retina are patterns of lights reflected from our environment. Since the world is 3-D, we have to perceive a 3-D sensation in order to discover what is present. The problem is that the 3-D world space is projected onto the retina of both eyes as 2-D images. However, the human brain can process subtle differences between two retinal 2-D images with a slightly different perspective to perceive 3-D. The difference between the two eyes' images is a result of the eyes' horizontal separation, this is referred to as *binocular disparity* or *retinal disparity*. The process in visual perception leading to the stereoscopic depth is referred to as *stereopsis* [1]. Figure 1.1(a) and 1.1(b) respectively illustrates the binocular projective geometry in the human visual system and the binocular disparity. Cues for depth perception can be divided in four categories [2] as

(a) Ocular information: Accommodation[1] and convergence.
(b) Stereoscopic information: Binocular disparity.
(c) Dynamic information: Motion parallax.
(d) Pictorial information: Occlusion and relative scale.

Computer stereo vision : The 3-D real world can be analyzed by more than two viewpoint images simultaneously captured at slightly different positions. Cameras taking images at different positions can generate a depth cue to determine the relative distances between objects, e.g. disparity and parallax of 3-D objects, as our eyes do [3]. Figure 1.2(a) and 1.2(b) represent the left image and the right image of stereo camera. A computer compares the images to find parts that match, and the distance between the best matches can be used to determine the depths of 3-D objects. An image with the best matches designates as *disparity map*. Figure 1.2(c) and 1.2(d) show disparity maps that are respectively generated by the left-to-right

1) Accommodation is the process of focusing on an object.

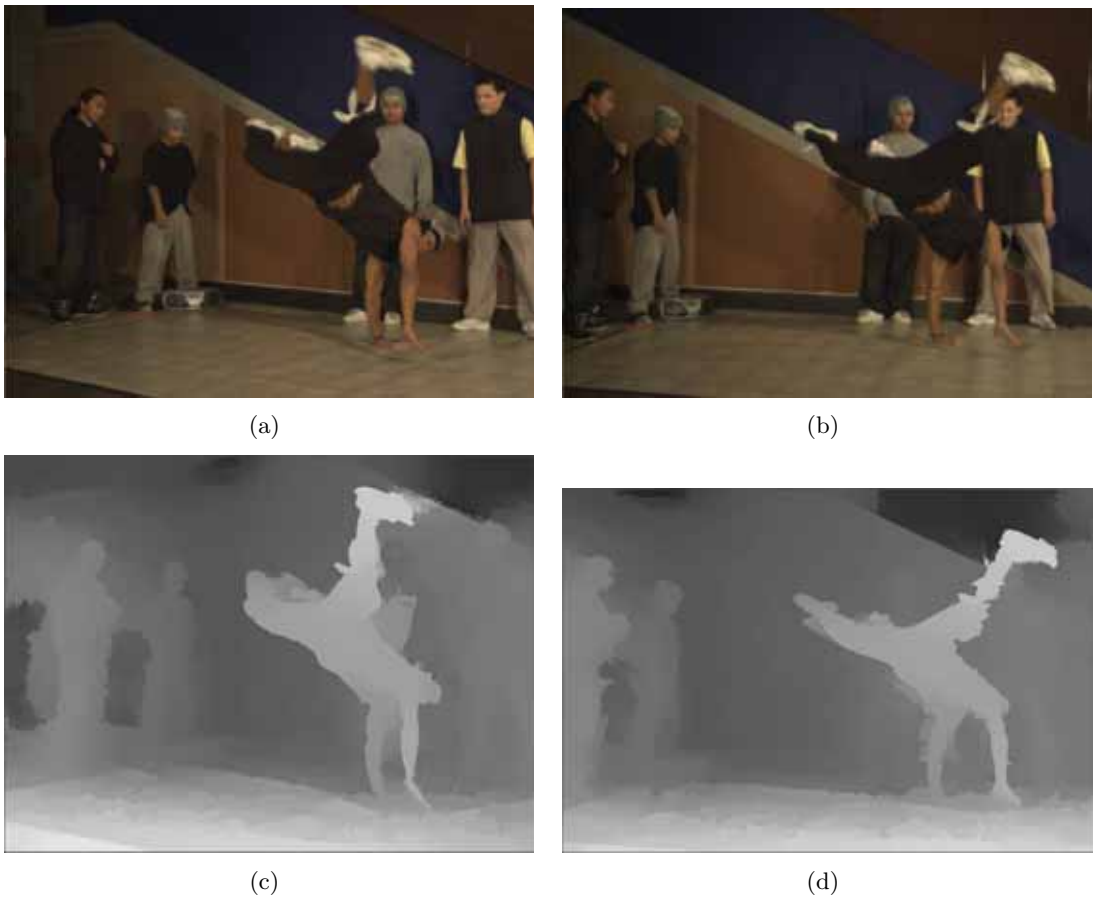


(a) (b)

Figure 1.1. Human visual system - (a) geometry of binocular projection by the human visual system, (b) definition of binocular disparity

and the right-to-left image matchings. The camera angles according to the relative camera distances for the fixed 3-D positions make a geometric relation referred as *epipolar geometry*. [89] The simplest case is when the camera image planes are on the common plane. Multiple camera images may be converted to be on the same image plane by reprojection through a linear transformation. This process is called *rectification*. [4] Disparity estimation is often performed on the rectified images and used to recover the 3-D scene structure. The process to find the depth of a 3-D point through measurements of side angles in a triangle formed by a 3-D point and two cameras is referred as *triangulation*. [5]

Understanding of motion : In computer vision, the understanding of motion can be divided into two categories: 2-D motion and 3-D motion. 2-D motion refers to the displacement of 2-D components in the image plane, e.g. pixels, edges, regions and image object, etc., that can be measured by matching or optical flow methods. 3-D motion refers to the estimation of rotation and translation of rigidly moving objects relative to the camera. If the scene is static, the measured 2-D motions totally derive from the camera motions that can be modeled as one of translational affine or projective. Even though a single 3-D motion is present, multiple 2-D motion models can be analyzed in the images due to the perspective effects, depth discontinuities, occlusions and etc. [25, 26] When the scene is dynamic, i.e. a moving object captured by a hand-held camera, it is difficult to separate the camera motions from the object motions because the object motions are not global but local and independent. In Figure 1.3(a), the motion trail of a dynamic scene taken by a static camera exemplifies the local and independent motions. The motion blurring effects in the motion trail are caused by



(a) (b)

(c) (d)

Figure 1.2. Disparity map - (a) left-viewpoint image of stereo camera, (b) right-viewpoint image of stereo camera, (c) disparity map using left-to-right matching, (d) disparity map using right-to-left matching [90]

a temporal motion. The camera motion may be mixed with the object motions. However, a set of synchronized multiple cameras can capture both camera motion and object motions as Figure 1.3(b) shows. The global motions in the image are induced by the 3-D camera locations in the world coordinate system and thus can be modeled as a 3-D motion. The relative distances from multiple cameras can be approximated by the affine motion model. Several 2-D and 3-D motion models are shown in Table 1.1, and the introduction of the models will be described in Chapter 3.1.

1.2 3-D Video

3-D video refers to *stereoscopic motion-picture images* in depth. Recording a 3-D video can be via synchronized dual cameras whereby the left and right eye views of a scene are recorded by the two camera views. 3-D graphic models and animations can be easily converted to a

(a) (b)

Figure 1.3. Spatio-temporal motion effects - (a) motion trail of a dynamic scene captured by a static camera, (b) motion trail of a scene captured by multiple cameras

Table 1.1. 2-D and 3-D motion models in computer vision

Motion model	Model equations	Model parameters
2-D translational	$\mathbf{m}_2 = \mathbf{m}_1 + \mathbf{t}_i$	$\{\mathbf{t}_i \in \mathbb{R}^2\}_{i=1}^n$
2-D similarity	$\mathbf{m}_2 = \alpha_i \mathbf{R}_i \mathbf{m}_1 + \mathbf{t}_i$	$\{\mathbf{R}_i \in SO(2), \alpha_i \in \mathbb{R}^+\}_{i=1}^n$
2-D affine	$\mathbf{m}_2 = \mathbf{A}_i \begin{bmatrix} \mathbf{m}_1 \\ 1 \end{bmatrix}$	$\{\mathbf{A}_i \in \mathbb{R}^{2\times 3}\}_{i=1}^n$
3-D translational	$\mathbf{m}_2^T \hat{\mathbf{t}}_i \mathbf{m}_1 = 0$	$\{\hat{\mathbf{t}}_i \in \mathbb{R}^3\}_{i=1}^n$
3-D rigid-body	$\mathbf{m}_2^T \mathbf{F}_i \mathbf{m}_1 = 0$	$\{\mathbf{F}_i \in \mathbb{R}^{3\times 3}\}_{i=1}^n$
3-D homography	$\mathbf{m}_2 \sim \mathbf{H}_i \mathbf{m}_1$	$\{\mathbf{H}_i \in \mathbb{R}^{3\times 3}\}_{i=1}^n$

stereo view or any number of views. These 3-D recordings are composited and edited for stereoscopic 3-D playback either in video using stereo glasses[2], or auto-stereoscopic displays[3]. A 3-D video give more sensory information to the viewer than a conventional 2-D video. The addition of binocular depth in a video gives people a higher sense of *being there* in a displayed scene. An increase in sensory information, through the addition of stereoscopic cues, may enhance the presence[4] of viewers. [6, 8, 9] From a different point of view, 3-D video also defines *spatial motion pictures* seen on normal TV or monitors with horizontal parallax which occupyies a 3-D volume. A study of presence [7] revealed that a moving sequence has relatively larger effects on presence ratings than the stereoscopic effect. Thus, a key feature of 3-D video is interactivity in the sense that a user has the possibility to choose an arbitrary

2) e.g., LCS shutter, polarized projection and anaglyph glasses, etc.
3) e.g., Lenticular lens, parallax barrier and cross-lenticular displays, etc.
4) Presence is a concept which is referred to the subjective experience of being in one place or environment even when one is situated in another or an unremarked sense of *being there and reacting to* in amediated environment. As the sense of presence increases, people become more aware of the mediated environment, and less aware of the environment in which they are physically located.

viewpoint within a visual real-world scene. Spatio-temporal effects such as freeze-and-rotate have been demonstrated in several recent movies, like The Matrix [10] or Romeo Must Die [11]. However, these effects can only be generated by using a large number of still cameras and involving a considerable amount of manual editing. The advance of computer vision techniques have solved the problem of the 3-D motion properties of objects in a scene and recently making it possible to produce automatic and interactive virtual-reality[5] using images and videos. According to of scene and camera configurations, 3-D video processing methods can be categorized as follows:

(a) Static scene analysis : Capturing multiple views of a static scene is relatively simple since only a single camera is needed. The camera can be horizontally moved to take multiple viewpoint images of the static scene. If the camera geometry can be established with the use of hardware which controls the movement of camera, e.g. a robotic arm or a similar mechanism like Crane and Jib arm, novel views can be synthesized in the camera geometry. A camera gantry is used in [12] to capture light field, which assumes that the camera locations form a uniform grid on a 2-D plane. In concentric mosaics [13], a camera is mounted on the tip of a rotating arm, which captures a series of images whose centers of projection are along a circle. A software approach that estimates the camera geometry is camera calibration. In the study of Lumigraph [14], a hand-held camera is used to capture a scene which contains three planar patterns to perform the camera calibration. In [15], a CCD camera with a range scanner attached to a spherical arm is used to capture images and the scene geometry simultaneously, and camera calibration is performed to determine the camera positions.

(b) Dynamic scene analysis : For the acquisition of a dynamic scene, an array of static cameras is needed in most cases. The cameras need to be synchronized if correspondence between images will be explored. An exceptional case is automatically self-reconfigurable camera arrays mounted on robotic servos as [16] shows. While the cameras are capturing a calibration pattern in the scene, the cameras adjust the pose to minimize the calibration error and to acquire more accurate images for rendering. Although a number of multi-view video systems allow for the realistic rendering of a time-varying scene, the capturing systems should be confined to dense cameras which are placed very close together for producing high-quality results. A dynamic light field camera using multiple baselines of a couple of centimeters does not need any scene geometry at all. [17] However, if the baseline is wide, the scene geometry that is predominantly computed by stereo methods should be calculated to compensate the sampling distortions according to the object depths.

5) Non-interactive 3-D video is passive virtual-reality as in theater presentations, whereas user-determined 3-D video is interactive virtual-reality.

1.3 Multi-view Approaches for 3-D Motion Analysis

Structure-from-Motion (SfM) [18, 19] and Multi-stereo [20–22] are approaches that establish the relative distances between 3-D cameras and scene points. SfM and Multi-stereo respectively use a monocular camera in motion and a set of multiple static cameras. Figure 1.4(a) and 1.4(b) illustrate SfM and Multi-stereo approaches.

Rigid-body motion : When the scene is static, both methods can establish a model of 3-D camera motion from the trajectories of a set of 2-D feature points or fits the 3-D structures to reliable 2-D correspondences in viewpoint images. These methods can be well understood when the relative 3-D positions between scene points and cameras can be defined by *a rigid-body motion* [23]. For a fixed camera position, the movement of a point

$$(\nabla \mathbf{m})^T \mathbf{v} + \mathbf{m}_t = 0 \tag{1.1}$$

where $\nabla \mathbf{m}$ is the spatial gradient of $\mathbf{m}$, $\mathbf{m}_t$ is the temporal derivative, and $\mathbf{v} = [du/dt\ dv/dt]$ is the velocity vector. If we consider a fixed scene point $\mathbf{m}$, two camera positions is always relative to the scene points. Since the scene points are static and fixed in the world coordinate, the movement of a camera is same to static multiview cameras that the translation $\mathbf{t}$ and rotation $\mathbf{R}$ of a camera is relative to the others.

$$\mathbf{m}(t+1) = \mathbf{R}\mathbf{m}(t) + \mathbf{t} \tag{1.2}$$

SfM methods are equivalent to multi-baseline stereos for a static scene.

Epipolar constraint : It is well-known that two viewpoints for a scene are related by *epipolar constraint* [24] and that multiple stereo views are related by the *multi-linear constraints* [52]. A scene point and the optical centers of two cameras form a plane called epipolar plane. The lines where the epipolar plane intersects the two image planes are the epipolar lines, and the projections of a point must lie on these lines. It satisfies the relationship:

$$\begin{bmatrix} (\mathbf{m'}_i)^T & 1 \end{bmatrix} \mathbf{F} \begin{bmatrix} \mathbf{m}_i \\ 1 \end{bmatrix} = 0 \tag{1.3}$$

where $\mathbf{m}_i$ and $\mathbf{m'}_i$ is the projected points to two views, and $\mathbf{F}$ is 3×3 matrix known as *fundamental matrix*. Such constraints can be used to estimate the *camera geometry*, e.g. translation and rotation parameters, and calculated by a linear method such as the 8-point algorithm (i.e. see Appendix A.5).

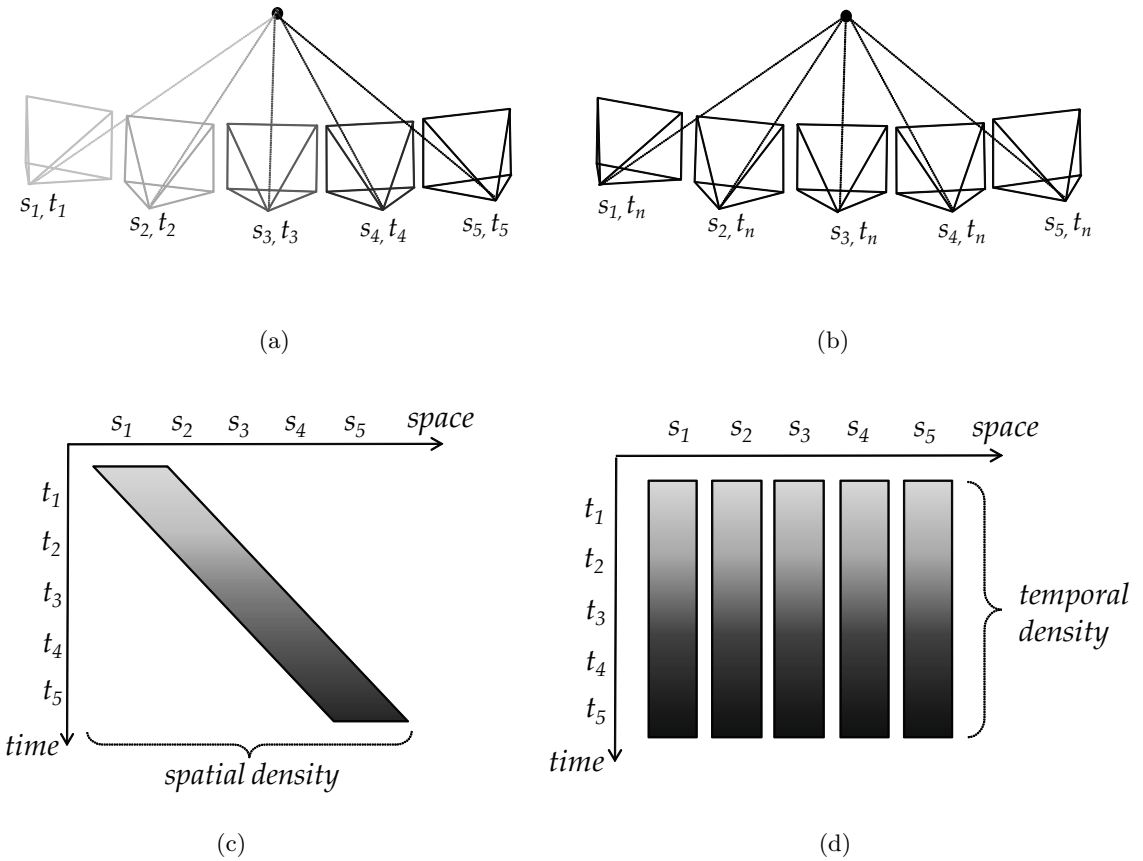


Figure 1.4. Relationship between SfM and multi-stereo methods - (a) illustration of SfM method, (b) illustration of multi-stereo method, (c) spatio-temporal notation for SfM method, (d) spatio-temporal notation for multi-stereo method

Spatial and temporal camera density : In general, the denser capturing of multiview images with a large number of cameras provides the more precise 3-D representation. The SfM (Structure-from-Motion) method is easier to obtain the spatial camera density than the multi-stereo method as Figure 1.4(c) shows. However, a SfM cannot handle a temporally independent moving object[6] since the positions of objects is not relative to the positions of a moving camera. Moreover, a particular motion analysis is very noise sensitive. The multi-stereo method can provide the relative 3-D positions of moving objects through temporal observations of static cameras as Figure 1.4(d) represents. Stereo methods provide more reliable 3-D information since the difference of stereo viewpoints is relatively larger than the viewpoint difference of camera motion for a 3-D point. However, a stereo method is difficult due to the ambiguity in the matching process of discrete pixels and the unbalanced camera

6) The handling of an independently moving object is known as the non-rigid motion problem.

brightness. If more than two cameras are fixed, and the camera poses are known, the SfM method can be cooperated with a multi-stereo method and provide complementary information to obtain 3-D scene structure. When a set of 3-D points $\mathbf{M}_i$ at a frame t is projected onto an image $I_i(t)$ captured by n viewpoints, the point's coordinates $\mathbf{M}_i^{I_n}(t) \sim \mathbf{m}_i(u, v)$ in the image coordinate system can be expressed in terms of a vector-valued nonlinear function g:

$$\mathbf{M}_i^{I_n}(t) = \begin{bmatrix} u \\ v \end{bmatrix} = g\left(I_n, \mathbf{M}_i(t)\right) \tag{1.4}$$

Given a set of feature points $\mathbf{m_i}$ taken by n viewpoints I_n (i.e. structure-from-stereo) or by different time frames t (i.e. structure-from-motion), we can reconstruct the points $\mathbf{M}_i(t)$ from a set of the projections $\mathbf{M}_i^{I_n}(t)$. In the past few years, a lot of attempts have been made to use multi-view cameras or a monocular camera in motion for a complex 3-D scene modeling. These attempts are made with many viewpoints from 3 to 100 camera positions to generate overlapping point clouds.

Fundamental problems : There are five fundamental problems to extract information from stereo and motion data. Our method addresses these problems.

(a) Image motion and disparity with an unknown camera translation allow us to infer object range only up to a scale ambiguity since image motion and disparity depend on the ratio of camera translation to object range. This problem can be solved by the robust camera geometry estimation in Chapter 3 and 4. The camera geometry with minimum distortions in collineation is estimated by a nonlinear bundle adjustment method. Chapter 4 shows that the camera geometry is closely connected by the view geometry, i.e. the epipolar geometry.

(b) Image motion and disparity tend towards zero near the focus of expansion (FOE)[7]. Since the object range is inversely proportional to image motion and disparity, the scene structure estimation is ill-conditioned near the focus of expansion. We constrain the direction of the parallax field to lie along the epipolar direction [93] and address this problem by the sceneflow estimation, i.e. joint disparity and optical flow estimation, described in Chapter 6.

(c) When more closely aligned the local image structure is with the epipolar directions, i.e. directions pointing towards the focus of expansion, the more ill-conditioned scene structure may be estimated in those regions. If there are some errors in the camera

7) The focus of expansion is the intersection of the translation vector of the camera with the image plane, and a group of motion trajectories is defined to be an image point towards the camera movement. [27] With a positive component of velocity along the optic axis, image points will appear to move away from the FOE and expand e.g. with those closer to the FOE moving slowly and those further away moving more rapidly.

geometry estimation, it causes large errors in the estimation of the scaled relative depth. This problem is solved by using the intersection of spatio-temporal epipolar lines and a robust outlier removal. Chapter 4 introduces the spatio-temporal epipolar geometry, and Chapter 5 describes the robust estimation method.

(d) The spatio-temporal baselines might be defined with respect to a monocular camera in motion[8], a static stereo camera[9], or some other combination of static and non-static cameras[10]. With a prior knowledge for the static and non-static cameras, the proposed method first utilizes the static scene assumption to make a combination of SfM and multi-stereo methods. Then, a multi-stereo method solves independently motion problem that cannot be solved in temporal frames. The results are shown in Chapter 6.

(e) A large baseline gives better depth estimates for distant objects. However, the increased problems of ambiguity and occlusion[11] are the drawbacks. This paper studies the large baseline problem by a feature-based estimation method with a robust statistic. In Chapter 5, we generalize that the all robust methods in 3-D vision can be described by error weights and efficiently remove outliers in feature-based estimates. When the camera geometry is accurately recovered by robustly detected feature points in a set of images, the optical-ray method described in Chapter 3 can combine the projection rays for an arbitrary camera viewpoint. However, the density of depth sampling is a fundamental problem in light field rendering. The proposed method compensates ray distortions according to scene geometry by using the dense 3-D sceneflow method. The optical-ray can be rearranged by depth-compensated image-based rendering as Chapter 4 depicts.

1.4 Proposed Method

The generalized video-based 3-D analysis method to solve the image-based modeling and rendering problems is proposed in this dissertation. Figure 1.5 represents the block diagram of the proposed method.

Camera geometry from uncalibrated cameras : If camera and viewing geometry are unknown[12], but the scene structure remains rigid and piecewisely planar, it is possible to deduce the viewing geometry from a number of spatio-temporal feature matches and to recover the scene structure. For a static scene, 3-D distance between two points $\mathbf{M}_i(t)$ and $\mathbf{M}_j(t)$

8) A monocular camera in motion is related to optical flow and SfM methods.
9) Stereo camera problem is related to disparity and triangulation methods.
10) This dissertation addresses this problem with the 3-D sceneflow method.
11) Visible to one camera but not to the other.
12) The unknown cameras and viewing geometry is the general condition of uncalibrated cameras.

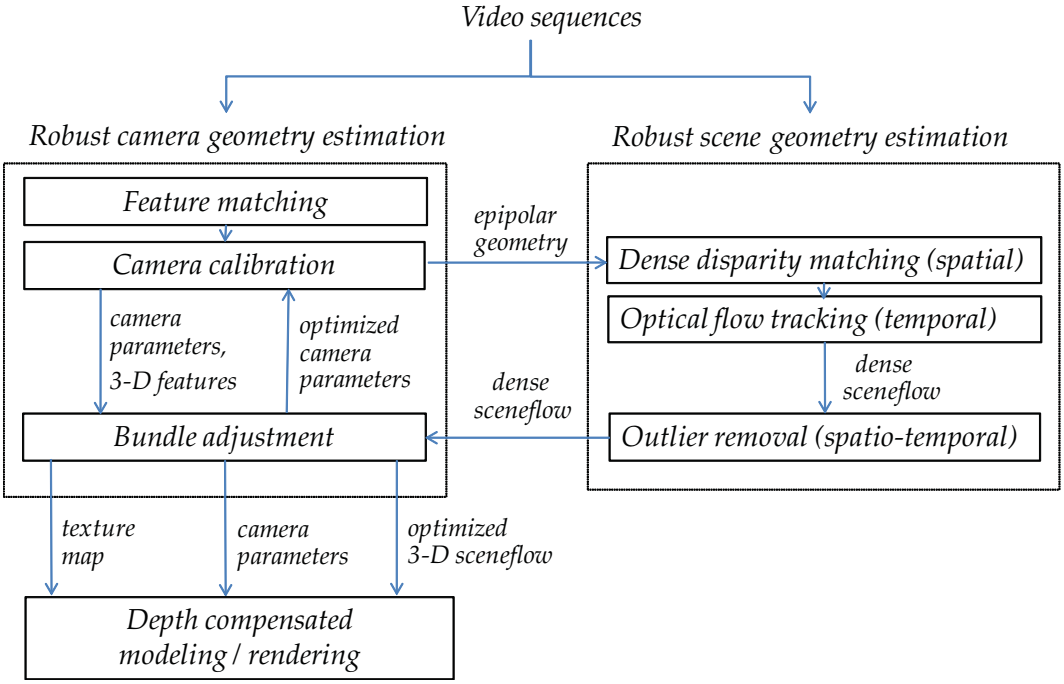


Figure 1.5. Block diagram of the proposed method

derived from the spatial matches should be temporally invariant.

$$\begin{aligned}&\left[\mathbf{M}_i(t)-\mathbf{M}_j(t)\right]^T\left[\mathbf{M}_i(t)-\mathbf{M}_j(t)\right]\\&=\left[\mathbf{M}_i(t+1)-\mathbf{M}'_j(t+1)\right]^T\left[\mathbf{M}_i(t+1)-\mathbf{M}'_j(t+1)\right]\end{aligned}\tag{1.5}$$

Using the projective relationship and the known image coordinates $\mathbf{m}_i(t)$, $\mathbf{m}_j(t)$, $\mathbf{m}_i(t+1)$ and $\mathbf{m}_j(t+1)$, this can be expressed for 4 unknown depth values $\mathbf{M}_i(t)$, $\mathbf{M}_j(t)$, $\mathbf{M}_i(t+1)$ and $\mathbf{M}_j(t+1)$. Viewpoints looking at a rigid structure are related by a set of collineations[13] and the constraints between different sets. Let $\mathbf{T}$ be an anti-symmetric matrix such that $\mathbf{Tx}=\mathbf{T}\times\mathbf{x}$ for all 3-D vectors, 3×3 essential matrix $\mathbf{E}=\mathbf{TR}$ is defined by the relationship

$$\begin{bmatrix}\mathbf{M}_i(t+1)^T & 1\end{bmatrix}\mathbf{P}'^{-T}\mathbf{E}\begin{bmatrix}\mathbf{M}_i(t)\\1\end{bmatrix}\mathbf{P}^{-1}=0\tag{1.6}$$

Collineation in view geometry : A collinearity constraint[14], well known from photogram-

13) A collineation is a one-to-one map from one projective space to another, or from a projective plane onto itself, such that the images of collinear points are themselves collinear. All isomorphisms induce a collineation.

14) Image points and projection centers are located on a straight line. Since an image is exactly a plane, a transformation of the plane transforms collinear points into collinear points. A projective collineation transforms every 1-D form projectively, and a perspective collineation is a collineation which leaves all

metry, can be used to reduced the geometric error in the projected ray [28]. If camera locations and 3-D points are subject to the collinearity constraints that can be expressed by a set of homographies, rays passing through the stereo viewpoint image can be represented by the epipolar geometry. In most cases, the view geometry includes several image distortions according to the projective distortions for the geometric structures. However, the linear view geometry is a basis of the nonlinear *scene geometry estimation*, e.g. disparity, parallax and depth, etc.

3-D sceneflow : Using epipolar constraints, a set of dense motions of rigid structure forms *3-D sceneflow* that can be estimated by the brightness-invariance of spatial correspondences $\mathbf{m}$ and $\mathbf{m}'$ in two images.

$$\mathbf{m}\left(u(t), v(t), t\right) = \mathbf{m}'\left(u'(t), v'(t), t\right) \tag{1.7}$$

and temporally tracked by the time-invariant in two frames t and $t+1$.

$$\mathbf{m}\left(u(t), v(t), t\right) = \mathbf{m}\left(u(t+1), v(t+1), t+1\right) \tag{1.8}$$

The 3-D sceneflow should satisfy the spatio-temporal invariance constraints since they are linked by the collinearity.

$$\mathbf{m}\left(u(t), v(t), t\right) = \mathbf{m}'\left(u'(t+1), v'(t+1), t+1\right) \tag{1.9}$$

Figure 1.6 illustrates the 3-D sceneflow derived by spatio-temporal correspondences $\mathbf{m}(t)$, $\mathbf{m}'(t)$, $\mathbf{m}(t+1)$ and $\mathbf{m}'(t+1)$. The collinearity is defined by the reverse of the projection on the homogeneous coordinates.

$$\mathbf{M} = \mathbf{P}^{-\mathbf{1}} \begin{bmatrix} \mathbf{m}(t) \\ 1 \end{bmatrix}, \quad \mathbf{M} = \mathbf{P}'^{-\mathbf{1}} \begin{bmatrix} \mathbf{m}'(t) \\ 1 \end{bmatrix}, \quad \mathbf{M} = \mathbf{P}'^{-\mathbf{1}} \begin{bmatrix} \mathbf{m}'(t+1) \\ 1 \end{bmatrix} \tag{1.10}$$

The collinearity between observed spatio-temporal correspondences and the camera parameters can be nonlinearly optimized by a bundle adjustment method [19, 29]. Although a feature-based method is suitable to obtain global image warping according to camera geometry, existing feature-based methods cannot produce a dense set of correspondences or depths that is necessary to remove the image distortion according to the geometric structure.

Depth compensated image-based modeling/rendering : In most image-based modeling and rendering techniques, the projective distortion problem can be solved by motion

lines through center and points of axis invariant.

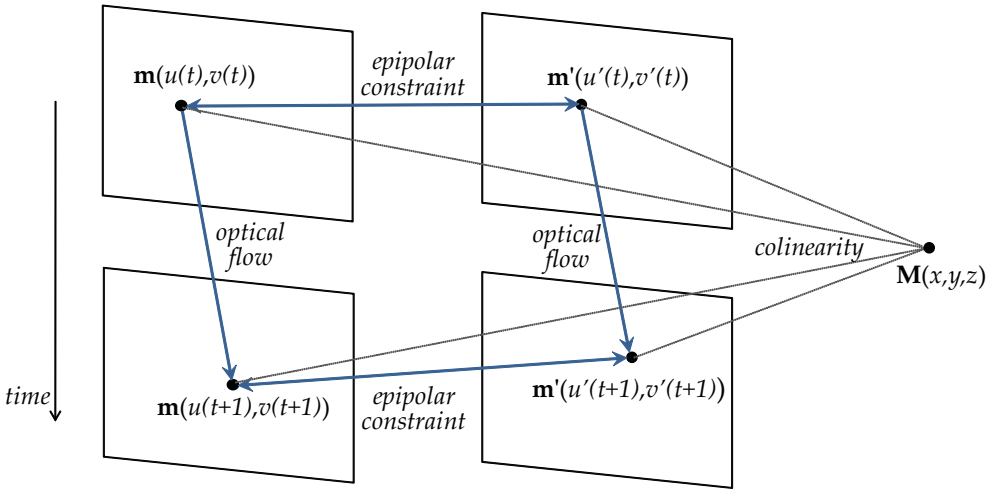


Figure 1.6. 3-D sceneflow method

segmentation or compensation methods. For example, divergent rays are occurred in a single ray projection model for several different depths. However, dense depth estimation [31, 32] include many difficult problems: handling of image noises, textureless regions, depth discontinuities and occlusions. In multi-stereo cases, more constraints can be added than a two camera case, because field of view depending on the placement of cameras can be separated by the matching process [22]. We can easily obtain a dense set of spatio-temporal correspondences, i.e. pixel-wise displacement field, and explicitly handle occlusions in some views with information of other views.

Outlier removal : However, the conventional dense matching methods do not preserve the spatio-temporal discontinuities, and object boundaries are often poorly localized. The proposed method employs an optical flow-based regularization using the robust statistic of error norms [33, 34] to eliminate outliers[15] from camera geometry and scene geometry estimates.

1.5 Background and Contributions

The presented works were carried out within the EU 3DTV project, a Network of Excellence funded by the European Commission, the 6th Framework Information Society Technologies (IST) Programme [91]. The primary objective is to align the interests and efforts of European researchers within a common effective research network for achieving full scale 3-D visual signal handling capabilities. Capturing 3-D visual information of a real scene and creating an

15) Outliers designate points with gross errors existing in the data due to physical imperfections in imaging sensors and errors in a low-level vision computation.

exact optical duplicate of it at a remote site instantaneously or at a later time are ultimate goals in visual communications. All core and peripheral components related to this goal are collectively referred as *3-D Television* (3DTV). Main functional components of 3DTV are

(a) Capture and representation of 3-D scene information,
(b) Complete definition of digital 3DTV signal,
(c) Storage and transmission of the signal,
(d) Display for the reproduced 3-D scene,

and there are numerous challenges in these components. Above all, lack of 3-D video contents is an important bottleneck to make a consumer market. Although there have been a lot of studies [12–15, 22, 31, 32, 39, 40, 42, 45, 46] aiming at a free-viewpoint video for dynamic scenes, they are restricted within indoor scenes or pre-segmented objects. This paper aims to generate automatically a 3-D contents generation from outdoor video scene. A key feature of 3-D contents is interactivity in the sense that a user should have the possibility to choose an arbitrary viewpoint for a real-world scene. Contributions to the functional components of the 3DTV project are summarized as follows:

(a) 3-D scene information: 3-D contents are generated by a set of natural videos. A system which can capture large numbers of videos in real-time is too complex to use in the outdoor environment. For a static scene, a video sequence captured by a non-professional photographer with a hand-held camera, i.e. hand-held video sequences, is very similar to multi-stereo with many baselines. The proposed method captures large multiple viewpoints with a combination of stereo camera and a monocular camera in motion. Multi-stereos gives us a temporal density of viewpoints and the camera movement allows a spatial density of viewpoints with a continuous motion.
(b) Definition of 3-D signal: 3-D signal is defined by stereo motion video, dense 3-D points and translation and rotation parameters of camera geometry. Dense depth reconstruction from multiple viewpoints is a core task in 3-D representation but the quality is still not at a satisfactory level. We exploit a standard feature-based calibration method to estimate the camera geometry, and the dense sceneflow estimation is performed in the spatio-temporal epipolar geometry. A nonlinear bundle adjustment technique is applied to optimize the collinearity between the camera geometry and the scene geometry. With camera motions, dense disparity maps between two cameras can be transformed into dense depths because every dense matches can be inverse projected to the 3-D structure and reprojected again to the reference image using the camera parameters.
(c) Data structure of 3-D signal: The data structure includes information of all rays projected from scene to camera, and the amount of multi-view image data is usually huge. Thus, compression and transmission of the data with less degradation and delay over limited

bandwidth are also challenging tasks. The redundancy in multi-view videos can be eliminated by spatio-temporal correspondences[16]. Similarly, the redundancy of projection rays can be removed by using the collineation between spatio-temporal correspondences in the sceneflow.

(d) Display of 3-D scene: Auto-stereoscopic display devices, that do not require special glasses to view it, need particular light control method to provide unsurpassed image quality and 3-D realism to viewers. For example, auto-stereoscopic display needs several viewpoint images to recover accurate depth feeling. The proposed method can synthesize a particular arbitrary viewpoint or a viewpoint of the user's eye position.

16) Disparity and motion estimation permit to increase the quality without the addition of redundant information. This idea is often employed in multi-view video coding (MVC) to reduce the inter-view redundancy efficiently.

Chapter 2

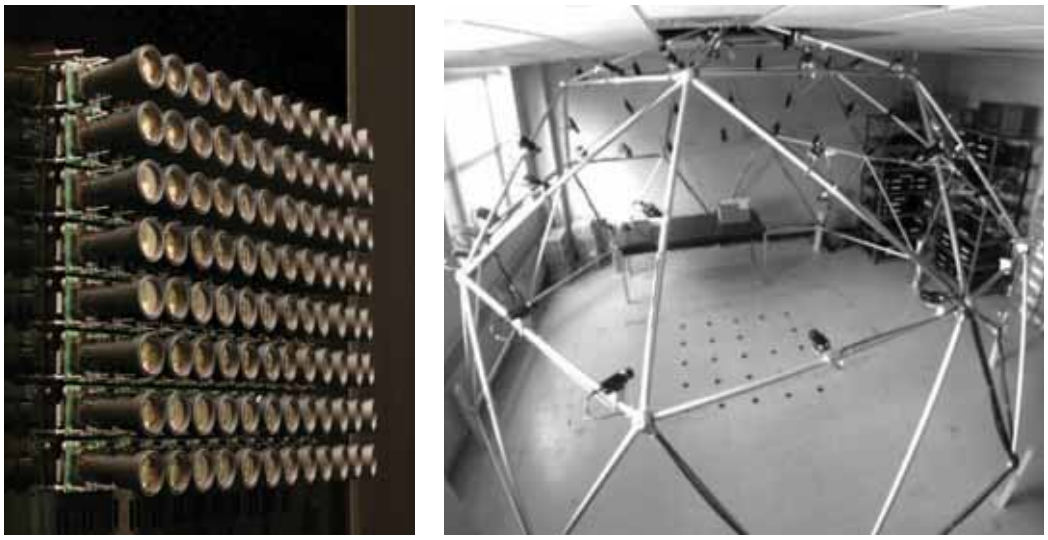
Related Works

We can understand the geometric relationship of viewpoints as camera geometry and scene geometry. The camera geometry is a definition for camera motion or poses that are usually obtained by a feature-based method. Since there are a minimum required number of points to solve the parametric equations, a lot of detected features are eliminated to increase the accuracy of estimated parameters, and thereby only sparse 3-D structure can be recovered in the result. When the camera parameters are known, it is important to recover dense scene structure for image-based rendering and modeling applications. Since the camera geometry allows only global warping information, the dense scene geometry estimation should be additionally performed to increase the 3-D sampling density. This chapter presents a review of related works.

2.1 Multiview Video Analysis

The simplest case of 3-D video is the stereoscopic video that is shot using two cameras with the parallel or toed-in camera configuration. The distance between two cameras is called *baseline*. The multi-view 3-D video is characterized by *depth-dependent displacement* between spatially and temporally successive images. Disparity and motion estimations are the process of computing the relative displacement of objects within a scene. Novel views can be synthesized by image interpolation with disparity and motion compensations [36, 37]. However, most conventional stereo algorithms have a number of significant weaknesses. In particular, the different images need to be color calibrated for reliable results. Moreover, establishing stereo correspondences is hard to accomplish in homogeneous regions with constant luminance or if the scene is not *Lambertian*[17]. In addition, many stereo techniques rely on closely spaced images, and in some cases employ significant user input for each image pair to guide the

17) The term *Lambertian* refers to a condition which reflects light equally in all directions on particular surfaces of objects within a scene. The effect is that the viewer's perspective any rotation, scaling or translation of an object does not result in a change of the shadows and highlights of the object's surface.



(a) (b)

Figure 2.1. Multiview camera configurations - (a) Stanford University Light field Camera with dense camera configuration, (b) Virtualized Reality^{TM} of Carnegie Mellon University with wide camera distribution

stereo process. New approaches have been developed to remedy this situation. The Middlebury Stereo [35] offers a web-based evaluation of recently developed stereo algorithms using a number of objective error metrics [38]. The error metrics cover aspects of reliability as the *Percentage of Bad Matching Pixels* compared to acceptable ground truth. Special attention is given to the types of regions involved, so that textureless, occluded and depth discontinuity regions are evaluated separately. Multiview video is an expansion of stereoscopic videos based on multiple baselines. The viewpoints are featured as *geometrically calibrated* and *temporally synchronized* video data. The camera setups range from dense configuration e.g. Stanford's Light field Camera [17] to intermediate camera spacing [69] and to wide camera distribution e.g. Virtualized Reality^{TM} [39] of Carnegie Mellon University. Figures 2.1(a) and 2.1(b) respectively represent Stanford's Light field Camera and Virtualized Reality^{TM}. The wider spacing between the cameras is more of a challenge in producing locally consistent geometries and thereby photorealistic views due to significant extended occlusions. Large video camera arrays [40, 42] and densely packed camera arrays [41, 43] easily enable users to view a distant 3-D world freely. A significantly denser camera configuration such as that of the Stanford's Light field Camera allows *synthetic aperture tracking* which projects images of a scene from different views on to a virtual focal surface, enabling us to see through occlusions in the scene.

2.2 Image-based Modeling and Rendering

Image- (and video-) based modeling and rendering are terms used to describe computer vision methods which generate novel views using a set of 2-D images (or videos) of a scene. Image- (and video-) based modeling utilizes a combination of visual-hull and multi-stereo matching to reconstruct a 3-D model, and renders the model for an arbitrary viewpoint. This is often called *Free-Viewpoint Video*. Since the density point cloud obtained by viewpoint matching is normally low, 3-D mesh generation or surface fitting is additionally considered.
Image- (and video-)based rendering method is based on a *ray-space representation* using camera calibration parameters, and virtual viewing of a scene is possible by warping the original images into the correct perspective given a viewing direction. Although dense sampling permits photorealistic rendering with just either a simple planar or rough geometric approximation, the disadvantage is the large number of images required for rendering and the viewing range limited near the camera path.
Other methods may utilize the concept of *multi-video-plus-depth* data representation [44]. Approaches in intermediate camera spacing are trying to reduce the number of required cameras and needs to compensate geometric aperture and parallax in depth. In [69], Zitnick et al. proposed a *Layered Depth Image* (LDI) representation using an 8 cameras configuration. However, this method still needs a quite dense camera setup for a limited viewing range, i.e. horizontal field of view of about 30°. In [46], sparsely placed and scalable *3-D video bricks*, which act as low-cost Z-cameras, are generated by scalability in terms of camera configurations. One single brick consists of a projector, two grayscale and one color camera to acquire stereo images, structured light, and texture images simultaneously. Then, a multi-window based matching algorithm with a subsequent sub-pixel disparity refinement was used to compute accurate disparity maps. About ten 3-D video bricks are needed to fully cover 360° in all dimensions. In [45], a dynamic point sample framework, called 3D video fragments, is proposed for real-time free-viewpoint video. By generalizing 2-D video pixels towards 3-D irregular point samples, this approach uses a 2-D video update scheme exploiting the spatio-temporal coherence of the video streams of multiple cameras for more complex 3-D model representations.

2.3 Camera Geometry Estimation

Camera geometry is the relationship between 3-D world coordinates and their corresponding 2-D image coordinates which can be established by a camera calibration. Once the geometry is established, 3-D information can be inferred from 2-D information and vice versa. Accurate camera calibration are a necessary prerequisite to extract precise and reliable 3-D metric information from images. A fundamental criterion for camera geometry estimation is based on the perspective and projective camera models:

(a) Perspective camera model (nonlinear): A camera model based on *perspective collineation*, where the implication is that the interior orientation is stable, and all departures from collinearity, linear and nonlinear can be accommodated. The collinearity-based model generally requires five or more point correspondences within a multiple image network. Due to its nonlinear nature, the model requires approximations for parameter values for the least-squares bundle adjustment in which the calibration parameters are recovered.

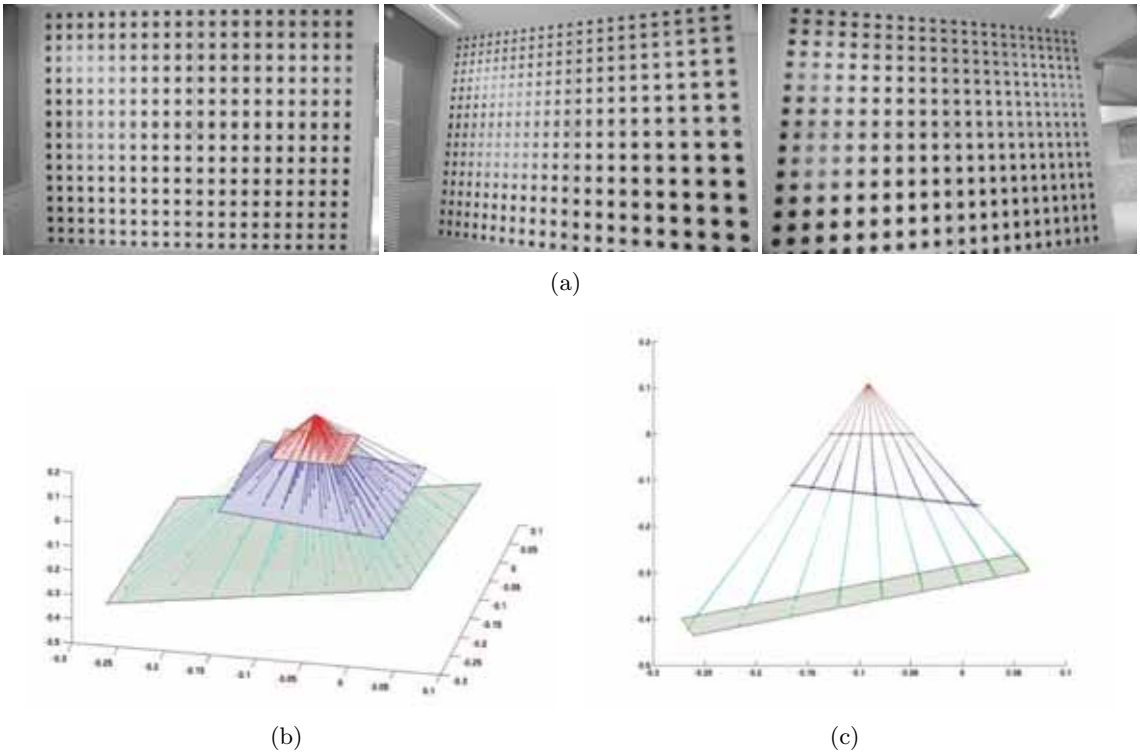
(b) Projective camera model (linear): A projective camera model supporting projective rather than Euclidean scene reconstruction. Such a model, characterized by the Essential matrix and Fundamental matrix models, can accommodate variable and unknown focal lengths, but needs a minimum of 6 point correspondences to facilitate a linear solution, which is invariably quite unstable[18]. Nonlinear image coordinate perturbations such as lens distortion are not easily dealt with in such models.

In an application involving multiple cameras, the calibration is necessary to guarantee geometric consistency across the different cameras. The parameters are categorized by intrinsic and extrinsic parameters respectively determining the internal camera geometric and optical characteristics, and the 3-D position and orientation of the camera frame relative to a certain world coordinate system.

Photogrammetric calibration : Photogrammetric calibration attracts less attention today since many respects of the method reached maturity in the mid 1980s. These are all based on the pinhole camera model and include terms for modeling radial distortion. Tsai's calibration method [47] requires $n > 8$ features points per image and solves the calibration problem with a set of n linear equations based on the radial alignment constraint. A second order radial distortion model is used while no decentering distortion terms are considered. The two-step method can cope with either a single image or multiple images of 3-D or a planar calibration grid shown in Figure 2.2(a) but grid point coordinates must be known. The model of Heikkila and Silven [48] first extracts initial estimates of the camera parameters using a closed form solution known as direct linear transformation (DLT), and then a nonlinear least-squares estimation employing a the Levenberg-Marquardt algorithm[19] is applied to refine the interior orientation and compute the distortion parameters. The model includes two coefficients for radial and decentering distortions and works with single or multiple images and with 2-D or 3-D calibration grids. Zhang's calibration method [49] requires a planar checkerboard grid to be placed at more than two different orientations in front of the camera. The method extracts corner points of the checkerboard pattern to compute a projective transformation between the image points of the different images, up to a scale factor. Afterwards, the camera

18) The equations often need normalization

19) The Levenberg-Marquardt algorithm provides a numerical solution to the problem of minimizing a function, generally nonlinear, over a space of parameters of the function. This minimization especially arise in least squares curve fitting and nonlinear programming.



(a)

(b) (c)

Figure 2.2. Collineation in camera calibration pattern - (a) Photogrammetric calibration using checkerboard pattern, (b) collineations of observed patterns, (c) 2-D view of collineations [28]

interior and exterior parameters are recovered using a closed-form solution, while the 3rd- and 5th-order radial distortion of a lens as a function of the radial distance on the lens surface are recovered within a linear least-squares solution. A final nonlinear minimization of the reprojection error, solved by a Levenberg-Marquardt method, refines all the recovered parameters. Finally, the camera parameters with minimum distortions in collineations can be obtained as Figure 2.2(b) and 2.2(c) show. However, the collineations are not always satisfactory because the camera parameters and image locations are known only approximately.

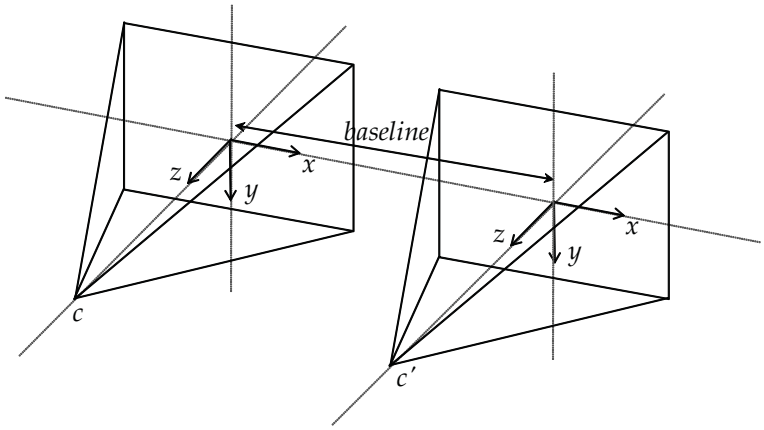
Self-calibration : Recent researches have focused on the problem of *self-calibration*. Self-calibration methods do not use any calibration grid and checkerboard pattern. The metric properties of the camera, i.e. determined up to an arbitrary Euclidean transformation and a scale factor, are recovered from a set of uncalibrated images, using constraints on camera parameters or on the imaged scene. In general, three types of constraints are applied to self-calibration: scene constraints, camera motion constraints, or constraints on the camera intrinsic parameters. In the case of an unknown camera motion and unknown scene, only constraints on the interior orientation [50, 51] can be used. The majority of self-calibration algorithms treat intrinsic camera parameters as constant but unknown [53–56, 56, 57]. Just by

moving a camera in a static scene, the rigidity of the scene provides in general two constraints on the cameras. If images are taken by the same camera with fixed internal parameters and correspondences between three images are sufficient to recover both the internal and external parameters, it is possible to reconstruct 3-D structure up to a similarity. The perspective projection can be applied to model collinearity in the calibration process. Geometric constraints are applied and projective structures are stratified to Euclidean ones. The parameters of the nonlinear distortion are computed by means of the bundle adjustment and extended by additional parameters functions that are supposed to model the systematic image errors. The basic mathematical model of bundle adjustment is provided by the nonlinear optimization, usually extended by correction terms for the interior orientation and radial and decentering lens distortion [58, 59]. The bundle adjustment method provides a simultaneous determination of all parameters along with estimates of the precision and reliability of the extracted calibration parameters. Correlations between the interior orientation and exterior orientation parameters, and the object point coordinates, along with their determinability also can be quantified [19, 29].

2.4 Scene Geometry Estimation

Multiview correspondence, or multiple views matching, is the fundamental problem of determining which parts of two or more views are projections of the same scene geometry. The output is a *disparity map* for each pair of cameras, giving the relative displacement, or disparity, of corresponding image elements. Disparity maps allow us to estimate the 3-D structure of the scene and the geometry of the cameras in space. A variety of constraints are used to obtain exact correspondences in complex image features. If camera calibration parameters are known, epipolar constraints can be used to reduce the matching process for finding corresponding points from 2-D searching to 1-D searching along corresponding epipolar lines. Image pairs are warped so that epipolar lines coincide with the image scan lines, and the pairwise disparity estimation establishes image correspondence between adjacent rectified image pairs and independent depth estimates for each camera viewpoint. The performance depends on the size of the baseline. Figure 2.3(a) represents a stereo camera setup with a baseline, and Figures 2.3(b) and 2.3(c) compare narrow baseline and wide baseline using the interpolated images of two views.

Characteristic of different baselines : Small baseline stereo has viewpoints where the baseline is much smaller than the observed average scene depth. This configuration is usually valid for hand-held image sequences or viewpoints taken by many densely located cameras. The advantages (+) and disadvantages (−) are



(a)

(b)

(c)

Figure 2.3. Stereo camera baselines - (a) stereo camera setup with a baseline, (b) narrow baseline, (c) wide baseline

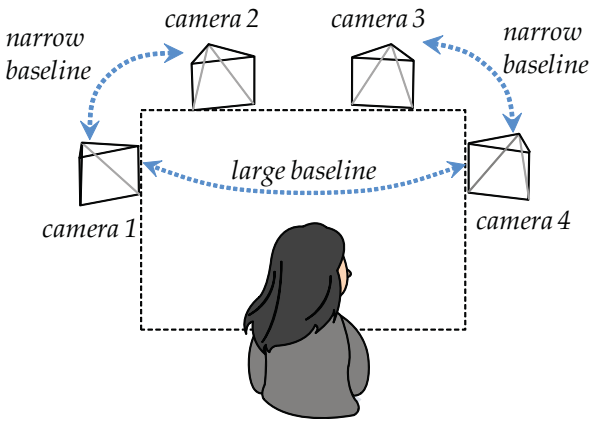
(+) easy correspondence estimation, since the views are similar,
(+) small regions of viewpoint related occlusions,
(−) small triangulation angle, hence large depth uncertainty.

The wide baseline stereo in contrast has the following features:

(+) large triangulation angle, thereby high depth accuracy and resolution,
(−) difficulty of correspondence estimation, since the views are not similar,
(−) large regions of viewpoint related occlusions.

Characteristic of multi-view baselines : A flexible multi-view scheme is needed to combine the advantages of narrow baseline and wide baseline stereo. Small baseline stereos in multiple viewpoints can link together and form a wide baseline stereo. The depth resolution is increased through the combination of multiple viewpoints and large global baseline while the matching is simplified through the narrow local baselines. The VIRTUE system

(a)



(b)

(c)

(d)

Figure 2.4. Combination of wide and narrow baselines - (a) the VIRTUE system [96], (b) a combination of wide and narrow baselines in the VIRTUE system, (c) a warped view using pre-calibrated camera parameters, (d) another warped view

[60] shown in Figure 2.4(a) is an example. Four cameras can only be positioned around a large plasma screen where the width of the screen has a large baseline. A stereo pair located at left (or right) corner of the screen has narrow baselines which is easier to generate a dense disparity map. Figure 2.4(b) illustrates the large and narrow baselines. For the large baseline, the system just use a warping method using pre-calibrated camera parameters as Figures 2.4(c) and 2.4(d) show. Since the dense disparity map can be inversely warped, i.e. back-projected, into the original cameras, this method may produce a dense depth map which allows a viewpoint fusion. The fusion of baselines has a number of advantages as follows.

(+) very dense depth maps for each viewpoint,
(+) high depth resolution through viewpoint fusion,
(+) small viewpoint dependent occlusions,

(+) texture enhancement or anti-aliasing.

Scene geometry estimation methods : The quality of results largely depends on the performance of dense disparity estimation. There are three related topics to scene geometry estimation as follows.

(a) Sparse disparity estimation: Determining a sparse set of correspondences among the images is usually performed as the first step in order to retrieve the structure of the scene and the motion of the camera, when nothing about the geometry of the imaging system is known yet and no geometric constraint can be used in order to help the search. Feature matching is frequently used to detect feature points independently in the two images, then match them using relational structure [63], maximal clique detection [64], or topology [66]. If it is difficult to detect features, template matching can be utilized to select templates or patches with some texture information in an image and then look for corresponding points in the other image using an image similarity (or dissimilarity) metric [61, 62]. Matching image features or templates are reconstructed through triangulation between two views. Features or templates that are also observed in the third view can then be used to determine the pose of this view in the reference defined by the two first views. The initial reconstruction is then refined and extended through a global least-squares minimization of all reprojection errors.

(b) Dense stereo matching: This is a concept that establishes correspondences of the whole pixels between two images. The matching image points must satisfy geometric constraints imposed by the algebraic structures such as the fundamental matrix for two views. A dense disparity map enables segmentation by depth discontinuity [68], which can be useful for layered scene representation in a detailed reconstruction of the 3-D shape. Physical and photometric constraints [67] for ordering of neighboring pixels[20], checking smoothness[21] [70], uniqueness of the match[22] and bidirectional uniqueness[23] [69], and detecting occlusions [71] can be added. These constraints are used to guide the correspondence towards the most probable scan-line match. The absence of transparent objects allows the use of disparity gradient limits, and the absence of occlusion can permit strong surface smoothness constraints. The matcher searches at each pixel in an image for maximum normalized cross-correlation in the other image by shifting a small measurement window along the corresponding scan-line. Some algorithms [72, 73] employs an adaptable size window or pyramidal estimation scheme to reliably deal with very large disparity ranges. Although large-baseline dense stereo can be of great importance for some virtual environment applications, large-baseline makes the correspondence more difficult [74].

20) If two points in two images match, then matches of nearby points should maintain the same order.
21) The disparities should change smoothly around each pixel.
22) Each pixel cannot match more than one pixel in any of the other images.
23) Left-to-right disparity and right-to-left disparity should be matched for a fixed 3-D position.

(c) Spatio-temporal stereo: This method improves the quality of spatial stereo using the spatio-temporal consistency of correspondences [76]. Due to the limited accuracy of geometric reconstruction and correspondence between views, conventional stereo algorithms often results in loss of visual quality compared to captured video and thereby visual artifacts. In a scene with static geometry that is viewed for multiple frames across time, correspondences between two frames can be established. Vedula et al. [75] introduced a sceneflow method, an extension of optical flow to 3-D volumetric reconstruction, to estimate temporal correspondence based on photo-consistency. Cheung et al. [77, 78] introduced a new representation of visual-hull which directly incorporated color for temporal matching to increase the effective number of cameras.

Chapter 3

Camera Geometry Estimation

To understand the 3-D camera geometry from a combination of 2-D videos captured by static and non-static cameras, we first introduce the basic concepts of projective geometry and the pinhole camera model. A pinhole camera is represented as a small hole through which light travels; an intensity of an object is formed on the camera's image plane through perspective projection. In order to determine how 3-D objects in the world geometrically appear in two-dimensional camera images, we need to define 3-D camera geometry in which to represent these objects: the world, camera and image coordinate systems. We apply the camera geometry to the plenoptic sampling using a ray projection.

3.1 Projective Geometry Estimation

The 3-D space we live in is well described as Euclidean in nature, and the definitive model of invariant displacement is *Euclidean geometry* (sometimes called parabolic geometry). For an n-D Euclidean space $\mathcal{R}^n$, points are expressed using n numbers, each giving a coordinate position in each dimension. In 3-D Euclidean geometry, the sides of objects have lengths, intersecting lines determining angles between them. Two lines are said to be parallel if they lie in the same plane and never meet. Moreover, these properties do not change when transformations in Euclidean space, i.e. translation and rotation, are applied as

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \alpha \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \tag{3.1}$$

where r_{ij} are coefficient of an orthonormal matrix, t_i is a translation. α denotes arbitrary scaling which is not considered when dealing with a pure Euclidean transformation instead of a metric one. The relative distance between each two points is expressed by a metric space where the distance shortest between adjacent points is given by a metric. Equation 3.1 can

be expressed by a matrix form:

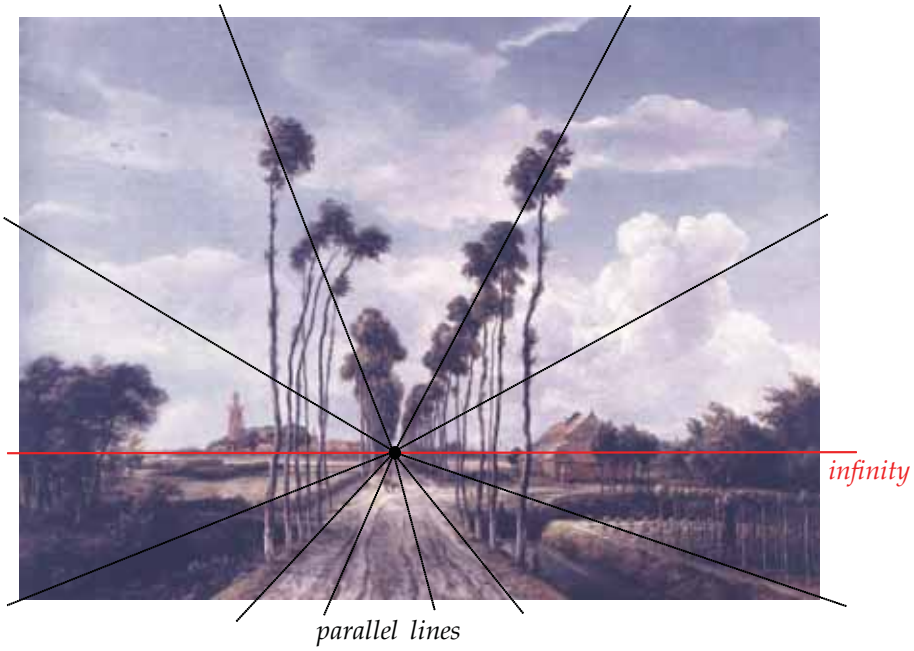
$$\mathbf{T}_{\mathcal{R}} = \begin{bmatrix} \alpha r_{11} & \alpha r_{12} & \alpha r_{13} & t_x \\ \alpha r_{21} & \alpha r_{22} & \alpha r_{23} & t_y \\ \alpha r_{31} & \alpha r_{32} & \alpha r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.2}$$

However, when we observe the 3-D world using our eyes, we do not see the world in a Euclidean form due the effect of perspective projection: that parallel lines meet at infinity. Figures 3.1(a) and 3.1(b) show the perspective effect using a 2-D picture and a 3-D sphere, respectively. When a Euclidean space is projected, some information of the world is visually lost. Hence, Euclidean geometry is insufficient for image processing with a camera (similarly to the case of an eye). Every point in an image represents a possible line of sight of an incoming light ray and any 3-D point along the ray projects to the same image point. In the image, lengths and angles are no longer preserved, and parallel lines may intersect. In Figure 3.1(a), parallel lines on the road converge at a point infinitely far away. The plane on which the road lies intersects the set of infinite points at the horizon line.

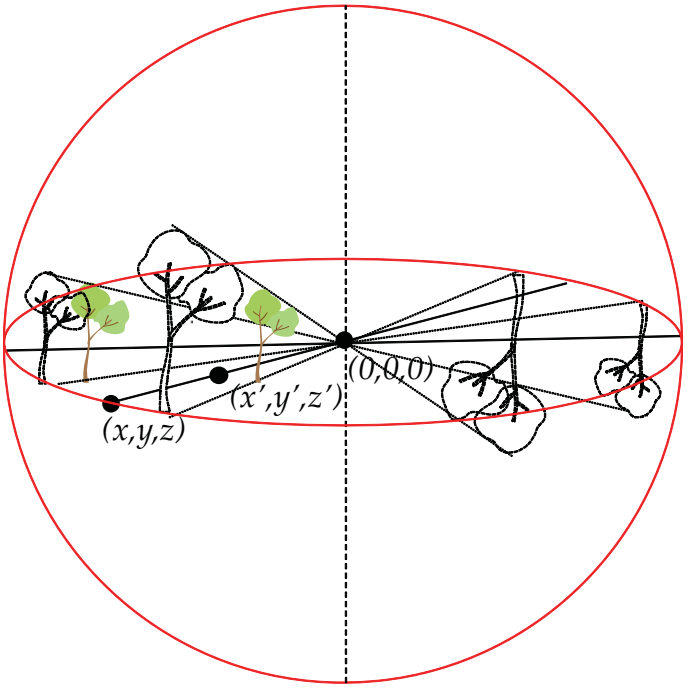
Projective geometry : It is a non-Euclidean geometry that formalizes the perspective art. It can be thought of as an extension of Euclidean geometry in which the direction of each line is subsumed within the line as an extra point, and in which a horizon of directions corresponding to coplanar lines is regarded as a line. Thus, two parallel lines towards the same direction will converge to a point infinitely far way on a horizon. The convergence at infinite distance serves as a definition of parallelism. Figure 3.1(b) illustrates the parallelism using a 3-D sphere. A set of all lines in the space is passing through the origin $(0, 0, 0)$. Two points (x, y, z) and (x', y', z') are equivalent if there is a nonzero real number α such that $x = \alpha x', y = \alpha y', z = \alpha z'$. Idealized directions are referred to as points at infinity, while idealized horizons are referred to as lines at infinity. Euclidean geometry is a subset of projective geometry. Projective geometry exists in any number of dimensions, just like Euclidean geometry. There are two geometries between them; affine and metric.

$$\text{Projective} \subset \text{Affine} \subset \text{Metric} \subset \text{Euclidean} \tag{3.3}$$

Hierarchy of geometry : Starting from Euclidean geometry in the hierarchy shown in Equation 3.3, each geometry belongs to the upper geometry. Each successive geometry has a less rigid space and hence also more transformations which leave the space invariant. Table 3.1 shows the allowed transformations of each geometry in the hierarchy [89]. Projective geometry allows a much larger group of transformations than just translations and rotations, a group which includes perspective projections. Of course, the drawback is that it preserves only the



(a)



(b)

Figure 3.1. Perspective projection and parallelism - (a) an example[88] of perspective projection, (b) illustration of the parallelism using a 3-D sphere

Table 3.1. Measures that remain invariant under the transformations in the hierarchy of geometries

	Geometries	Euclidean	Metric	Affine	Projective
Transformations	rotation	◯	◯	◯	◯
	translation	◯	◯	◯	◯
	uniform scaling	·	◯	◯	◯
	nonuniform scaling	·	·	◯	◯
	shear	·	·	◯	◯
	perspective projection	·	·	·	◯
	composition of projections	·	·	·	◯
Invariants	length	◯	·	·	·
	angle	◯	◯	·	·
	ratio of length	◯	◯	·	·
	parallelism	◯	◯	◯	·
	incidence	◯	◯	◯	◯
	cross ratio	◯	◯	◯	◯

properties of Euclidean structure which remains invariant to projection. However, projective transformations preserve type (i.e. points remain points and lines remain lines), incidence (i.e. whether a point lies on a line), and a measure known as the cross ratio. The projective line, which is denote by $\mathcal{P}^1$, is analogous to a 1-D Euclidean world $\mathcal{R}^1$. The projective plane $\mathcal{P}^2$ corresponds to the Euclidean plane $\mathcal{R}^2$ and the projective space $\mathcal{P}^3$ is related to 3-D Euclidean space $\mathcal{R}^3$. The visual imaging using human eyes or cameras is represented by a projection from $\mathcal{P}^3$ to $\mathcal{P}^2$, from 3-D space to the 2-D image plane.

Affinity : An affine geometry $\mathcal{A}$ is a geometry in which properties are preserved by parallel projection from one plane to another. Affine transformations are a subgroup of projective transformations which can be represented by homogeneous notation. Geometrically, affinities preserve collinearity. An affine transformation has the following form, which is the general form of Equation 3.1.

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} a_{14} \\ a_{24} \\ a_{34} \end{bmatrix} \tag{3.4}$$

Equation 3.4 can be represented by a single matrix form.

$$\mathbf{T}_{\mathcal{A}} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.5}$$

It transforms parallel lines into parallel lines and preserve ratios of distances along parallel lines. In the same manner, an affine transformation must be a projective transformation which maps the infinite plane Π_∞ to the infinite plane, i.e. $\mathbf{T}_{\mathcal{A}}\Pi_\infty = \Pi_\infty$. A mapping from projective space to affine space in $\mathcal{P}^3$ can be determined as

$$\mathcal{P}^3 \to \mathcal{A}^3 : (x, y, z, w) \to \left(\frac{x}{w}, \frac{y}{w}, \frac{z}{w}, 1\right) \tag{3.6}$$

where $w \neq 0$. If the camera is at the origin $(0, 0, 0)$, the ray represented by homogeneous coordinates (x, y, z, w) is that passing through the 3-D point (x, y, z). $(x, y, z, 0)$ is a 3-D point that defines a perfectly normal optical-ray. This ray does not correspond to any finite pixel since it is parallel to the plane $w = 1$ and thus has no finite intersection with it. The homogeneous vector $(x, y, z, 0)$ represents ideal points at infinity for each 3-D direction in 3-D projective space.

$$\lim_{w\to 0}\left(\frac{x}{w}, \frac{y}{w}, \frac{z}{w}, 1\right) \sim \lim_{w\to 0}(x, y, z, w) = (x, y, z, 0) \tag{3.7}$$

The infinite plane is infinite points of the world $(x, y, z, 0)$ and it should be always fixed for all 3-D points (x, y, z, w) in $\mathcal{P}^3$. Accordingly, the plane can be defined by a coordinate of 0: $\Pi_\infty = (0, 0, 0, 1)$. Points on Π_∞ are ideal points of the form $(x, y, z, 0)^T$ satisfying an equation as

$$(0, 0, 0, 1)^T(x, y, z, 0)^T = 0 \tag{3.8}$$

n-D projective space : In an n-D projective space $\mathcal{P}^n$, the homogeneous and projective coordinate can be represented by a point with $n + 1$ coordinate vectors $\mathbf{x} = (x_1, \cdots, x_{n+1})^T$ which is satisfied with the following condition.

$$^{\forall}\mathbf{x} \in \mathcal{P}^n, \alpha \neq 0 \Rightarrow \alpha\mathbf{x} = \mathbf{x} \tag{3.9}$$

where α is the scale factor with non-zero scalar. The homogeneous coordinates of a vector $\mathbf{x} = [x, y, \cdots]^T$ are denoted by $\tilde{\mathbf{x}}$, i.e., $\tilde{\mathbf{x}} = [x, y, \cdots, 1]^T$. For example, a point (x, y) of the

Euclidean plane $\mathcal{P}^2$ in the projective plane can be represented by adding a third coordinate of 1 at the end: $(x, y, 1)$ since the point $(x, y, 1)$ is the same as the point $(\alpha x, \alpha y, \alpha)$ for any non-zero α. The equation is unaffected by scaling and hence the coordinates are called the homogeneous coordinates of the point. A linear transformation of an n-D projective space $\mathcal{P}^n$ into itself is represented by $(n+1)\times(n+1)$ matrix $\mathbf{A}$ where $det(\mathbf{A}) \neq 0$ denotes invertible i.e. $\mathbf{A}^{-T} = (\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1}$. Similarly, a projective mapping from $\mathcal{P}^m$ to $\mathcal{P}^n$ is expressed by $(m+1) \times (n+1)$ invertible square matrix.

3.2 Pinhole Camera Model

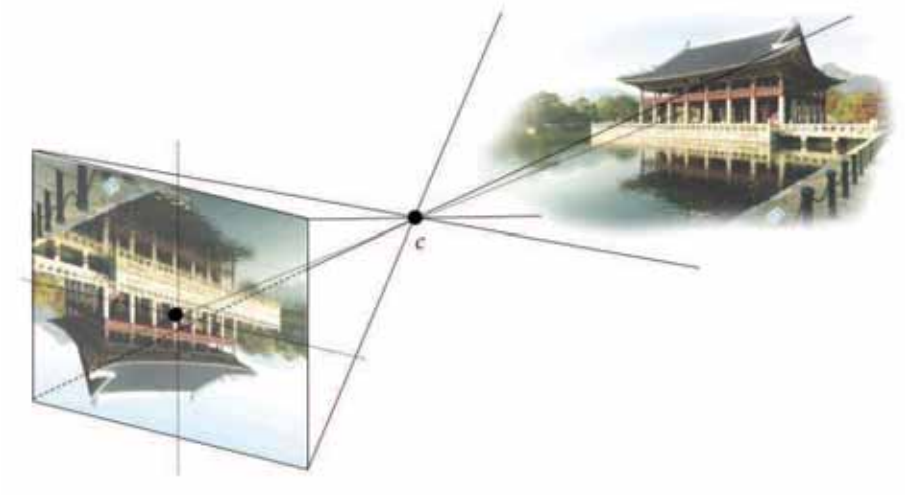
A lot of models of a camera, depending on the type of physical and optical property, exist. However, in computer vision, a lens-based camera is widely described by a pinhole model due to the simplicity[24]. The pinhole-camera is the simplest approximation of a perspective and the perspective camera model corresponds to an ideal pinhole camera. The geometric process for image formation in a pinhole camera is represented in Figure 3.2. Two labeled planes I and F are respectively referred to as the *retinal plane* or *image plane*, and the *focal plane*. Image formation involves the projection of scene points $\mathbf{M} = [x, y, z]^T$ in $\mathcal{P}^3$, i.e. world points, to image points $\mathbf{m} = [u, v]^T$ in $\mathcal{P}^2$, i.e. retinal plane. The line passing through the optical center c and orthogonal to the retinal plane I is called as the *optical axis* or *principal ray*. The intersection of the principal axis and the image plane is the *principal point*. The projection of scene points $\mathbf{M}$ is represented by a line passing through the *optical center* c at a distance f (known as the *focal length*) and a point on the retinal plane I. The *normalized image plane* is coplanar with the plane I and it is located at a unit distance (i.e. $f = 1$) from the optical center c. It can be written as a linear mapping from 3-D to 2-D.

$$\begin{bmatrix} u \\ v \\ f \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{3.10}$$

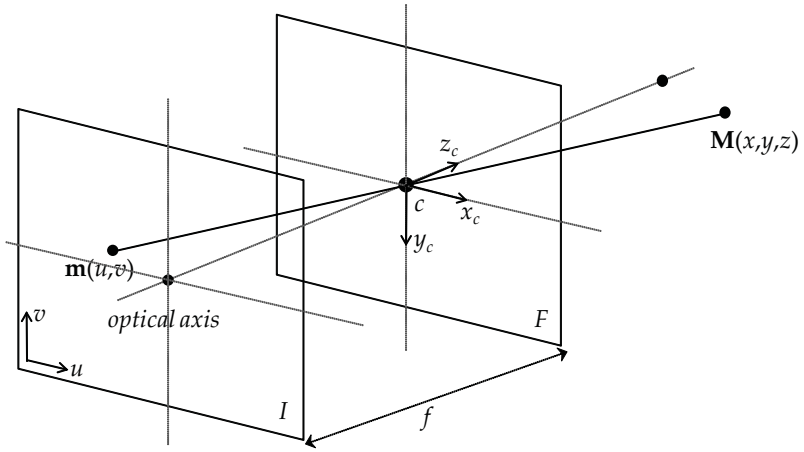
If a point $\mathbf{M}$ is at the optical center c, i.e. $z = 0$, this projection process is undefined. The focal length f of the camera is not necessary to be 1 and different values of f just correspond to different scaling of the images on optical axis (i.e. z-direction) as $\alpha \propto f/z$. The perspective projection from $\mathbf{M}$ to the image plane as point $\mathbf{m}$ is represented as

$$u = f\frac{x}{z} \quad \text{and} \quad v = -f\frac{y}{z} \tag{3.11}$$

24) Chapter 7 will introduce a confocal constraint in a real lens model to recover a pin-hole camera model for stereo analysis.



(a)



(b)

Figure 3.2. Pinhole camera model - (a) ray projection in a pinhole camera, (b) pinhole camera model

The new system into which the image coordinates **m** are to be transformed need not to be orthogonal. If $f = 1$ is assumed as different values of f corresponding to different scaling of

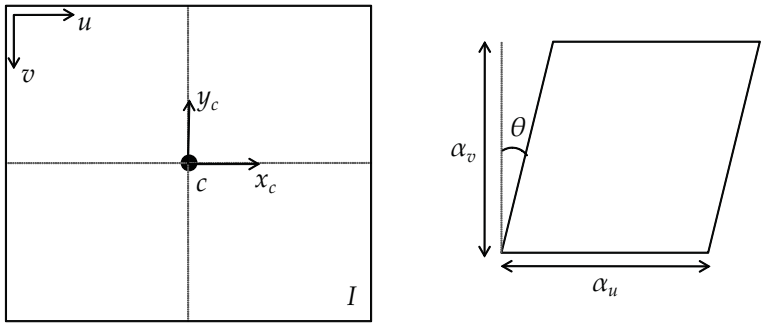


Figure 3.3. Camera calibration and skew factor

the image, the above equations are represented using homogeneous coordinates.

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{3.12}$$

Camera calibration matrix : In real images, the origin of the image coordinates is not the principal point and the scaling along each image axis is different, thus a transformation on the F plane (x_c, y_c) depending on focal length f is applied to take the center to the origin. Image distortion is caused when the intersection of the optical axis with the retinal plane is not at the optical center c. The transformation (x_c, y_c) is applied to the ideal image coordinates to obtain the actually observed image coordinates (u, v). 3×3 upper triangular matrix $\mathbf{C}$, called *camera calibration matrix* (or intrinsic camera parameter) provides the transformation between an image point and a ray in Euclidean 3-space. Parameters in the matrix $\mathbf{C}$ do not depend on the position and orientation of the camera in the space and they are thus called *intrinsic camera parameters* [87]. Once $\mathbf{C}$ is known, the camera is termed *calibrated*. A calibrated camera acts as a direction sensor which ables to measure the direction of rays.

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha_u & s_\theta & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ f \end{bmatrix} = \mathbf{C} \begin{bmatrix} x_c \\ y_c \\ f \end{bmatrix} \tag{3.13}$$

$s_\theta = tan(\theta) f k_v$ is the *skew factor* by non-orthogonality between the image axis. Since the width and height of the camera sensor are generally different, different scales in the image u and v directions $\alpha_u = f k_u$ and $\alpha_v = -f k_v$ are used[25]. k_u and k_v denote the units of

25) The v-coordinate is opposite to the y-coordinate.

$[pixels/length]$.

$$\begin{aligned} k_u x_c &= u - u_0 \\ k_v y_c &= v_0 - v \end{aligned} \tag{3.14}$$

The principal point (u_0, v_0) is the point where the optic axis intersects the image plane. Using the different scale units, the camera calibration matrix $\mathbf{C}$ is also written as

$$\mathbf{C} = \begin{bmatrix} fk_u & s_\theta & u_0 \\ 0 & -fk_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.15}$$

k_u and k_v are not individually considered and are often combined with f to calculate α_u and α_v. The ratio between α_u and α_v i.e. α_v/α_u designates the *aspect ratio* which is used to represent the skew factor as Figure 3.3 represents.

3.3 Complete 3-D Camera Model

The camera calibration simply assumes that the center of a pin-hole camera is located at the center of the world coordinate system and not far away from the optical axis. However, it is insufficient in a case with more than one camera or a specific world coordinate system since the world coordinate system does not usually coincide with the perspective reference frames. Since a camera can be placed at the center of the world coordinate system by a rigid transformation as Figure 3.4 shows, the 3-D coordinates undergo a Euclidean motion $\mathbf{M} = \mathbf{R}\mathbf{M}_c + \mathbf{t}$ described by the different orientations $\mathbf{R}$ and translation $\mathbf{t}$ of cameras. The parameters that define the position and orientation of the camera reference frame with respect to known world reference frame are called *extrinsic camera parameters* [87]. The homogeneous notation aims to express the Euclidean motion in linear form.

$$\tilde{\mathbf{M}} = \mathbf{T}\tilde{\mathbf{M}}_c \tag{3.16}$$

where $\tilde{\mathbf{M}} = [\ \mathbf{M}^T \quad 1]^T$, $\tilde{\mathbf{M}}_c = [\ \mathbf{M}_c^T \quad 1]^T$. $\mathbf{T}$ is 4×4 matrix referred to as *homogeneous transformation matrix.*

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0_3^T & 1 \end{bmatrix} \tag{3.17}$$

This equation can be simply expressed in the camera frame with the same rotation and translation as an inverted rotation and translation applied to the cameras due to the *Gauge*

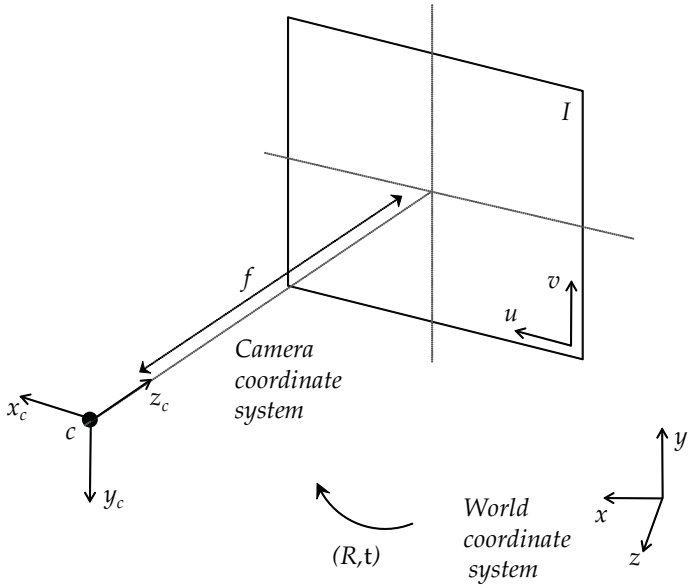


Figure 3.4. Transformation between the camera and world coordinates

freedom $\mathbf{m} \sim \mathbf{M}_c\mathbf{T}^{-1}\mathbf{T}\mathbf{M}$ for any projectivity $\mathbf{T}$ (i.e. the structure of the 3-D points $\mathbf{M}$ is jointly projective for the 3-D cameras $\mathbf{M}_c$).

$$\tilde{\mathbf{M}}_c = \begin{bmatrix} \mathbf{R}^T & -\mathbf{R}^T\mathbf{t} \\ 0_3^T & 1 \end{bmatrix} \tilde{\mathbf{M}} \tag{3.18}$$

The center of projection is given by the null-space of $\mathbf{M}$ and conveniently defines the camera center at the cameras' position. A combination form of Equation 3.12, 3.13 and 3.18 makes the complete expression of perspective projection of a pinhole camera which has calibration, position and orientation.

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha_u & s_\theta & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{R}^T & -\mathbf{R}^T\mathbf{t} \\ 0_3^T & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{3.19}$$

The coordinates of a 3-D point in a world coordinate system $\mathbf{M}$ is projected to the retinal image coordinates $\mathbf{m}$ are inherently nonlinear. Converting $\mathbf{m}$ and $\mathbf{M}$ to homogeneous coordinates $\tilde{\mathbf{m}}$ and $\tilde{\mathbf{M}}$ makes them linear. Equation 3.19 can be compactly expressed as

$$\tilde{\mathbf{m}} \simeq \mathbf{C}[\mathbf{R}^T| - \mathbf{R}^T\mathbf{t}]\tilde{\mathbf{M}} \tag{3.20}$$

or

$$\tilde{\mathbf{m}} \simeq \mathbf{P}\tilde{\mathbf{M}} \tag{3.21}$$

where $\mathbf{P}$ is a 3×4 matrix, of rank 3, called the *perspective projection matrix* or *camera matrix* which folds the 3×3 camera calibration matrix $\mathbf{C}$ mapping the image coordinates to the retinal image coordinates and the 3-D camera pose $(\mathbf{R}, \mathbf{t})$ i.e. displacement from the world coordinate system to the camera coordinate system.

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & 1 \end{bmatrix} \tag{3.22}$$

This matrix is defined up to a scale factor only, and hence has 11 independent entries which represents the linearity between known 3-D points (x_i, y_i, z_i) and measured feature positions (u_i, v_i).

$$\begin{aligned} u_i &= \frac{p_{11}x_i + p_{12}y_i + p_{13}z_i + p_{14}}{p_{31}x_i + p_{32}y_i + p_{33}z_i + 1} \\ v_i &= \frac{p_{21}x_i + p_{22}y_i + p_{23}z_i + p_{24}}{p_{31}x_i + p_{32}y_i + p_{33}z_i + 1} \end{aligned} \tag{3.23}$$

Homography : For two cameras with camera calibration parameters $\mathbf{C}$ and $\mathbf{C}'$ looking at points $\tilde{\mathbf{M}}_i$ on a plane, we choose $\mathbf{R} = \mathbf{I}$ and $\mathbf{t} = 0$, i.e. see Equation 3.12, for the first camera where $\mathbf{I}$ is 3×3 *identity matrix*. The projections of $\tilde{\mathbf{m}}_i$ in the first camera to a point $\tilde{\mathbf{m}}'_i$ in the second camera are written as

$$\tilde{\mathbf{m}}_\mathbf{i} = \mathbf{C}\mathbf{H}\mathbf{C}'^{-\mathbf{1}}\tilde{\mathbf{m}}' \tag{3.24}$$

where $\mathbf{H} = \mathbf{R} - \mathbf{t}\tilde{\mathbf{n}}^\mathbf{T}/d$ is called 3×3 *homography matrix*, and $\mathbf{R}$ and $\mathbf{t}$ is the rotation matrix and the translation vectors from the first camera to the second camera. $\tilde{\mathbf{n}}$ and d are the normal vectors of the plane and the distance to the plane respectively. With at least 4 coplanar features, a homography can be defined in 2-D space as a mapping between a point on a plane, as seen from the first camera, to the same point on the plane as seen from the second camera. This has many practical applications for compositing 2-D or 3-D objects into images or a video seamlessly by fitting the virtual camera to the real camera.

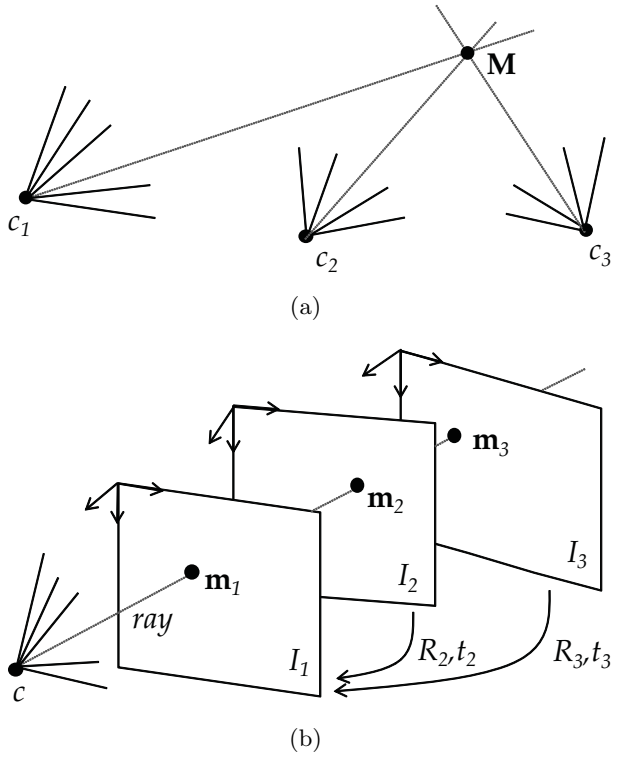


Figure 3.5. Plenoptic function - (a) plenoptic function, (b) the optical-ray in a collineation for arbitrary camera poses

3.4 Ray-space Definition using Optical Flow

The *ray-space* is the space of all light rays in 3-D which is one of the intrinsic parts of all of image-based rendering methods. The light field theories have been made towards the complete mathematical model called *plenoptic function*[26] [79, 80] using a ray projection. The projection can be geometrically represented by a ray through the optical center c_i and the point in space that is being projected onto the image plane I_i as Figure 3.5(a) shows. The optical-ray of an image point $\mathbf{m_i}$ is the locus of points in space that projects onto a camera c shown in Figure 3.5(b). The optical-ray can be described as a parametric line passing through the camera projection center c and a point at infinity that projects onto $\mathbf{m}$:

$$\tilde{\mathbf{M}} = \begin{bmatrix} -\mathbf{P}_{3\times3}^{-1}\mathbf{P}_4 \\ 1 \end{bmatrix} + \alpha \begin{bmatrix} -\mathbf{P}_{3\times3}^{-1}\tilde{\mathbf{m}} \\ 0 \end{bmatrix} \tag{3.25}$$

where $\mathbf{P}$ is normalized and represented by the block form $\mathbf{P} = [\mathbf{P_{3\times3}}\,|\mathbf{P_4}]$ with the first three rows and the first three columns. $\tilde{c} = \mathbf{P}_{3\times3}^{-1}\mathbf{P}_4$ is a three-vector containing the affine, i.e.

26) The plenoptic function is the 5-D function representing the intensity or chromaticity of the light observed from every position and direction in 3-D space.

(a)

(b)

Figure 3.6. Space-time ray-space - (a) illustration of space-time ray-space, (b) ray-space in scanlines

non-homogeneous coordinates of the optical center c, and the fourth column of $\mathbf{P}$ is $\mathbf{P_4}$. The scale parameter α in the equation of the optical-ray correspond to the depth of the point $\tilde{\mathbf{M}}$.

However, the ray-space did not consider a geometric localization problem with structure-from-motion, because the theory is based on the nature of the global approximation of the projection. Here, we define new terms of the ray-space which defines a geometric localization using a fluid in space-time. When an image sequence is taken by a hand-held camera, the space-time ray-space can be constructed from captured by scan-lines as Figure 3.6(a) shows. Each horizontal line in Figure 3.6(b) corresponds to a scan-line of a frame of the image sequences, and its vertical coordinate indicates the corresponding camera position. Since the camera motion is continuous, the ray-space forms a space-time flow.

A fluid in space-time $\mathbb{R}^4$ is a bundle $p : \mathbb{R}^4 \rightarrow \mathbb{R}^3$ (i.e. a projection mapping) and a 3-form $\tilde{\alpha}$ on $\mathcal{R}^3$ with the ray particles of the fluid, with density $\tilde{\alpha}$. Due to the well-known conservation law for fluids, the flow is described by means of 3-form in $\mathbb{R}^4$ or 2-form in $\mathcal{R}^3$. The pull-back

p^* of the density 3-form to $\mathbb{R}^4$ is defined as $\alpha = p^*(\tilde{\alpha})$,

$$d\alpha = dp^*(\tilde{\alpha}) = p^*(d\tilde{\alpha}) = 0 \tag{3.26}$$

where α has a closed form $d\alpha = 0$ since each object can be separated in the space.
Now, we consider light field created by a 3-D Lambertian surface $\mathcal{S}$ which reflects light equally in all directions. Each point on the surface defines a pencil of light rays with equal brightness. The surface can be parameterized by

$$(u_1, u_2, f(u_1, u_2)) \in \mathbb{R}^3 \tag{3.27}$$

where $f(u_1, u_2)$ is a function.
Since intensity is energy density on the 2-D surface, the brightness is 2-D representation with a 2-form $\tilde{L}$ of the surface $\mathcal{S}$ which is decomposable. The condition for decomposability is $L \wedge L = 0$, and L can be integrated to get the total energy $\int_{\mathcal{S}} \tilde{L}$ radiated from the surface $\mathcal{S}$. This is the coordinate-free expression for the integral of a 2-form over the manifold $\mathcal{S}$ on which it is defined. f is closed in the surface of an object.

$$dL = dp^*(\tilde{L}) = p^*(d\tilde{L}) = 0 \tag{3.28}$$

The ray-space is equivalent to the tangent bundle to the sphere $\mathcal{S}^2$ where can be expressed by $\mathcal{G}_2^4$ using the Grassmann manifold[27]. $\mathcal{G}_2^4$ can be represented by the space of skew symmetric tensors $\mathbf{T}$ belonging to the hyperbolic quadric which satisfies the Plücker condition[28] for the decomposability $\mathbf{T} \wedge \mathbf{T}$. $p : \mathcal{G}_2^4 \rightarrow \mathcal{S}$ is a bundle and any section on it is a camera.
If $e_{\mu\nu} = e_\mu \wedge e_\nu$ are basis vectors in $\mathbb{R}^4$, a skew symmetric tensor is

$$\mathbf{T} = a_1 e_{01} + a_2 e_{02} + a_3 e_{03} + b_1 e_{23} + b_2 e_{31} + b_3 e_{12} = (\vec{a}, \vec{b}) \tag{3.29}$$

where $\vec{a}$ is the direction and $\vec{b}$ is is the moment of the line which satisfies the Plücker condition $\vec{a} \cdot \vec{b} = 0$ and $\vec{b} = \vec{m} \times \vec{a}$ for any point $\vec{m}$ on the line. Equation 3.27 can be rewritten by

$$\vec{a}, \underbrace{(u_1, u_2, f(u_1, u_2))}_{=\vec{m}} \times \vec{a} = \vec{a}, \vec{b}(u_1, u_2) \tag{3.30}$$

By the field-equation theory, the boundary condition problem with $dL = 0$ derive the *Laplace equation* [81]. A ray-space well corresponds to the *optical flow* with brightness conservation.

27) A Grassmann manifold is a certain collection of vector subspaces of a vector space. Grassmann coordinates describe k-D linear subspaces, or flats, in a k-D Euclidean space. In particular, $\mathcal{G}_k^n$ is the Grassmann manifold of k-D subspaces of the vector space $\mathbb{R}^n$.

28) Plücker coordinates satisfy a quadratic constraint which establish a one-to-one correspondence between the 4-D space of lines in P^3 and points on a quadric in P^5, i.e. projective 5-space.

The boundary condition in a single image is closely related to the gradient flow and the divergence of the gradient field[29]. The details of a dense 3-D optical flow considering the boundary condition will be described in Chapter 6.

3.5 Experimental Results

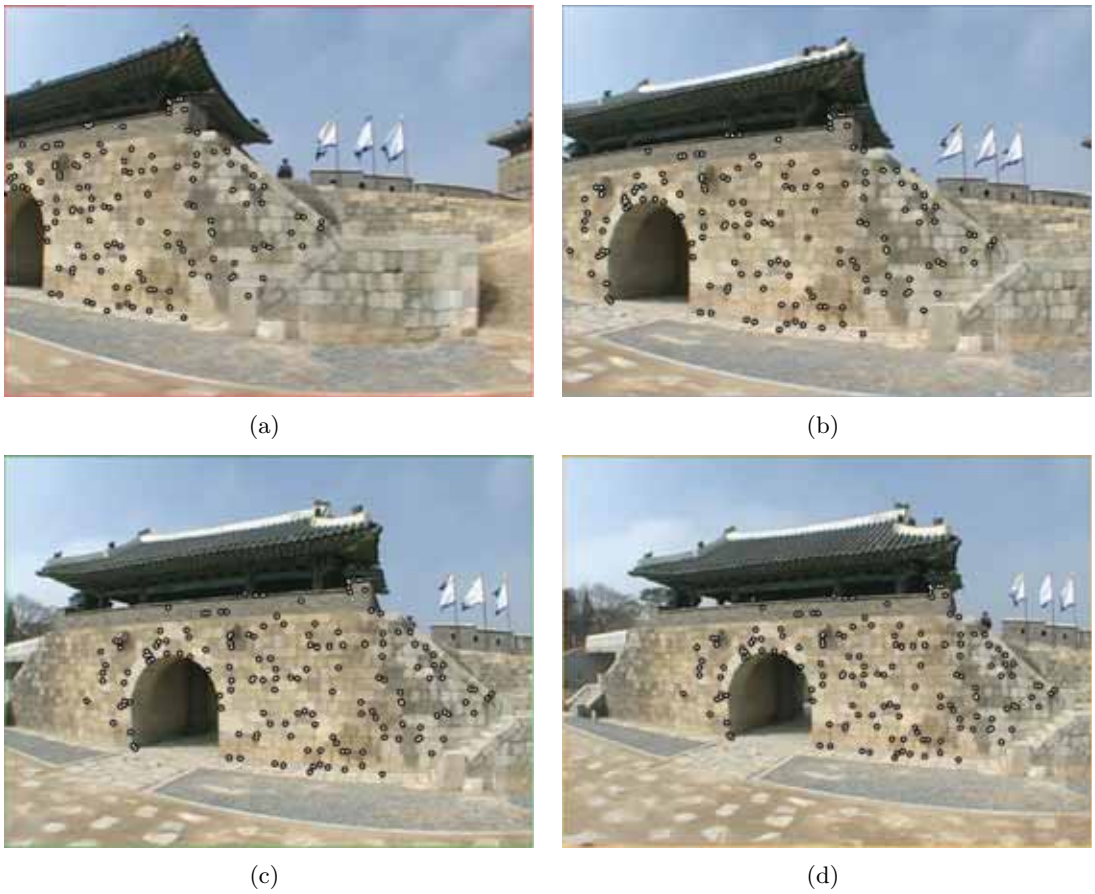
Camera pose estimation with planar region detection : In this chapter, the camera pose estimation from a natural image sequence is simulated by a scheme that relies on multiple images of a planar region in different poses. The planar calibration method is already developed well in photogrammetric calibration studies[30] and provides an estimate of the camera pose with respect to each calibration object pose. Although at least 4 points are needed to estimate the relative camera pose, it is difficult to check the reliability of the 4 points. We estimate the camera pose from coplanar features in natural video sequences.

Starting from some seed points, regions of interest are detected in the image sequence by an iterative region-growing scheme. The low-pass filtered image, i.e. scale-space, is used to avoid local minima problems for the region growing. Features are automatically detected in the growing regions and tracked in two consecutive image frames by the KLT tracker [85, 86]. The tracker minimizes the error between local changes of the image and the original image.

$$E = \int\int_W [I(\mathbf{A}\mathbf{x} + \mathbf{t}) - I(\mathbf{x})]w(\mathbf{x})d\mathbf{x} \tag{3.31}$$

$\mathbf{x} = (u, v)$ is intensity of pixel in image I, and $W \times W$ is the window size of correlation. Affine transform matrix $\mathbf{A}$ and translation $\mathbf{t}$ are linearly approximate to the changes of images.
The plane-induced collineation can be estimated by the homography from the 4 detected features in the regions. We compute the homography between the features and the calibrated points on the planar object. The planarity constraint is enforced to assure that only features of planar region are selected as seeds in the expanded region. A perspective transformation can be approximated with affine transformations that is only valid locally for the planar regions. Features in non-planar regions are eliminated by the distortion of point patterns of non-planar regions that cannot be approximated to an affine transformation. If features are far from the initial seeds, features are checked to update the homography. When the correlation value becomes large and the features are not selected any more, the iteration is stopped.
A test image sequence is captured by a hand-held camera in the outdoor environment and consists of 90 frames with the image resolution of 800×592 pixels. The camera translates fast and rotates smoothly. The detected and tracked features on a planar object for 4 frames are shown in Figure 3.7. The feature points can be partially occluded. The features' positions

29) A gradient field is a conservative field.
30) see Chapter 2.4



(a) (b)

(c) (d)

Figure 3.7. Feature tracking in a planar surface - (a) Camera 1, (b) camera 2, (c) camera 3, (d) camera 4

corresponding to the same face are coplanar. Since the camera parameters are calculated from the feature correspondences, the accuracy of the feature points is very important and can be visually evaluated in Figure 3.7. Figure 3.8 represents a long trajectory of feature tracks through the whole sequence, and this proves the endurance of the matching method. Figure 3.9 shows the relative poses of the 4 cameras when the planar features are located at the centralized coordinate. The border lines of the 4 frame images in Figure 3.7 are marked by different colors for visualization purpose, and the same colors are applied to the 4 cameras shown in Figure 3.8. Figure 3.10 represents the camera calibration result for all frames. The relative Euclidean distances between the camera path and 3-D points are shown by using the 3-D plots with X-Z axis and X-Y axis. The X-Z and X-Y plots respectively visualize the relative depth between cameras and 3-D points and the vertical camera movements. The results represent that the estimated 2-D motion parameters using feature points can be closely related by a collinearity of the feature locations.

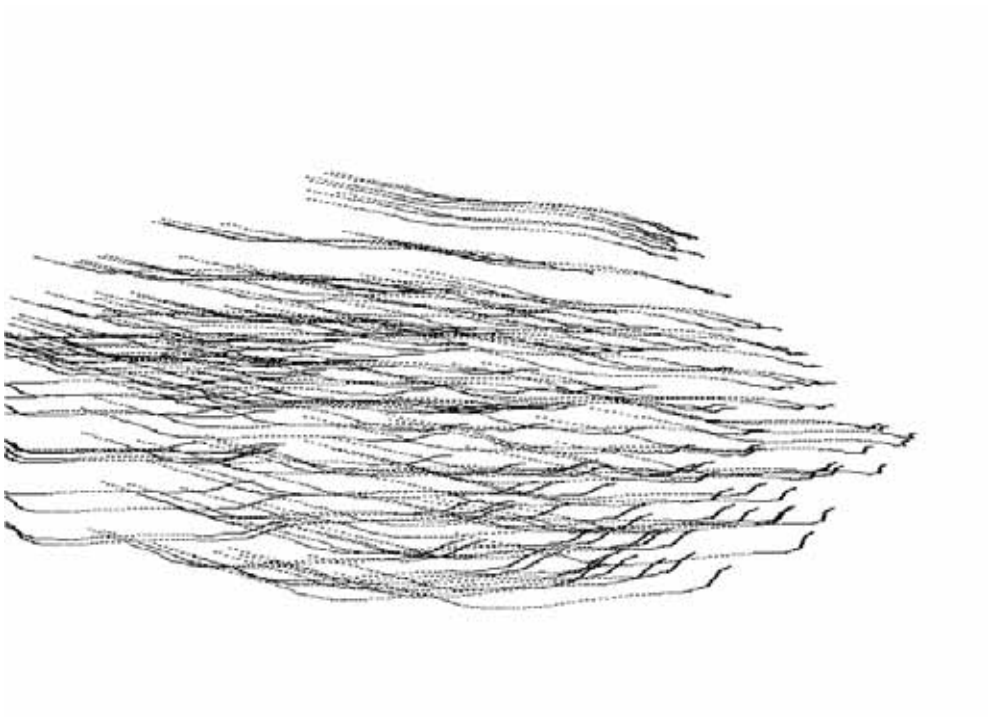


Figure 3.8. Trajectory of moving features by camera motion

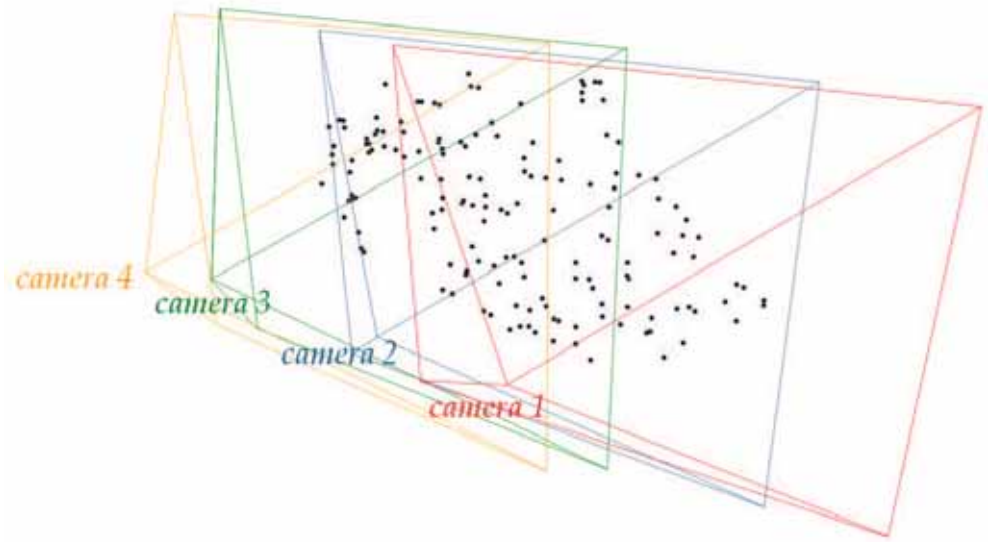
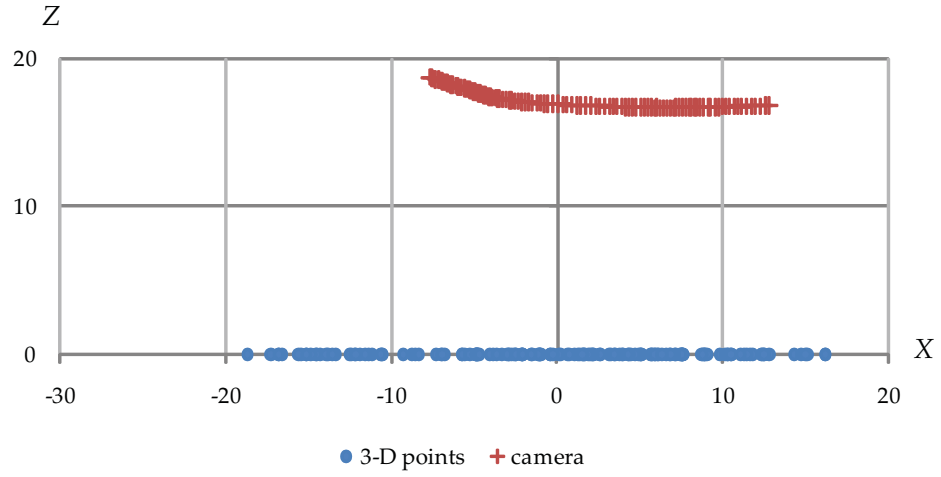


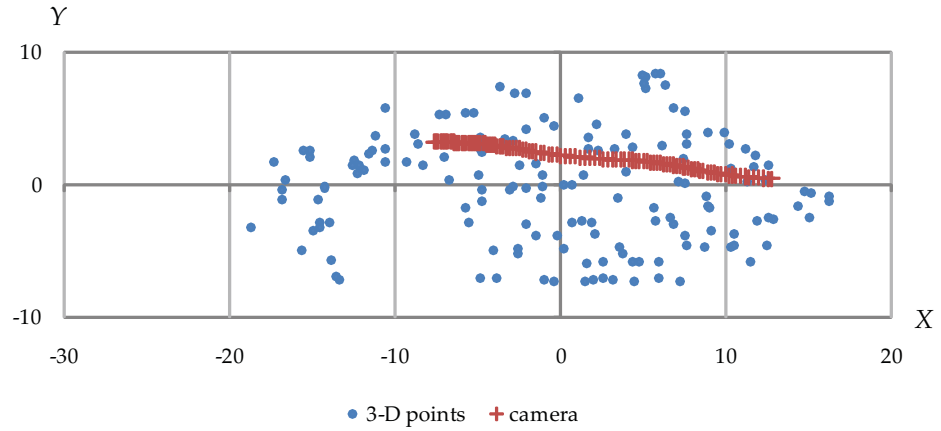
Figure 3.9. Camera calibration results for 4 frames

Optical-ray representation : For the evaluation of the estimated parameters, the optical-ray[31] for an arbitrary camera center is applied. Figure 3.11(a) represents a simple space-time

31) The details of the optical-ray method was described in described in the chapter 3.4.



(a)



(b)

Figure 3.10. Relative 3-D position between cameras and tracked points - (a) plot with X-Z axis, (b) plot with X-Y axis

ray-space without a calibration-based warping. The space-time ray-space is u-t or u-z sampling of an video sequence $(I(u,v),t)$ or layered depth image $(I(u,v),Z)$ where $I(u,v)$ is an image frame, and t and Z respectively denote time and depth coordinate. The position u of a scan-line is shown in Figure 3.6(a). Since the camera is moving, this causes motion blurring artifacts, and the distortions in the t-axis, in the ray-space show the translational and rotational camera movement. An interpolation of a video sequence captured by a moving camera causes a serious motion blurring artifacts as Figure 3.12(a) represents. If we know the accurate camera parameters, the optical-ray can be warped and rearranged into the projection ray of a virtual camera, i.e. arbitrarily or user-defined camera, as Figure 3.12(b) exemplifies, and an interpolation of the rearranged ray does not make motion blurring artifacts. Fig-

(a)

(b)

Figure 3.11. Optical-ray - (a) scanline optical-ray (b) rearranged, i.e. warped, optical-ray using camera calibration parameters

ure 3.12(b) is the simulation results of using the rearranged optical-ray method.

Fourier spectral analysis of ray-space : The plenoptic sampling method [82–84] introduced a Fourier spectral analysis of light field signals to determine the optimal sampling. We use the spectral analysis to compare the spatial variations in Figures 3.12(a) and 3.12(b). We transform the u-t domain corresponding to images to the Fourier spectral domain $\mathcal{F}_u$-$\mathcal{F}_t$. We remark that it is not a transform of the image $I(u, v)$. Figures 3.12(c) and 3.12(d) show the Fourier spectral domain images respectively corresponding to Figure 3.11(a) and 3.11(b)İn the Fourier spectral domain images, each point represents a particular frequency in the spatial time domain image, and thus the tilted lines in the spectrum visualize the spatial variation in the ray-space as Figure 3.13 illustrates. u and $u + \Delta u$ are the light samples on the horizontal scan-line of image and t is the time. Since we use a pin-hole camera model that defines a

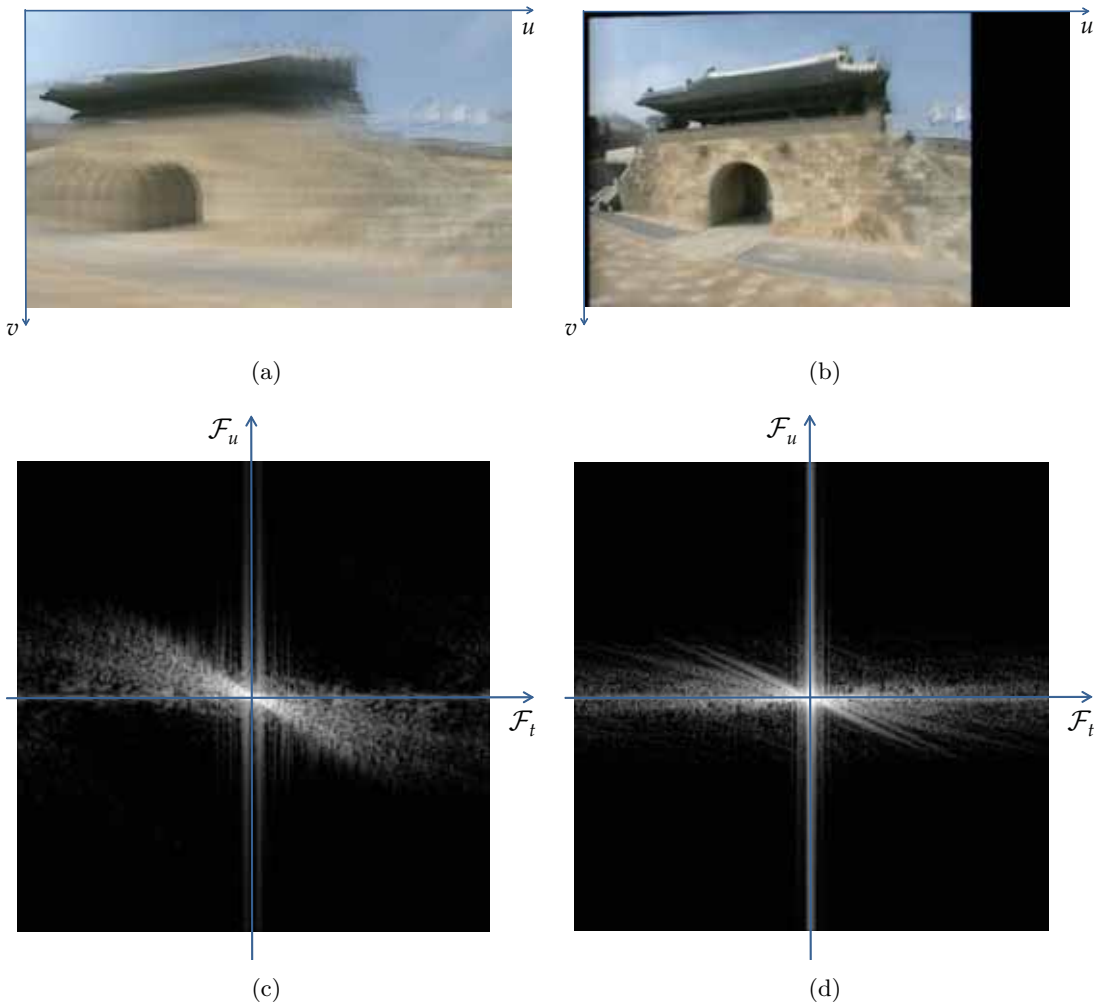


Figure 3.12. Virtual viewpoint in ray-space - (a) simple temporal interpolation (b) temporal interpolation with rearranged optical-ray, (c) Fourier spectral analysis of the ray-space for Figure 3.11a, (d) Fourier spectral analysis of the ray-space for Figure 3.11b

global projection ray, all light samples are assumed to be at the same depth as a point on the focal plane. However, the real scene consists of 3-D points with different depths Z, and some rays diverge from the global ray projection and cause blurring artifacts. The spatial variation of the rays is laid between the minimum depth Z_{min} and the maximum depth Z_{max}, and the spectral representation is bound to the range between maximum and minimum tilts. Although a global image warping with estimated camera parameters can overcome some distortions in Figure 3.12(b), the sampling positions in depth should be approximated to compensate the distortions caused by the different depths. The sampling position can be estimated as a disparity $u - (u + \Delta u) = f\mathbf{t}/z$ between the two image coordinates u and $u + \Delta u$. Additionally, an additional nonlinear parameter optimization such as the Marquardt-Levenberg method

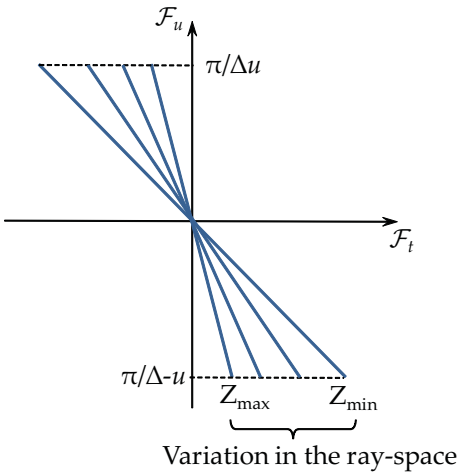


Figure 3.13. Fourier spectral analysis of ray-space

or bundle adjustment method may increase the accuracy since they can fit the projection parameters to the more reliable sampling positions.

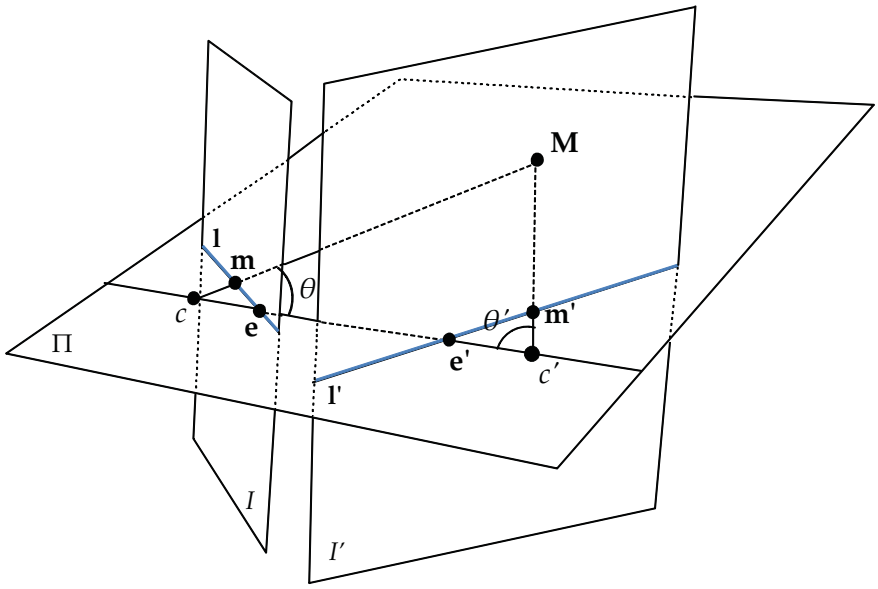
Chapter 4

Spatio-temporal Collineation of View Geometry

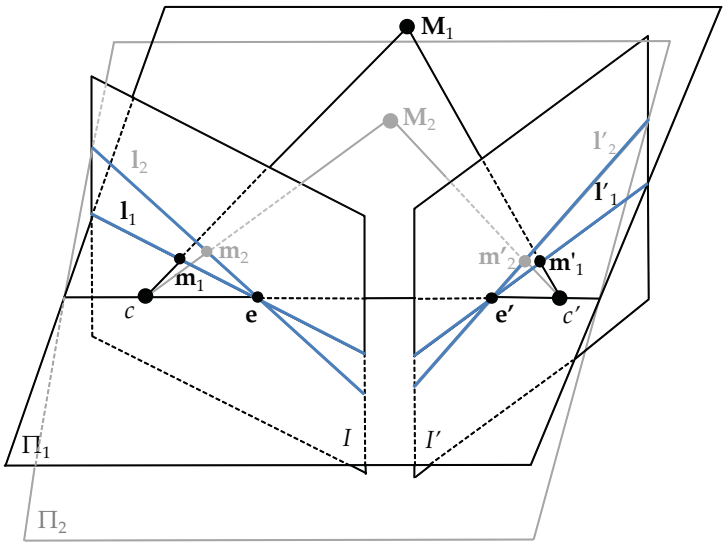
The camera geometry is a model of the direction with parallel rays between 3-D points and the pin-hole camera. The set of all projections along each line parallel to a given direction denotes one projection of the object. However, it is not accurate in most cases since the viewing geometry includes several imaging errors with geometric distortion. Epipolar geometry defining different camera geometries and relative orientation of image bundles should be studied for the geometric corrections of viewing geometry with divergent rays. 2-D projective pencils of epipolar lines derived from image point homologies are two formed bundles of rays in infinite variations of perspective positions in space. When the scene is rigid, the linearized motion in a video sequence defines temporal epipolar geometry. For a stereo video, the spatio-temporal view geometry has intersections of spatial and temporal epipolar lines at reliable corresponding points.

4.1 Epipolar Geometry

The *epipolar geometry* is the intrinsic projective geometry which exists between any two perspective images of a single rigid object/scene. When two cameras look at a 3-D scene, there is a geometric relation between the 3-D points and their projections onto the 2-D images that lead to constraints between the image points. These relations are derived based on the assumption that the cameras can be sufficiently well approximated by the pinhole camera model [89]. Figure 4.1(a) represents two cameras with the optical centers c and c$'$ of the first and second cameras, respectively. Given a physical point $\mathbf{M}$, $\mathbf{m}$ and $\mathbf{m}'$ are its image points on the first and second image planes I and I', respectively. An image point $\mathbf{m}$ in the first image plane I and its corresponding point $\mathbf{m}'$ in the second image plane I' are constrained to lie on a line $\mathbf{l}'$ called the *epipolar line*. If $\mathbf{m}$ and $\mathbf{m}'$ correspond to a 3-D point $\mathbf{M}$ in space,



(a)



(b)

Figure 4.1. Epipolar geometry - (a) relationship among a world point, epipoles and epipolar lines, (b) a pencil of epipolar lines in each image centered on the epipole.

the line $\mathbf{l}'$ is formed by the intersection of the epipolar plane Π which are defined by three 3-D points $\mathbf{M}$, c and c′ and the camera rotation angles θ and θ'. An image point $\mathbf{m}$ may correspond to an arbitrary point on the semi-line c$\mathbf{M}$, and the projection of c$\mathbf{M}$ to I' is the

line $\mathbf{l}'$. Furthermore, the intersection of the line cc$'$ with the second image plane I' is called the *epipole* of image plane I', denoted by $\mathbf{e}'$. The epipole $\mathbf{e}'$ is also the common point passed through by all epipolar lines of the points in the first image plane I. Both epipoles $\mathbf{e}$ and $\mathbf{e}'$ and both optical centers c and c$'$ lie on a single line since each focal point (i.e. optical centers c and c$'$) is projected onto a distinct point into the other camera's image plane.
Figure 4.1(b) shows epipolar planes which intersect each image plane where it forms the epipolar lines. All epipolar lines form a pencil containing the epipoles $\mathbf{e}$ and $\mathbf{e}'$ which is the intersection of the line cc$'$ with the image planes I and I'. Due to the symmetry of the epipolar geometry, all points lying on $\mathbf{l'_1}$ (and $\mathbf{l'_2}$) must lie on the epipolar line $\mathbf{l_1}$ (and $\mathbf{l_2}$), which is the intersection of the plane Π_1 (and Π_2) with the image plane I. These planes must intersect the second image plane I' at a common point, which is $\mathbf{e}'$ and all epipolar lines e.g. $\mathbf{l'_1}$ and $\mathbf{l'_2}$ in the second image plane I' pass through the epipole $\mathbf{e}$, which is the intersection of the line cc$'$ with the image plane I. This is the co-planarity constraint in solving *structure from motion*(SfM) problem when the intrinsic camera parameters are known.

4.2 Fundamental Matrix

The epipolar geometry can be represented by a 3×3 singular matrix. If intrinsic camera parameters are known and image coordinates have been normalized in advance, this matrix is the essential matrix $\mathbf{E}$. Otherwise, the matrix is the *fundamental matrix* $\mathbf{F}$ which is an algebraic representation of the epipolar geometry for two camera systems. If it is assumed that the world coordinate system coincides with the first camera coordinate system, the relation of two cameras can be defined by the pinhole camera model.

$$\tilde{\mathbf{m}} = \mathbf{C}[\ \mathbf{I}\quad 0\ ]\tilde{\mathbf{M}} \quad \text{and} \quad \tilde{\mathbf{m}}' = \mathbf{C}'[\ \mathbf{R}\quad \mathbf{t}\ ]\tilde{\mathbf{M}} \tag{4.1}$$

Eliminating $\tilde{\mathbf{M}}$ in the above two equations, the following equation is obtained:

$$\tilde{\mathbf{m}}'^{T}\mathbf{C}'^{-T}[\mathbf{t}]_{\times}\mathbf{R}\mathbf{C}^{-1}\tilde{\mathbf{m}} = 0 \tag{4.2}$$

where $(\mathbf{R}, \mathbf{t})$ is the rotation and translation which brings points expressed in the first camera coordinate system to the second one. $[\mathbf{t}]_{\times}$ is the anti-symmetric matrix defined by $\mathbf{t}$ such that $[\mathbf{t}]_{\times}\mathbf{x} = \mathbf{t} \times \mathbf{x}$ for all 3-D vector $\mathbf{x}$. Equation 3.24 can be rewritten by the fundamental matrix $\mathbf{F} = \mathbf{C}'^{-T}[\mathbf{t}]_{\times}\mathbf{R}\mathbf{C}^{-1}$.

$$\tilde{\mathbf{m}}'^{T}\mathbf{F}\tilde{\mathbf{m}} = 0 \tag{4.3}$$

If the second camera coordinate system is chosen as the world one, Equation 4.3 becomes $\tilde{\mathbf{m}}^T\mathbf{F}'\tilde{\mathbf{m}}' = 0$ with $\mathbf{F}' = \mathbf{C}^{-T}[\mathbf{t}']_\times\mathbf{R}'\mathbf{C}'^{-1}$, where $(\mathbf{R}', \mathbf{t}')$ transforms points from the second camera coordinate system to the first and the relation between $(\mathbf{R}, \mathbf{t})$ and $(\mathbf{R}', \mathbf{t}')$ is denoted by $\mathbf{R}' = \mathbf{R}^\mathbf{T}$, and $\mathbf{t}' = -\mathbf{R}^T\mathbf{t}$.
Some important properties of the fundamental matrix [89] are

(i) Transpose: If $\mathbf{F}$ is the fundamental matrix of the pair of cameras $(\mathrm{c}, \mathrm{c}')$, $\mathbf{F}^T$ is also the fundamental matrix of the pair in the opposite order: $(\mathrm{c}, \mathrm{c}')$.

(ii) Epipolar lines: For any point $\mathbf{m}$ in the first image plane, the corresponding epipolar line is $\mathbf{l}' = \mathbf{F}\mathbf{m}$. Similarly, $\mathbf{l} = \mathbf{F}\mathbf{m}'$ represents the epipolar line corresponding to $\mathbf{m}'$ in the second image plane.

(iii) Epipoles: For any point $\mathbf{m}$, the epipolar line $\mathbf{l}' = \mathbf{F}\mathbf{m}$ contains the epipole $\mathbf{e}'$. Thus $\mathbf{e}'$ satisfies $\mathbf{e}'^T(\mathbf{F}\mathbf{m}) = (\mathbf{e}'^T\mathbf{F})\mathbf{m} = 0$ for all $\mathbf{m}$. It follows that $\mathbf{e}'^T\mathbf{F} = 0$, i.e., $\mathbf{e}'$ is the left null-vector of $\mathbf{F}$.

(iv) Similarly, $\mathbf{F}\mathbf{e} = 0$, which means that $\mathbf{e}$ is the right null-vector of $\mathbf{F}$.

(v) 7 degrees of freedom: A 3×3 homogeneous matrix has eight independent ratios. However, $\mathbf{F}$ satisfies the constraint $\det(\mathbf{F}) = 0$, which removes one degree of freedom. $\mathbf{F}$ is of rank 2 and the common scaling is not significant. There are only 7 independent parameters among the 9 elements of the fundamental matrix.

(vi) Correlation: Assume a point $\mathbf{m}$ in the first image plane defines a line $\mathbf{l}' = \mathbf{F}\mathbf{m}$ in the second image plane, the epipolar line of $\mathbf{m}$. If $\mathbf{l}$ and $\mathbf{l}'$ are corresponding epipolar lines, any point $\mathbf{m}$ on $\mathbf{l}$ is mapped to the same line $\mathbf{l}'$.

(vii) Search space reduction: For each point in one image, the corresponding point in the other image must lie on a known epipolar line and the search space for a correspondence is reduced from 2-D image to 1-D epipolar line. This is called the *epipolar constraint*.

4.3 Fundamental Matrix and Camera Projections

The fundamental matrix may be algebraically derived in terms of two camera projection matrix, $\mathbf{P}$,$\mathbf{P}'$. The following formulation is due to Xu and Zhang [106]. A point $\tilde{\mathbf{m}}$ in the first image is matched to a point $\tilde{\mathbf{m}}'$ in the second image, and the optical-ray back-projected from $\mathbf{m}$ by $\mathbf{P}$ and $\tilde{\mathbf{m}}'$ by $\mathbf{P}'$ is obtained by solving $\alpha\tilde{\mathbf{m}} = \mathbf{P}\tilde{\mathbf{M}}$ and $\alpha'\tilde{\mathbf{m}}' = \mathbf{P}'\tilde{\mathbf{M}}$ where α is the scale factor. For the first camera, the optical-ray is parameterized by the scalar λ as

$$\tilde{\mathbf{M}}(\lambda) = \alpha\mathbf{P}^+\tilde{\mathbf{m}} + \lambda\mathbf{c} \tag{4.4}$$

where $\mathbf{P}^+$ is the pseudo-inverse of matrix $\mathbf{P}$, i.e. $\mathbf{P}\mathbf{P}^+ = \mathbf{I}$, and c is the camera center which all optical rays converge, defined by $\mathbf{P}\mathrm{c} = 0$, i.e. null-vector. $\lambda = 0$ and $\lambda = \infty$ respectively denotes points on the ray and the first camera center c. Similarly, an optical-ray is easily

defined by projecting these points onto the second camera as $\mathbf{P}'\mathbf{P}^{+}\tilde{\mathbf{m}}$ and $\mathbf{P}'\mathrm{c}$. The epipolar line $\mathrm{l}' = (\mathbf{P}'\mathrm{c}) \times (\mathbf{P}'\mathbf{P}^{+}\tilde{\mathbf{m}})$ is the line joining two projected points. The point $\mathbf{P}'\mathrm{c}$ is the epipole $\mathbf{e}'$ in the second image which is the projection of the first camera center. the epipolar line can be defined by

$$\mathrm{l}' = \left[\mathbf{e}'\right]_{\times} (\mathbf{P}'\mathbf{P}^{+})\tilde{\mathbf{m}} = \mathbf{F}\tilde{\mathbf{m}} \tag{4.5}$$

where the fundamental matrix is

$$\mathbf{F} = \left[\mathbf{e}'\right]_{\times} \mathbf{P}'\mathbf{P}^{+} \tag{4.6}$$

where $[.]_{\times}$ denotes 3×3 skew symmetric matrix representing the vector cross product[32]. Eliminating s and s' by multiplying $\tilde{\mathbf{m}}'^{T}$ from the left, i.e. equivalent to a dot product, we have

$$\tilde{\mathbf{m}}'^{T}\mathbf{F}\tilde{\mathbf{m}} = 0 \tag{4.7}$$

The use of the pseudo-inverse of the projection matrix is valid for both full perspective projection as well as affine cameras. However, this derivation breaks down in the case where the two cameras has the common camera center of both $\mathbf{P}$ and $\mathbf{P}'$, and thereby $\mathbf{P}'\mathbf{c} = \mathbf{0}$.

4.4 Fundamental Matrix Estimation with Parallax

If two sets of image points are the projections of a plane in space, they are related by a homography $\mathbf{H}$. Shashua and Avidan [97] showed that the fundamental matrix and the homography is related by $\mathbf{F} = [\tilde{\mathbf{e}}']_{\times}\mathbf{H}$. For a point which does not belong to the plane, an epipolar line is defined by

$$\mathbf{l}' = \tilde{\mathbf{m}}' \times \mathbf{H}\tilde{\mathbf{m}} \tag{4.8}$$

This is essentially the same formula to Equation 4.5, because the homography $\mathbf{H}$ having the explicit form $\mathbf{H} = \mathbf{P}'\mathbf{P}^{+}$ in terms of the two camera matrix. Since a constraint on the epipole is given as $\tilde{\mathbf{e}}'^{\mathbf{T}}\mathbf{l}' = \mathbf{0}$, two such points are sufficient to estimate the epipole $\mathbf{e}'$.
Given the fundamental matrix $\mathbf{F}$ and the epipoles $\mathbf{e}$ and $\mathbf{e}'$ in an image pair, the entire group

32) For a vector $\mathbf{a}$, $[\mathbf{a}]_{\times}$ is such that $[\mathbf{a}]_{\times}\mathbf{b} = \mathbf{a} \times \mathbf{b}, \forall \mathbf{b}$, and $[\mathbf{a}]_{\times}^{T} = -[\mathbf{a}]_{\times}$. The skew symmetric matrix has a form as

$$[\mathbf{a}]_{\times} = \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix}$$

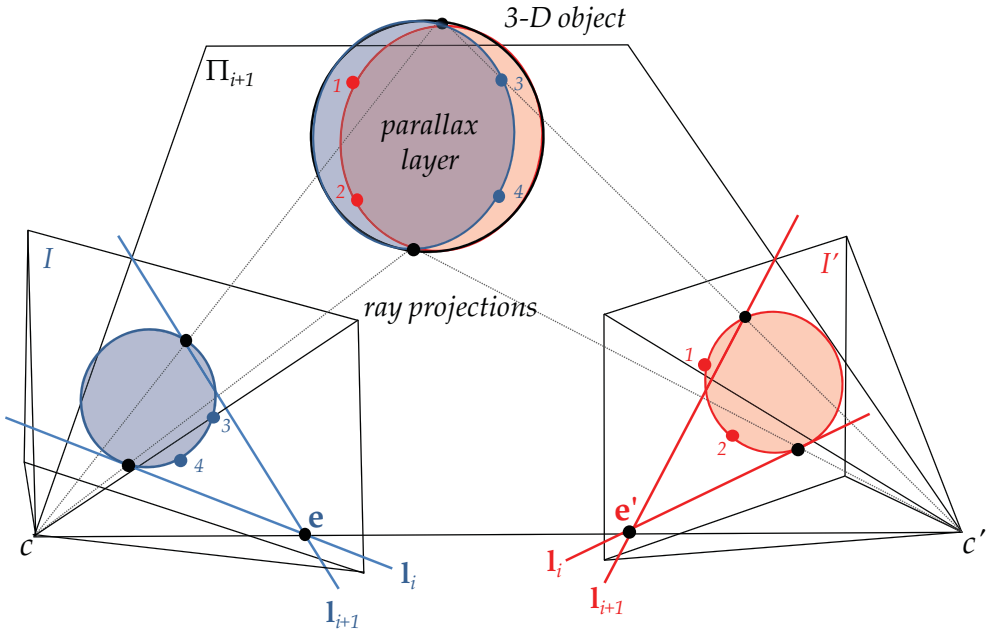


Figure 4.2. Parallax in projection rays

of possible homography matrices between two images lies in a 4-D subspace. The family of homography matrices from the first view to the second view are spanned by 4 homography matrices $\mathbf{H_1}, \cdots, \mathbf{H_4}$ whose planes do not all coincide with a single point.

$$\mathbf{H}_i = [\varepsilon_i]_\times \mathbf{F}, \quad i = 1, 2, 3 \quad \text{and} \quad \mathbf{H}_4 = \mathbf{e}'\delta^T \tag{4.9}$$

ε_i are the identity vector, $\varepsilon_1 = [1, 0, 0]^T$, $\varepsilon_2 = [0, 1, 0]^T$, $\varepsilon_3 = [0, 0, 1]^T$. δ is a vector such that $\delta^T\mathbf{e} \neq \mathbf{0}$. These 4 homography matrices are referred to as *primitive homographies* and allow any other homography $\mathbf{H}$ to be expressed as a linear combination $\mathbf{H} = \sum_{i=1}^{4} \alpha_i \mathbf{H}_i$, $\alpha_i \in \mathcal{R}$. For a 3-D plane Π inducing a homography $\mathbf{H}$ between two images, any 3-D point $\mathbf{M}$, which is not on Pi and projects to image points $\tilde{\mathbf{m}}$ and $\tilde{\mathbf{m}}'$, satisfies the equation $\tilde{\mathbf{m}}' = \mathbf{H}\tilde{\mathbf{m}} + \mathbf{e}'$ with arbitrary scaling, i.e. since $\mathbf{H}$ and $\mathbf{e}'$ are homogeneous entities, defined up to an arbitrary scale factor λ.

$$\tilde{\mathbf{m}}' = \mathbf{H}\tilde{\mathbf{m}} + \lambda\mathbf{e}' \tag{4.10}$$

where the first term is the homography induced by Π, and the second term involves *parallax* due to the deviation of the actual 3-D location from Π. The term λ designates as *relative affine structure*[98] depends on $\tilde{\mathbf{m}}$ but is invariant to the choice of the second image. The point $\tilde{\mathbf{m}}$ is to establish a common relative scale between $\mathbf{H}$ and $\mathbf{e}'$. Given $\tilde{\mathbf{m}}$, $\tilde{\mathbf{m}}'$, $\mathbf{H}$ and $\mathbf{e}'$, the term λ corresponding to $\tilde{\mathbf{M}}$ can be computed by cross-multiplying both sides of Equation 4.10

with $\tilde{\mathbf{m}}'$.

$$\lambda = \frac{\left(\mathbf{H}\tilde{\mathbf{m}} \times \tilde{\mathbf{m}}'\right)^T \left(\tilde{\mathbf{m}}' \times \mathbf{e}'\right)}{\left\|\tilde{\mathbf{m}}' \times \mathbf{e}'\right\|^2} \tag{4.11}$$

The parallax layers is selected to lie in-between 3-D points approximated by a set of matching point pairs $(\tilde{\mathbf{m}}_i, \tilde{\mathbf{m}}'_i)$ $i = 1, \cdots, n$ in two frames. Figure 4.2 illustrates parallax in projection rays. The two boundary lines, i.e. red and blue lines on a 3-D object surface, include points a, b, c and d with the same parallax, and the lines are projected to the silhouettes of the object in the viewpoint images. Any planar homography defined between the image pair I and I' can be expressed as the linear combination of the 4 primitive homographies. We choose primitive homographies with 4 coefficients μ_j minimizing $\tilde{\mathbf{m}}'_i \simeq \sum_{j=1}^{4} (\mu_j \mathbf{H}_j)\tilde{\mathbf{m}}$.

4.5 Essential Matrix Estimation

The *essential matrix* [24] is a specialization of the fundamental matrix for the case of normalized image coordinates. Both the fundamental matrix and essential matrix can be used for establishing constraints between matching image points, but the essential matrix can only be used in relation to calibrated cameras since the inner camera parameters must be known in order to achieve the normalization. The cameras which are calibrated for the essential matrix are useful for determining both the relative position $\mathbf{t}$ and rotation $\mathbf{R}$ between the cameras and the 3-D position of corresponding image points. The essential matrix $\mathbf{E}$ can be defined by the equation

$$\mathbf{E} = \mathbf{R}[\mathbf{t}]_\times \tag{4.12}$$

where $\mathbf{R}$ is the rotation matrix of the camera and $[\mathbf{t}]_\times$ denotes the skew-symmetry matrix of the translation vector $\mathbf{t}$. The skew-symmetric matrix must have two singular values which are equal and one which is zero. The multiplication of the rotation matrix does not change the singular values which means that also the essential matrix has two singular values which are equal and one which is zero. The essential matrix has a relationship with point correspondences, as shown in the following equation:

$$\widehat{\mathbf{m}}'^T \mathbf{E} \widehat{\mathbf{m}} = 0 \tag{4.13}$$

where $\widehat{\mathbf{m}}$ and $\widehat{\mathbf{m}}'$ are respectively representations of the point correspondences $\mathbf{m}$ and $\mathbf{m}'$ using normalized image coordinates. These point correspondences are extracted from a pair of images taken by calibrated cameras. The following equation depicts the relationship between

the essential matrix $\mathbf{E}$ and the fundamental matrix $\mathbf{F}$:

$$\mathbf{E} = \mathbf{C}_2^T \mathbf{F} \mathbf{C}_1 \tag{4.14}$$

where $\mathbf{C}_1$ and $\mathbf{C}_2$ represent the matrices of intrinsic camera parameters. Since the essential matrix is closely related to the fundamental matrix, it contains most of its properties, except that it has only five degrees of freedom. The nearest rank 2 matrix to $\mathbf{E}$ given by $\hat{\mathbf{E}}$ minimizes the Frobenius norm of $\mathbf{E} - \hat{\mathbf{E}}$ subject to the constraint $\det\hat{\mathbf{E}} = 0$. Let

$$\mathbf{E} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{4.15}$$

be the singular value decomposition (SVD) of matrix $\mathbf{E}$, where $\mathbf{D} = \mathrm{diag}(\sqrt{\lambda}, \sqrt{\lambda}, 0)$ is a three diagonal entry which consists of two nonzero identical and one zero values[33], and $\mathbf{U}$ and $\mathbf{V}$ are respectively 3×3 orthogonal and diagonal matrix. The two nonzero singular values, i.e. $\sqrt{\lambda}$, of $\mathbf{E}$ must be equal, known as the *Huang-Faugeras constraint*. The rotation and translation information of camera motion can be retrieved by splitting the essential matrix, i.e. see Appendix A.6.

4.6 Maximum and Minimum Bounds of Epipolar Range

When the calibration is exact, the search for corresponding points are to be restricted to epipolar lines. However, an inexact calibration causes a need for a wide search range of image matching. In this case, the maximum bounds of epipolar geometry may be utilized to determine a specific range of stereo geometries. In each viewpoint image, all possible epipolar geometric ranges can be represented as the region in the other image formed from the union of all associated epipolar lines. The search range for correspondences can be restricted by imposing limits on the maximum bounds. This method is also efficient for a space variant sampling such as light field. For any point $\mathbf{x}$ in the stereo geometry shown in Figure 4.1(a), the regions r and r' of the corresponding epipolar lines across images I and I' are represented as

$$r = r' \frac{f \sin(\theta) + \mathbf{x} \cos(\theta)}{f \sin(\theta') - \mathbf{x}' \cos(\theta')} \tag{4.16}$$

The range of camera rotation angles is approximated as $\theta' \in [\theta_{\min}, \pi - \theta_{\min}]$ where the minimum angle $\theta_{\min}$ is relative to the baseline. The maximum bound of translation is restricted to a maximum horizontal disparity as $|\mathbf{x} - \mathbf{x}'| \leq D$ due to the epipolar constraint. For the known image size, the maximum and minimum bounds $r_{\max}$ and $r_{\min}$ can be approximated

33) This is rank-2 constraint since all the epipolar lines intersecting in a unique epipole.

as

$$r_{\max} \approx r' \frac{\sqrt{f^2 + \mathbf{x'}^2}}{f \sin(\theta_{\min}) - \mathbf{x'} \cos(\theta_{\min})} \tag{4.17}$$

and

$$r_{\max} \approx r' \frac{\sqrt{f^2 + \mathbf{x'}^2}}{f \sin(\theta_{\min}) - \mathbf{x'} \cos(\theta_{\min})} \tag{4.18}$$

4.7 Epipolar Rectification

Rectification is a fascinating process to make the epipolar geometry between two viewpoints coincident and parallel with the scan-lines of the images. It is efficient for dense depth estimation by reducing the search space from a 2-D image to a 1-D epipolar line. Rectification can be classified in two categories: planar or standard rectification [99, 100] and epipolar rectification [101, 102]. The planar or standard rectification that is mostly used for stereo vision systems uses a planar projective mapping which consists of finding parallel planes with the baseline, i.e. the line passing through the camera centers. This is accomplished by applying a homography to each image that maps the epipole to a point at infinity. With this kind of transformation, rectified images have the following properties:

(a) All epipolar lines are parallel to the horizontal coordinate axis,
(b) Corresponding points have the same vertical displacement.

However, this approach would fail if the epipoles are located within the images or would result in very large images when the epipoles are close to the image borders. In single camera setups, forward camera motion must also be considered. Therefore, Pollefeys et al.[101] and Oram [102] introduced an epipolar rectification approach that reparameterizes the images around epipole using polar coordinates. The epipolar geometry between a pair of images must be known beforehand. The first step is to determine the extreme epipolar lines $\mathbf{l_1}$, $\mathbf{l_3}$ and $\mathbf{l_2}$, $\mathbf{l_4}$, i.e. the epipolar lines passing the outer image corners in both images, as Figures 4.3(a) and 4.3(b) show. The epipolar line transfers $\mathbf{l'} \sim \mathbf{H^{-}Tl}$ and $\mathbf{l} \sim \mathbf{H^{T}l'}$ are calculated as

$$\mathbf{H} \approx \left[\mathbf{e'}\right]_{\times} \mathbf{F} - \mathbf{e'a^{T}} \tag{4.19}$$

where $\mathbf{a}$ is an arbitrary 3 vector such that $\det \mathbf{H} \neq 0$. The first term is a transformation taking the epipole to a point at infinity, and the second term represents 1-D projective transformation of the image along all the epipolar lines.[34]. Figure 4.3(c) represents the epipolar line transfer using time t. The common region can be determined by transforming these lines to the other

34) Equation 4.19 has a form of $I + (1,0,0)^T \mathbf{a^T}$ that can be written as an affine transformation

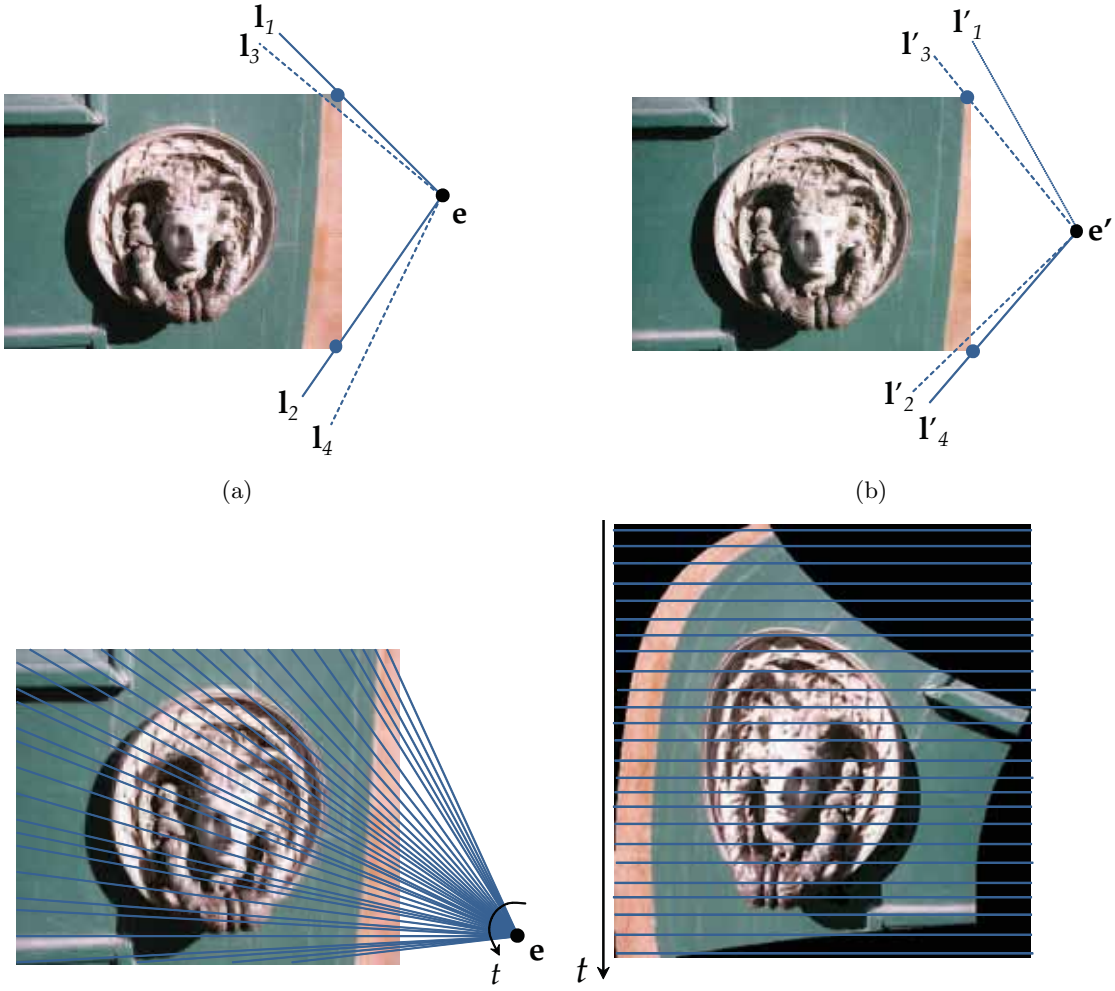


Figure 4.3. Epipolar rectification - (a) the 1st camera image with the extreme epipolar lines, (b) the 2nd camera image with the extreme epipolar lines, (c) the epipolar line transfer, (d) the common region

image as Figure 4.3(d) represents. To avoid pixel loss, each row of the rectified image is obtained by scanning the consecutively orienting epipolar lines $\mathbf{l}_{i-1}$ and $\mathbf{l}_i$ of this region at a minimum angle φ as Figure 4.4 illustrates.

$$\varphi_{i-1} = \arctan\left(\frac{1}{\gamma_{i-1}}\right) \tag{4.20}$$

$$\mathbf{A} = \begin{bmatrix} a & b & c \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

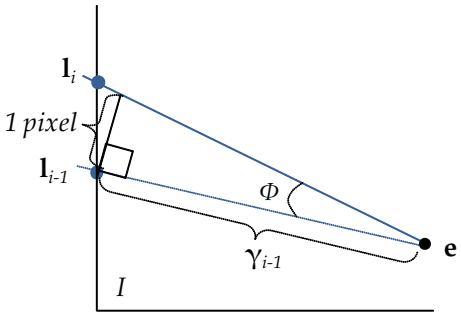


Figure 4.4. Minimum angle to avoid pixel loss in epipolar rectification

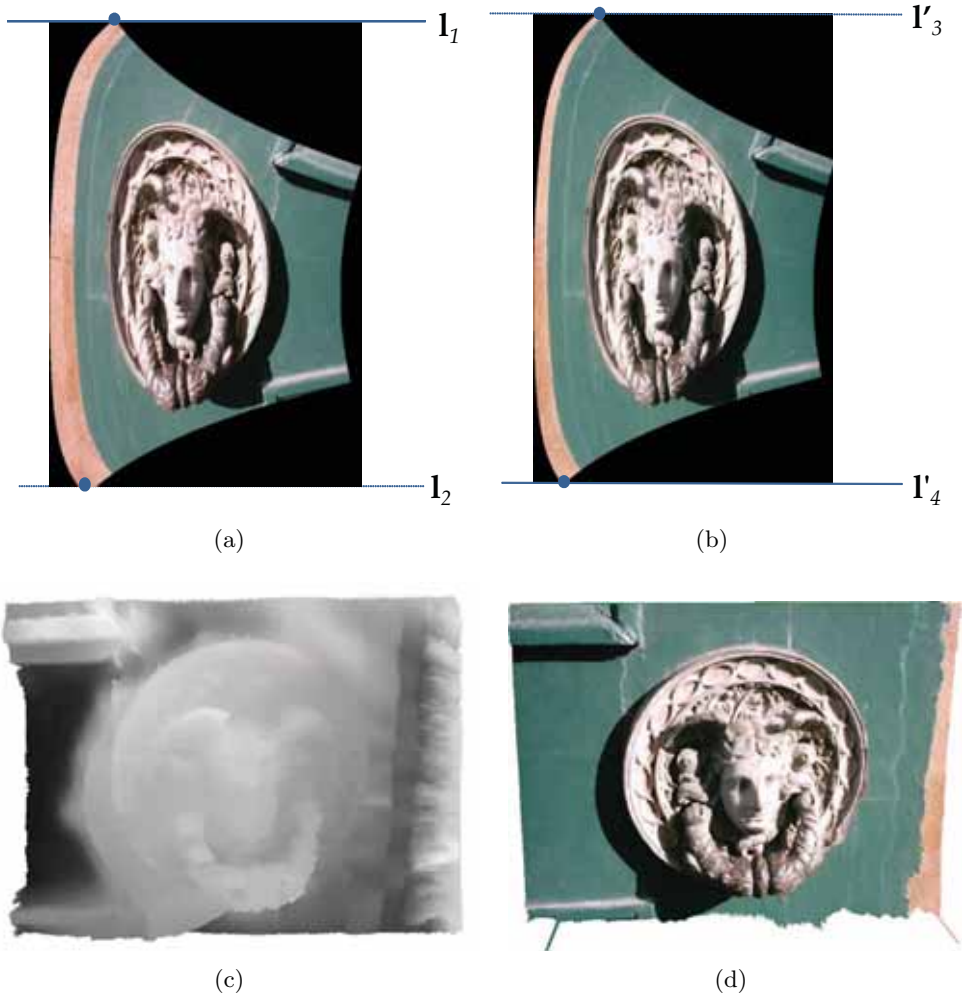


Figure 4.5. Rectified images - (a) rectified first image, (b) rectified second image, (c) dense disparity matching in two rectified images, (d) 3-D view

where φ_{i-1} is the distance between the image border and the epipole of line $\mathbf{l}_{i-1}$. Since a row of the rectified image corresponds to a specific angle, the inverse rectification is possible using look-up tables. Unlike planar rectification, the maximum sizes of the rectified images are fixed and given with

$$\begin{cases} width_{rectifiedimage} = \sqrt{width^2_{originalimage} + height^2_{originalimage}} \\ height_{rectifiedimage} = 2 \cdot (width_{originalimage} + height_{originalimage}) \end{cases} \tag{4.21}$$

Figures 4.5(a) and 4.5(b) represent the rectified stereo images. The size of original stereo images is 589×446, and the size of rectified images is 634×954. A dense stereo matching with consistency checking is applied to the rectified images, and inverse rectification of left-to-right and right-to-left disparity maps recovers dense 3-D depths of the scene as Figure 4.5(c) shows. The details of the dense stereo matching and consistency checking will be described in Chapter 6. Figure 4.5(d) is a 3-D model recovered from the dense 3-D depth.

4.8 Temporal Epipolar Geometry

A temporal linearity can be defined by the geometric relationship of the tracked point between two frames. The spatio-temporal space can be represented by the brightness-invariance concept shown in Equations 1.7, 1.8 and 1.9.

$$I(u, v, t) = I(u + d_u, v + d_v, t + 1) \tag{4.22}$$

where d is the optical flow, i.e. $d_u = u' - u$ and $d_v = v' - v$ where $(u', v', t + 1)$ is the correspondence of (u, v, t). Expanding Equation 4.22 using Taylor's series and retaining only the linear term as

$$I(u + d_u, v + d_v, t + 1) = I(u, v, t) + I_u + I_v + I_t \tag{4.23}$$

where I_u, I_v and I_t denote the spatial and temporal derivatives of image: $I_u = \frac{\partial I}{\partial x}u$, $I_v = \frac{\partial I}{\partial y}v$ and $I_t = \frac{\partial I}{\partial t}u$. These can be rewritten as the equation of a line:

$$dv(u, v, t) = -\frac{I_u(u, v, t)}{I_v(u, v, t)} du(u, v, t) - \frac{I_t(u, v, t)}{I_v(u, v, t)} \tag{4.24}$$

Since $d_u = u' - u$, and $d_v = v' - v$, this linear constraint can be defined as

$$v' - v = -\frac{I_u(u, v, t)}{I_v(u, v, t)} u' - \left(\frac{I_t(u, v, t)}{I_v(u, v, t)} - v - \frac{I_u(u, v, t)}{I_v(u, v, t)} u \right) \tag{4.25}$$

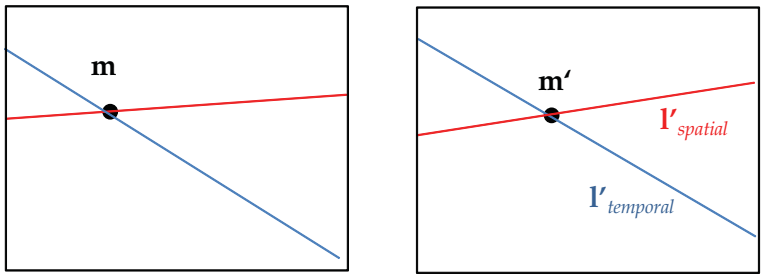


Figure 4.6. Spatio-temporal epipolar geometry

In homogeneous coordinates, we can obtain the *temporal line equation* of the optical flow.

$$\left[u', v', 1\right]^T \left[I_u, I_v, I_t - I_u u - I_v v\right] = \left[u', v', 1\right]^T \mathbf{l}'_{temporal} = 0 \tag{4.26}$$

where $\mathbf{l}'_{temporal} = \mathbf{F}\left[u, v, 1\right]^T$ denotes the temporal epipolar line. Spatial and temporal epipolar lines intersect at spatio-temporal correspondences as Figure 4.6 illustrates.

$$\tilde{\mathbf{m}}' = \mathbf{l}'_{spatial} \times \mathbf{l}'_{temporal} \quad \text{and} \quad \lambda\tilde{\mathbf{m}}' = \left[u', v', 1\right] \tag{4.27}$$

where λ is an arbitrary scale factor defined by Equation 4.10 and 4.11. Finally we can get the spatio-temporal collineation as

$$u' = \frac{(f_4 u + f_5 v + f_6)(I_t - I_u u - I_v v) - (f_7 u + f_8 v + f_9) I_v}{(f_1 u + f_2 v + f_3) I_v - (f_4 u + f_5 v + f_6) I_u} \tag{4.28}$$

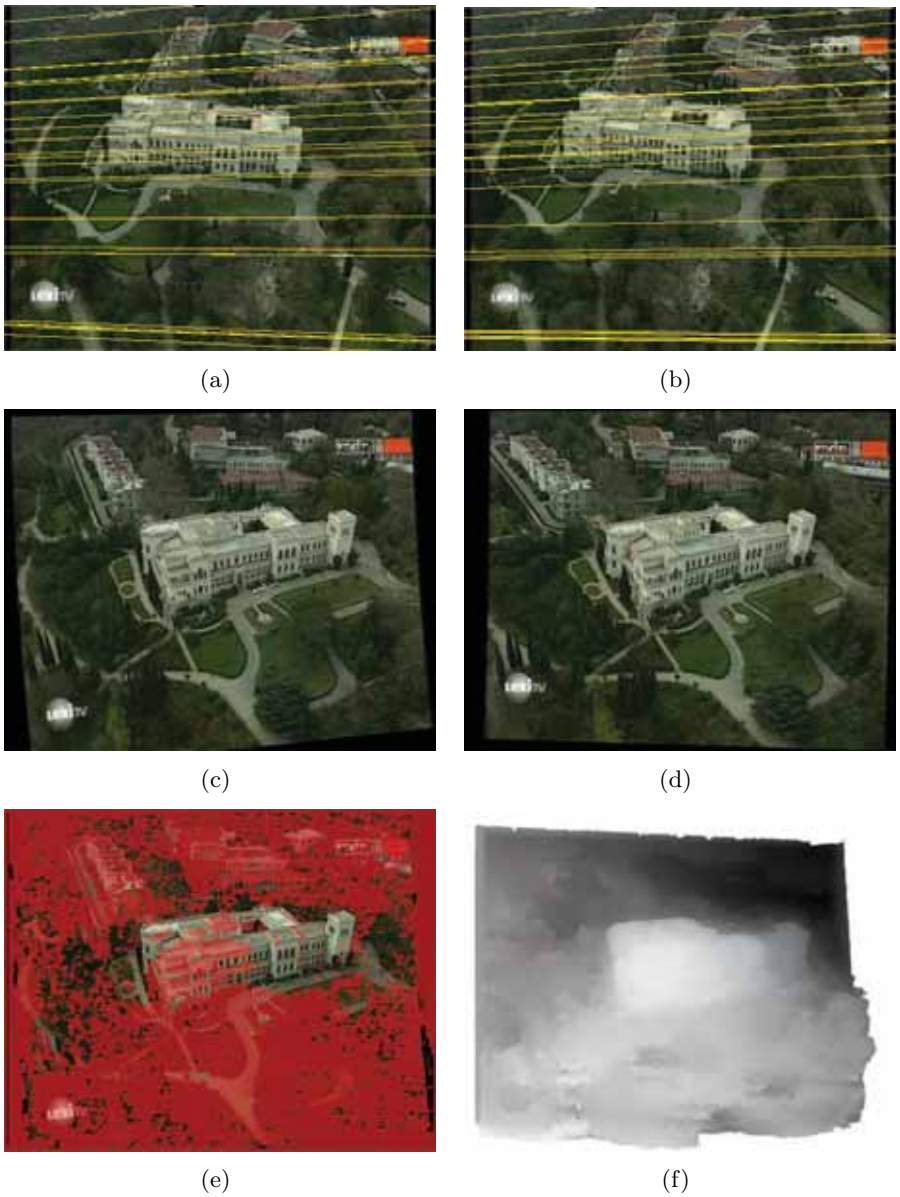
and

$$v' = \frac{(f_7 u + f_8 v + f_9) I_u - (f_1 u + f_2 v + f_3)(I_t - I_u u - I_v v)}{(f_1 u + f_2 v + f_3) I_v - (f_4 u + f_5 v + f_6) I_u} \tag{4.29}$$

4.9 Experimental Results

Epipolar geometry estimation of two views : When we know the geometry of the camera system as a priori, the set of all linear projection rays generates viewing geometry that includes several imaging errors with geometric distortions. These distortions are shown as divergent rays or camera aperture in the viewpoints and solved by a compensation using image correspondences. However, every image point in a viewpoint image is a potential match candidate for every point in the other viewpoint image, making a 2-D and large search problematic. Epipolar geometry provides the convenience of the epipolar constraint which reduces the search problem to 1-D.

Figures 4.7(a) and 4.7(b) show the images from two frames of an image sequence captured

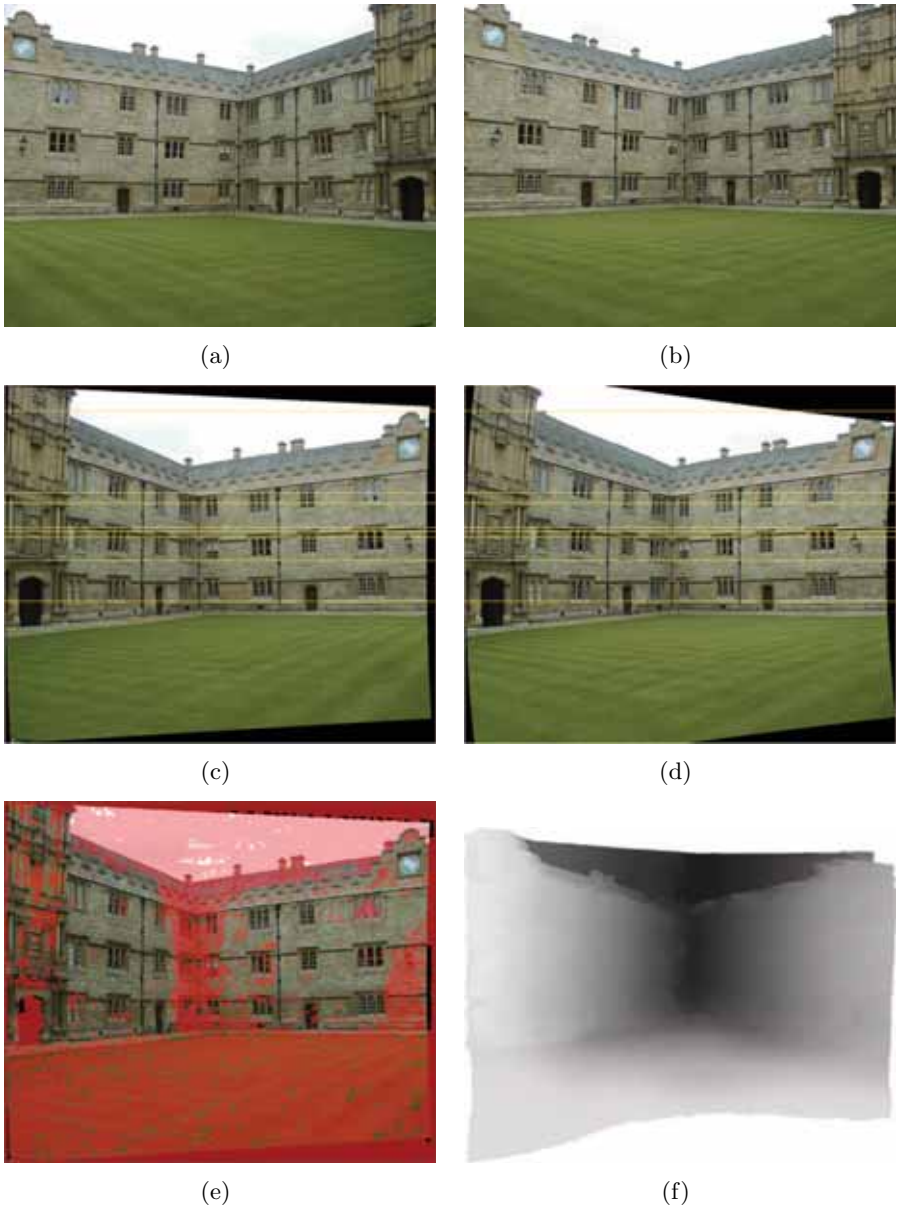


(a) (b)

(c) (d)

(e) (f)

Figure 4.7. Epipolar geometry estimation between two temporal frames - (a) epipolar lines of the 95th frame of a sequence captured by a moving monocular camera, (b) epipolar lines of the 115th frame, (c) the rectified 95th frame, (d) the rectified 115th frame, (e) the segmented region by the maximum bound of divergent rays, (f) the inverse rectified dense correspondences



(a) (b)

(c) (d)

(e) (f)

Figure 4.8. Epipolar geometry estimation between uncalibrated stereo images - (a) the left image captured of an uncalibrated stereo image pair, (b) the right image, (c) the rectified left image, (d) the rectified right image, (e) the segmented region by the maximum bound of divergent rays, (f) the inverse rectified dense correspondences

(a) (b)

Figure 4.9. Recovered 3-D model - (a) the recovered 3-D model using results in Figure 4.7, (b) the recovered model using Figure 4.8

with a video camera from a moving helicopter. The motion of the helicopter induces visual changes and large occlusion in a monocular sequence. The image resolution is 720×576 and the epipolar geometry is estimated using frames of the 95th and the 115th of the sequence. Features are tracked by the KLT tracker that was introduced in Equation 3.31 to establish temporal corresponding pairs. Compared to stereo matching, the motion correspondence is much easier to solve since tracking techniques can be much more reliable than large-baseline matching between image frames. However, the estimation using two consecutive frames alone is very noise sensitive and inaccurate. Hence, we use a frame-by-frame tracking of long sequence and use remote frames to estimate viewing geometry. For a large distribution of sparse features on the images, we eliminate unstable features using the RANSAC (random sampling consensus) method and calculate the remaining error for the obtained fundamental matrix. The matrix is iteratively updated to select the best with a minimum error. The yellow lines in Figures 4.7(a) and 4.7(b) are the epipolar lines, and each line includes a feature at least. Figures 4.7(c) and 4.7(d) represent the rectified images reformed from the epipolar lines. The size of the rectified images is 754×599. The object region can be segmented by thresholding the maximum bound of divergent rays as Figure 4.7(e) shows. The linear projection ray is generally stable in the background, and a ray from a foreground object is not the same to the linear projection ray. Figure 4.7(f) is the result from a dense matching with a maximum bound constraint.

Figures 4.8(a) and 4.8(b) show the left and right images captured by uncalibrated stereo cameras with toed-in setup. It is more difficult to establish correspondences between stereo images than motion. However, the view geometry estimation from stereo images has a smaller influence of noise than temporal frames since the spatial correspondences allow the relatively larger displacement than two temporal frames. The image size is 1024×768. Without representing epipolar lines on the original images, the evaluation can be possible on the rec-

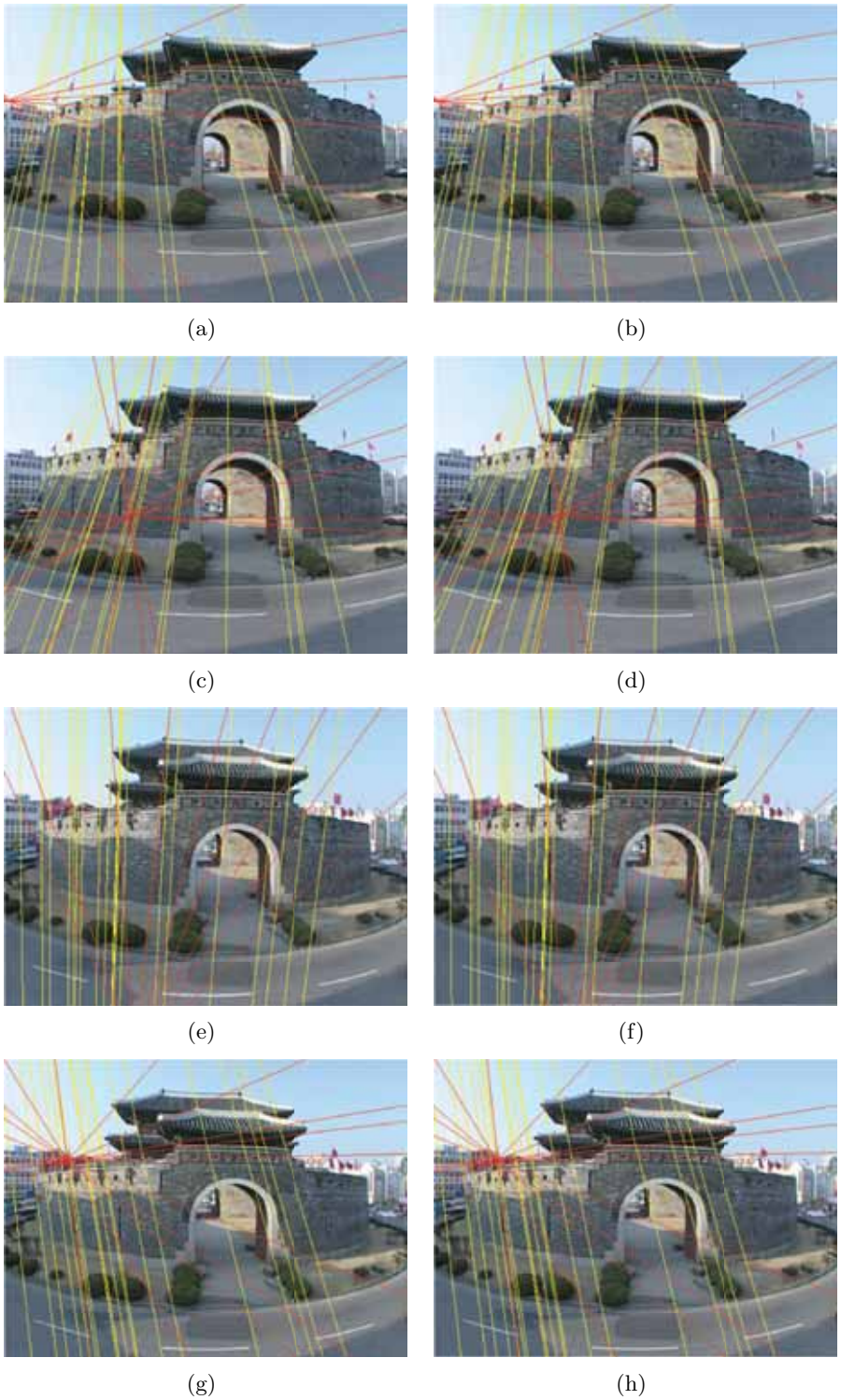


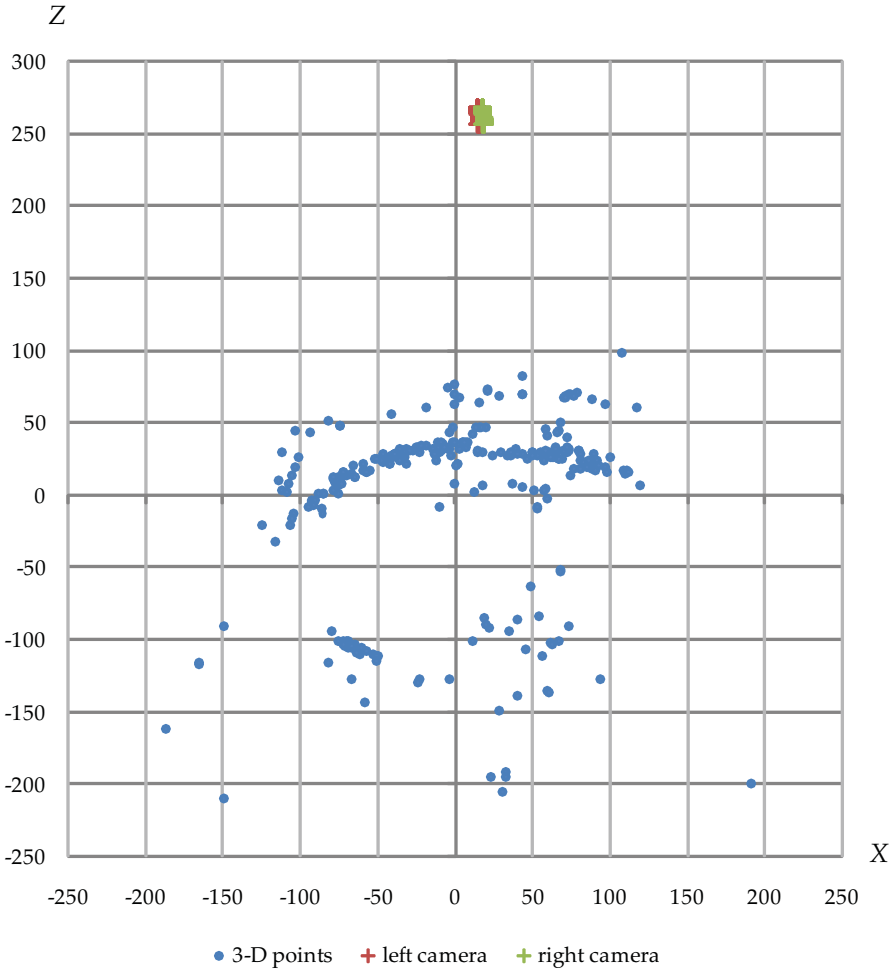
Figure 4.10. Spatio-temporal epipolar geometry using the sceneflow - (a) the epipolar lines of the left 16th frame of a stereo video sequence, (b) the 16th frame of the right image sequence, (c) the epipolar lines of left 51th frame, (d) the 51th frame of the right image sequence, (e) the epipolar lines of left 94th frame, (f) the right 94th frame, (g) the epipolar line of left 137th frame, (f) the right 137th frame

tified image. The yellow lines are the rectified epipolar lines that are parallel and satisfy the basic epipolar constraints. This evaluation of rectified epipolar lines is also possible in Figures 4.7(c) and 4.7(d)The rectified image has a resolution of 1047×880. Figures 4.8(c) and 4.8(d) respectively show the segmented object regions by a threshold of maximum bound and the inversely rectified dense correspondences. The 3-D models are recovered by a image-based modeling scheme using consistency checking of a dense depth map and a texture map. When the image points of dense disparity and viewpoints are projected into the depth map and texture map, a lot of points are overlapped. The samples are selectively interpolated by the Euclidean distance between the samples. Figures 4.9(a) and 4.9(b) show 3-D models recovered from Figures 4.7 and Figure 4.8 respectively.

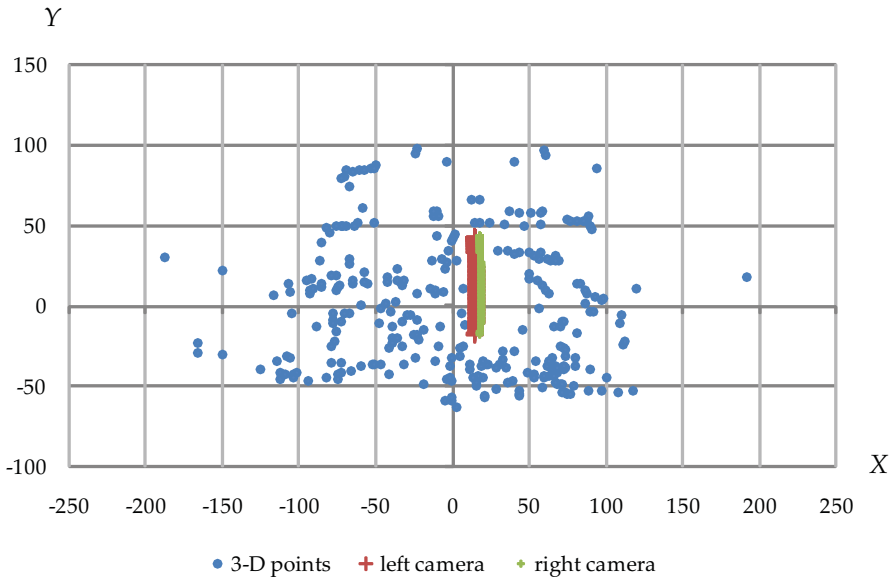
Spatio-temporal epipolar geometry estimation : It is difficult to establish feature correspondences between stereo images. The image brightness of a point in the scene does not remain constant from one camera viewpoint to another but remains constant over time for a single camera. Thus, we can apply optical flow to stereo image sequences, and the joint correspondences in the spatio-temporal image frames make very strong multi-linear constraints. We use two hand-held cameras to capture a scene with unknown camera geometry. Features are tracked by the KLT tracker and spatially matched with outlier removal using RANSAC. This method has the advantages of spatially large baseline and temporally dense correspondences. Figure 4.10 shows the estimated epipolar lines of temporal frames of a stereo video sequence. The epipolar lines with red color show the epipolar constraints of stereo camera, and the yellow epipolar lines represent the linear constraints for camera motion. The video has a resolution of 800×592 and 152 frames. Once we know the fundamental matrix and the position of the epipoles, the distances between correspondences can be measured, based on the epipolar constraints. 3-D points of a scene can be reconstructed as Figure 4.11 shows since the distances can be converted to the depths by triangulation. The centralized coordinate system[35] derived from the depth from viewpoints allows us to know the 3-D camera locations that are relative to the reconstructed 3-D points. Figures 4.12(a) and 4.12(b) represent the dense depth map and the 3-D model recovered from a stereo video in Figure 4.10 by using spatio-temporal dense correspondences with the multi-linear constraint. The $X - Z$ plot shows the relative depth between cameras and 3-D points, and the $X - Y$ plot shows the vertical movements of cameras.

Optical-ray-space representation using the dense 3-D sceneflow : We use the spatio-temporal view geometry for the virtual view synthesis. For the calibrated cameras and an arbitrarily defined virtual camera, the optical-ray in the 3-D space can be warped and ar-

35) The 3-D camera positions can be measured by the Euclidean distance between the 3-D points and camera positions measured from the world center $(0, 0, 0)$.

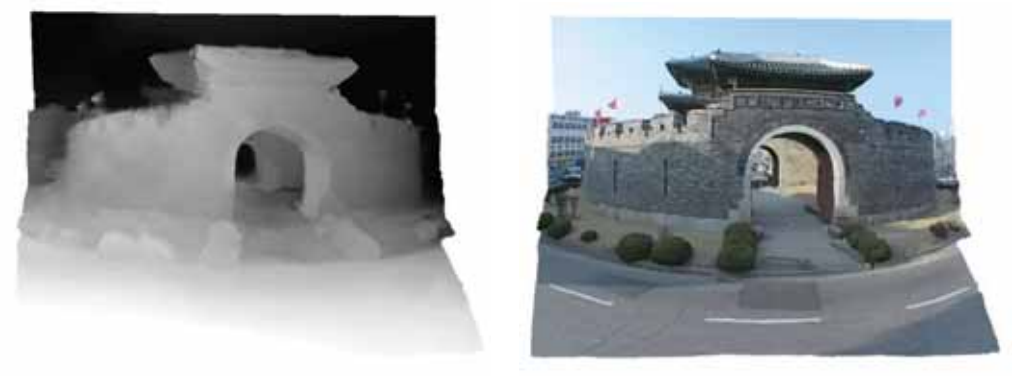


(a)



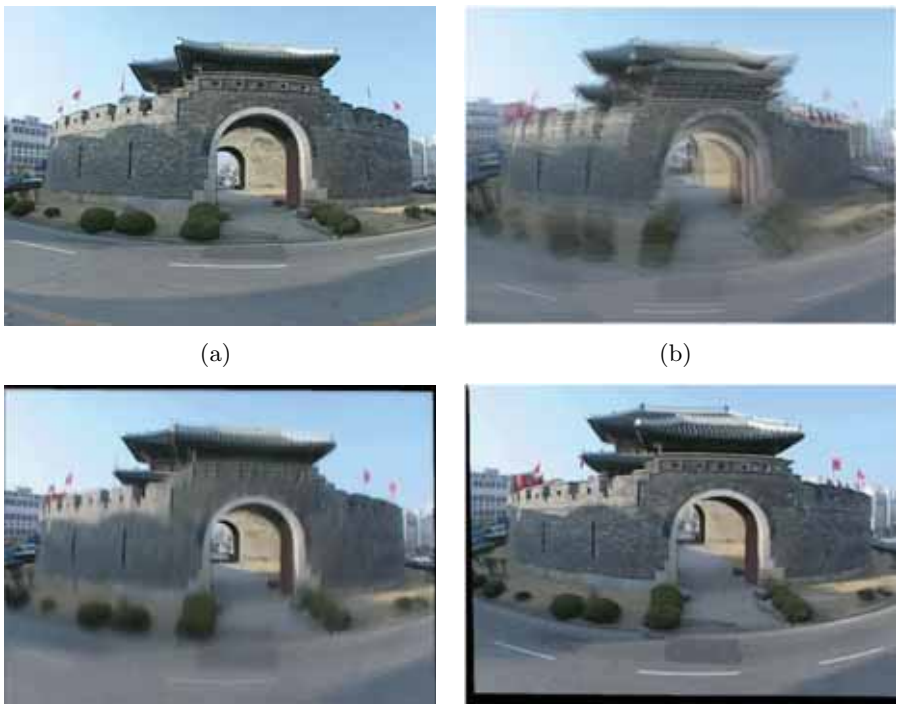
(b)

Figure 4.11. Relative 3-D position of cameras and tracked points for all stereo frames - (a) plot with X-Z axes, (b) plot with X-Y axes



(a) (b)

Figure 4.12. Recovered depth and 3-D model using sceneflow - (a) recovered dense depth map, (b) recovered 3-D model



(a) (b)

(c) (d)

Figure 4.13. Spatio-temporal interpolation in optical-ray - (a) the 1st frame of the stereo video sequence, (b) the spatio-temporally interpolated image, (c) spatio-temporal interpolation in optical-ray for an user-defined camera, (d) depth compensated spatio-temporal interpolation in optical-ray

ranged from the spatio-temporal image projections to the ray projection of the virtual camera. Figure 4.13(a) shows the 1st frame of the stereo video sequence, and the simplest temporal interpolation of 152 frames results in the blurring artifacts as the representation in Figure 4.13(b). However, the temporal interpolation in the optical-ray generates a view of near-infinite plane with only small distortions. Ray projections from the object surface to the camera center do not always fit to the global ray projection because the camera parameters define one global ray projection although the 3-D points of objects are laid on different depth planes. The reconstructed image still has some divergent rays as Figure 4.13(c) exemplifies. Thus, we use the dense 3-D sceneflow to approximate the sampling locations on the object surface which compensates the projective distortions of the divergent rays. Figure 4.13(d) shows the result that most distortions are eliminated.

In Figure 4.14, we compare the vertical sections of the optical-ray-spaces corresponding to Figures 4.13(b) and 4.13(d). A vertical scan-line is used to reconstruct the ray-space because the camera has large vertical movements, and Figure 4.14(a) shows the position of a vertical scan-line. The ray-space shown in Figure 4.14(b) has a lot of temporal variations rays that may cause blurring artifacts in the virtual view. The proposed method results in well rearranged rays as Figure 4.14(c) shows. An error is located in the upper part of the image, i.e. the roof of the castle, that is partially occluded and slowly appears due to the sampling error in depths. Sceneflow in the sky should be near zero since the depth of the sky should be near infinite. Since the sky in the image is a homogeneous region that is fundamentally difficult to establish a correspondence, the depth derived from sceneflow may not correspond to the true structure. This problem can be avoided by using pre-segmentation or a high image sampling rate, and we use the latter[36]. Figures 4.14(d) and 4.14(e) are the Fourier spectral domain images that visualize the frequency of spatial variations in the ray-spaces shown in Figures 4.14(b) and 4.14(c). We can transform the ray-space $v - t$ to Fourier spectral domain $\mathcal{F}_v$-$\mathcal{F}_t$. The Fourier spectral analysis of ray-space was already introduced with Figure 3.13 in Chapter 3.5. Figure 4.14(d) indicates larger ray variations in the ray-space than Figure 4.14(e).

36) Object segmentation is widely used in image-based modeling and tele-immersion studies to reduce the computation time or to restrict errors. However, it is very difficult to extract automatically an object of interest that consists of several image regions from a natural scene without prior information.

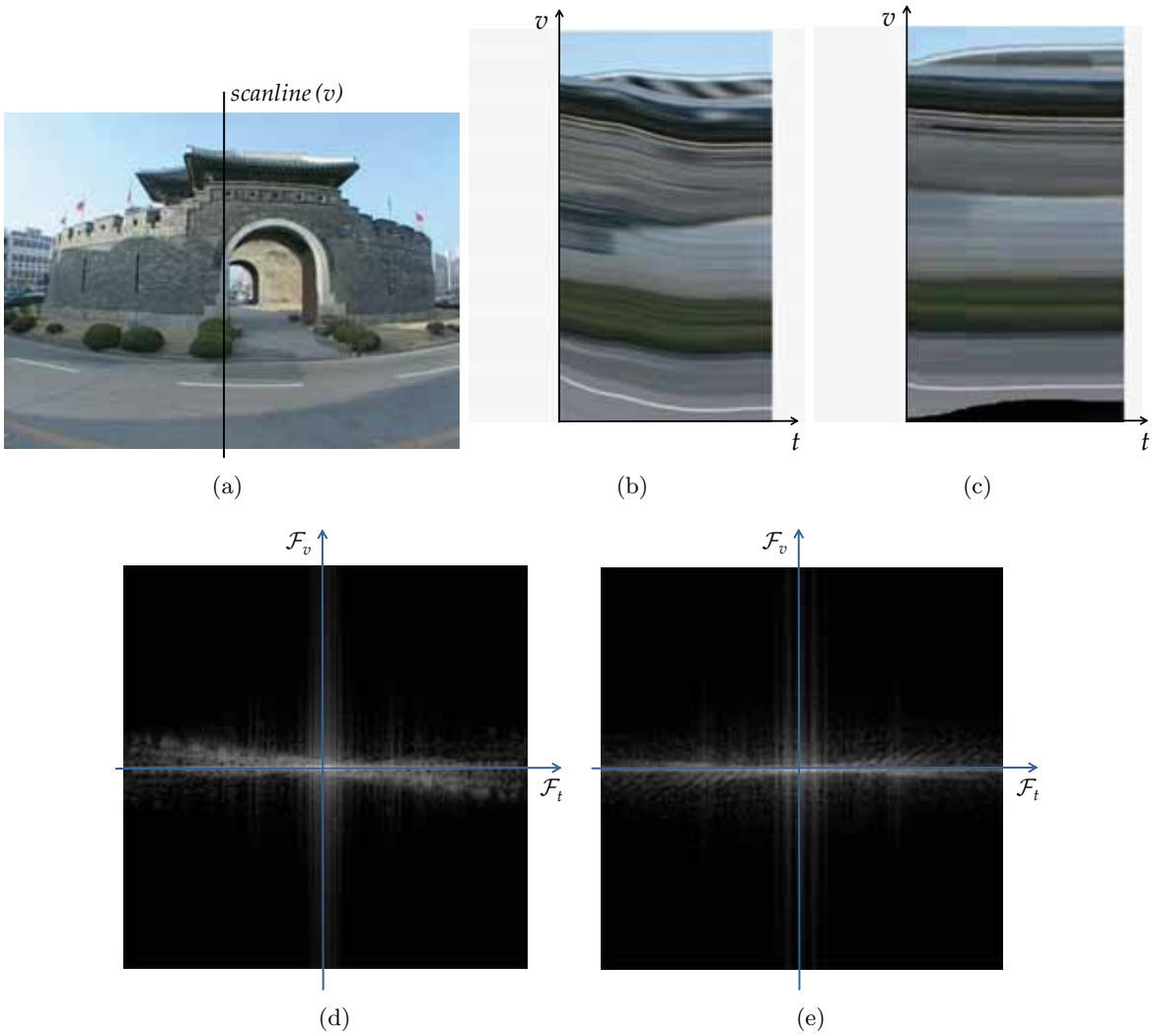


Figure 4.14. Spatio-temporal ray-space - (a) A vertical scan-line (b) ray-space of Figure 4.13b, (c) ray-space of Figure 4.13d, (d) Fourier spectral domain of Figure 4.14b, (e) Fourier spectral domain of Figure 4.14c

Chapter 5

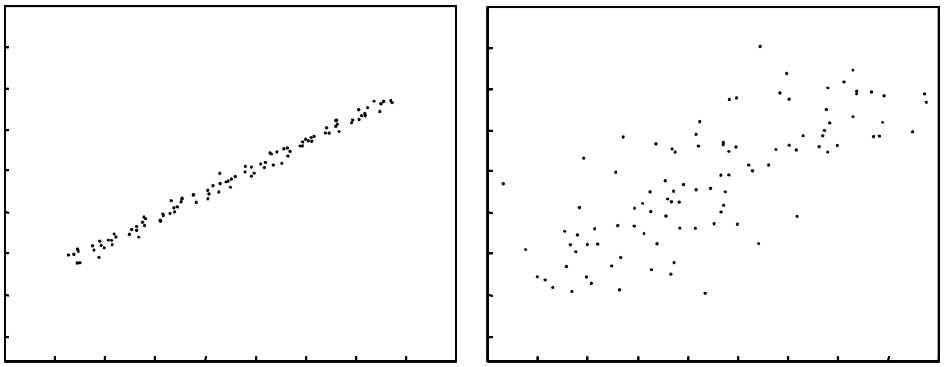
Unified Representation of Robust Estimations

Estimation techniques in 3-D computer vision applications need to estimate accurate parameters despite noise and measurement errors that may be caused by specular highlights, saturation of sensors, or mistakes in establishing correspondences. In the last twenty years, many robust techniques have been applied to the standard problems in computer vision. There are robust approaches for performing surface fitting [107], motion estimation [108, 109], epipolar geometry [110, 111], image alignment [112], segmentation [113, 114], edge detection [115, 116], etc. In this dissertation, the final targets are obtaining dense depth compensated video-based modeling and rendering, and there are two important problems:

(a) Dense image-to-image correspondences are not accurate because the matching process is unreliable, and a parameter estimation based on the data are heavily over-constrained. Robust parameter estimation of the epipolar geometry is important since the parameters can be important constraints to reconstruct the scene geometry up to a projective transformation.

(b) Even when we can obtain accurate camera parameters using some reliable low level features, e.g. edge, line and corner etc., it is not sufficient to reform rays diverged from the optical axis due to the nonlinear scene geometry. Robust estimation of dense correspondence is still important.

When errors are large, it is very difficult to discriminate inliers from data including a lot of outliers. This chapter states clearly what is meant by robustness and describes how the estimator can be robust. And, a unified representation of robust estimator will be introduced for a robust parameter estimation and robust regularization of dense scene geometry.



(a) (b)

Figure 5.1. The importance of the scale - (a) data with a small scale of noise, (b) data with a large scale of noise

5.1 Robustness of the Estimator

Due to physical imperfections in imaging sensors and errors in low-level vision computations such as detection and matching algorithms, the image data may always contain outliers. In the estimation of the fundamental matrix as described in previous chapters, extracting the correct matches is difficult since the bad location of detected features[37] and the false matches[38] always exist in the correspondence procedure. In the estimation of the fundamental matrix, the location error of a point of interest is assumed to exhibit Gaussian behavior. This assumption is reasonable when the localization error for most points of interest is small, i.e. within one or two pixels. However some points are possibly incorrectly localized with more than three pixels, the presence of outliers severely affects the precision of the fundamental matrix. The existence of outliers increases interests in robust estimators which can determine inlier simultaneously with the estimation of parameters.

A robust estimator considers the relation between the data points and reasonable model candidates as a *scale* and outliers can be removed by different contributions of scales for the data of interest [33, 34]. Figures 5.1(a) and 5.1(b) exemplify the importance of scaling using the same data only in two different scales of the noise. Data in a large scale is more difficult to define the data of interest than a small scale. A typical example of the non-robust estimator is the total least squares estimation whose objective function can be expressed as

$$\int_{i=1,\cdots,n} d_i^2 \tag{5.1}$$

37) The bad location means the detected feature points are not correctly localized.
38) False matches are caused by the false heuristics in the establishment of correspondences.

where d_i is the distance between measurement and its true value and the parameter estimates are obtained by minimizing the sum of squared residuals. If we assume that the noise corrupting the data is of zero mean, which yields an unbiased parameter estimate. If the noise variance is known, a minimum variance parameter estimate can be obtained by choosing appropriate weights on the data. A robust estimator includes an additional assumption in the objective function [117] as

$$\int_{i=1,\cdots n} \left(\frac{1}{\sigma} d_i\right)^2 \tag{5.2}$$

where the function is homogeneous if the scale σ of the inlier noise does not play any role in the main estimation process. Instead of $\left(\frac{1}{\sigma}\right)^2$ in Equation 5.2, a robust estimator employs a *bounded loss function* ρ_σ as

$$\int_{i=1,\cdots,n} \rho_\sigma \left(d_i^2\right) \tag{5.3}$$

$\rho_\sigma(d)$ depends on σ and satisfies the following properties:

(a) Symmetric, $\rho_\sigma(d) = \rho_\sigma(-d)$,
(b) Positive-definite function with a unique minimum at zero, $\rho_\sigma(d) > 0$ and
(c) Less increasing with $|d|$ than square.

When $\rho_\sigma(u) = u^2$ yields the total least squares solution and $\rho_\sigma(d) = |d|$ yields the least absolute values (L^1) regression. However, we are more interested in the bounded loss functions which are characterized as

$$\begin{cases} 0 \leq \rho_\sigma(d) \leq 1 \\ \rho_\sigma(d), \quad |u| > 1 \end{cases} \tag{5.4}$$

5.2 Robust Estimation with Random Sampling

The objective functions in most robust estimators often have many local extrema. The optimization procedure starts from several initial positions and finds the desired solution, which locates in the candidate space of all possible distinct elemental subsets. In numerical technique, the random sampling is often used to obtain a close to optimal solution since the number of possibly distinct elemental subsets in the data can be very large. It reduces the amount of computations in the outlier rejection capability of the implemented robust estimator. Most robust estimators with random sampling run through the following procedure:

(1) Select p parameters, i.e. the number of point correspondences in each subsample, randomly.
(2) Calculate the residual to the model for each parameter.

Table 5.1. Several kinds of ρ, ψ and w functions.

	$g(d)$	$\psi(d) = g(d)d$	$\rho(d) = \int g(d)d\,dx$
Least − squares	1	d	$d^2/2$
Huber $\begin{cases} \lvert d\rvert \le c \\ \lvert d\rvert > c \end{cases}$	$\begin{matrix} 1 \\ c/\lvert d\rvert \end{matrix}$	$\begin{matrix} d \\ c \cdot \mathrm{sign}(d) \end{matrix}$	$\begin{matrix} d^2/2 \\ c(\lvert d\rvert - c/2) \end{matrix}$
Lorentzian	$\frac{1}{1+(d/c)^2}$	$\frac{d}{1+(d/c)^2}$	$\frac{c^2}{2}\log\left(1+(\frac{d}{c})^2\right)$
Tukey $\begin{cases} \lvert d\rvert \le c \\ \lvert d\rvert > c \end{cases}$	$\begin{matrix} 1-\left[1-\left(\frac{d}{c}\right)^2\right]^2 \\ 0 \end{matrix}$	$\begin{matrix} d\left[1-(1-(\frac{d}{c})^2)^2\right] \\ 0 \end{matrix}$	$\begin{matrix} \frac{c^2}{6}\left[1-(1-(\frac{d}{c})^2)^3\right] \\ (c^2/6) \end{matrix}$

(3) The efficiency is defined as the ratio between the lowest achievable variance for the estimated parameters and the actual variance provided by the given method. Sort the squared residuals and select a scale which acts as a bounded loss function for the estimated model.

(4) The scale performs a consistency check of point correspondences which are considered as outliers.

(5) Repeat the previous steps t times, i.e. t is the number of subsamples, and calculate the probability P that none of these subsets contains only inliers:

$$P = 1 - [1 - (1 - \varepsilon)^p]^t \tag{5.5}$$

where ε is the maximal proportion of outliers. By requiring that the probability P must be near 1, t can be determined by the measurement of the outlier rejection capability for given values of p and ε:

$$t = \frac{\log(1-P)}{\log|1-(1-\varepsilon)^p|} \tag{5.6}$$

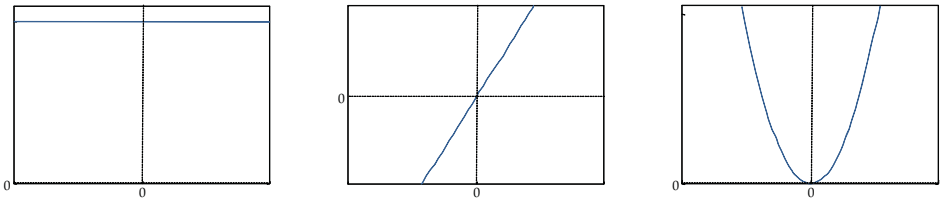
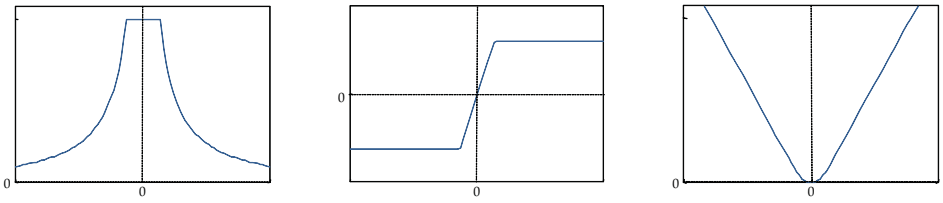
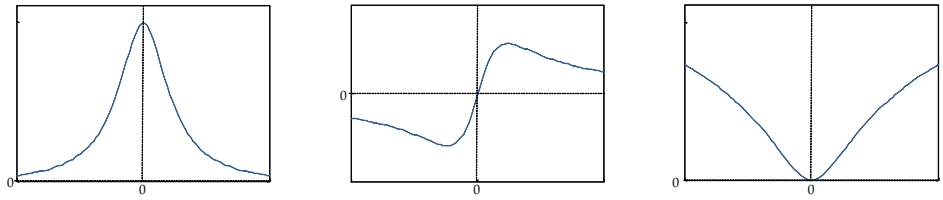
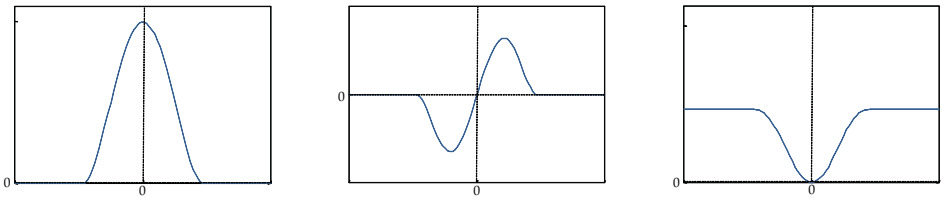
5.3 Unified Representation of Robust Fundamental Matrix Estimation

The theory of robust estimators was developed in statistics and three famous methods, M-estimators [117, 118] and least median of squares (LMedS) [119, 120] and random sample consensus (RANSAC) [110, 120–122], were successfully applied to solve a lot of computer vision problems. All of these robust estimators can be regarded as particular cases of M-estimators[39] with auxiliary scales.

Instead of the function in Equation 5.3, we can solve the following iterated reweighted least-

39) The M-estimator was first proposed in 1964, by Peter Huber [34] as a maximum likelihood estimator which depends on the assumed distribution family of the data being at least approximately true.

Table 5.2. Graphs of ρ, ψ and w functions.

$w(d)$	$\psi(d) = w(d)d$	$\rho(d) = \int w(d)d\,dx$
		
Least-squares(non-robust)		
		
Huber		
		
Lorentzian		
		
Tukey		

squares problem based on $\rho(d_i)$ to estimate the parameter vector.

$$\rho(d_i) = \int_{i=1,\cdots n} g(d_i^{(k-1)}) d_i^2 \tag{5.7}$$

where $g(d) = \frac{\psi(d)}{d}$ is the *redescending weight function* that higher weights are given to the measurements with small residuals. The derivative $\psi(d) = \frac{\partial \rho(d)}{\partial d}$ is called the *influence function*. The influence function measures the influence of a datum on the value of the parameter. The objective function of the parameter vector which is minimized should have a unique minimum. This requires that the ρ-function is convex. The superscript $^{(k)}$ indicates

the iteration number. In the weight function, only those data points that are at a distance less than from the current fit are taken into account. There are several kinds of redescending weight functions [123, 124] as Table 5.1 shows. Table 5.2 represents the graph of the functions. For the least-squares with $\rho(d) = d^2/2$, the influence function $\psi(d) = d$ is that the influence of a datum on the estimate increases linearly with the size of its error. However, Tukey's biweight function considers some estimated standard deviation σ of inlier errors.

$$\rho(d_i) = \int g(d_i) d_i dx = \begin{cases} \frac{c^2}{6}\left(1 - \left[1 - \left(\frac{d_i}{c\sigma}\right)^2\right]^3\right) & \text{if } |d_i| \leq c\sigma \\ (c^2/6) & \text{otherwise} \end{cases} \tag{5.8}$$

where the constant $c = 4.6851$ is the normalization factor for a Gaussian distribution and the weight function w is

$$g(d_i) = \begin{cases} \left[1 - (d/c)^2\right]^2 & \text{if } |d_i| \leq c\sigma \\ 0 & \text{otherwise} \end{cases} \tag{5.9}$$

In the redescending M-estimators, the scale parameter σ can be estimated at every iteration and acts as a threshold for the inlier/outlier separation.
The scale of the robust estimator can be utilized for the estimation of the epipolar geometry. The fundamental matrix estimation proceeds the following steps:

(1) A Monte Carlo type technique is used to draw t random subsamples of correspondences.
(2) For each subsample, indexed by J, we compute the fundamental matrix $\mathbf{F}_J$.
(3) For each $\mathbf{F}_J$, residuals d_J are determined with respect to the whole set of point correspondences.
(4) Retain the estimate $\mathbf{F}_J$ for which d_J is minimal among all t d_J.
(5) Assign a weight g_i based on σ for each correspondence.
(6) Refine the fundamental matrix $\mathbf{F}$ by bringing all $\mathbf{F}_J$ into the bounded loss function $\int_{i=1,\cdots,n} \rho_i$

$$\min \sum_{i=1}^{n} g_i d_i$$

where $\rho_i = g_i d_i$

Median Absolute Deviation (MAD) estimator : Given n-point correspondences, i.e. $[(\mathbf{m}_i, \mathbf{m}'_i) \,|i = 1, \cdots, n]$, the robust standard deviation estimate is related to the median of the absolute values of the residuals [117]. Median Absolute Deviation (MAD) estimator measures the scale based on the spread of the residuals around the median.

$$c \cdot \underset{i=1,\ldots,n}{\text{median}} |d_i| \tag{5.10}$$

where d_i denotes the distance between the i-th points and epipolar lines.

$$d_i = d\left(\tilde{\mathbf{m}}_i', \mathbf{F}_J\tilde{\mathbf{m}}_i\right) + d\left(\tilde{\mathbf{m}}_i, \mathbf{F}_J^T\tilde{\mathbf{m}}_i'\right) \tag{5.11}$$

Least Median of Squares (LMedS) estimator : The LMedS method uses the squared residuals d^2 instead of absolute residual d.

$$d_i = \left[d^2\left(\tilde{\mathbf{m}}_i', \mathbf{F}_J\tilde{\mathbf{m}}_i\right) + d^2\left(\tilde{\mathbf{m}}_i, \mathbf{F}_J^T\tilde{\mathbf{m}}_i'\right)\right] \tag{5.12}$$

A weight is applied to each correspondence.

$$g_i = \begin{cases} 1 & \text{if } r_i^2 \leq (2.5\hat{\sigma})^2 \\ 0 & \text{otherwise} \end{cases} \tag{5.13}$$

where $\hat{\sigma} = 1.4826\left[1 + 5/(n-t)\right]\sqrt{d_J}$ and the constant 1.4826 is set to achieve the same efficiency as least-squares in the presence of only Gaussian noise[40]. Point correspondences having $w_i = 0$ are considered as outliers. Although this outlier removal is very similar to the RANSAC method, a major difference is that the LMedS the scale is computed from a condition set on the percentage of inliers, and RANSAC requires a prefixed threshold provided by the user.

RANdom SAmple Consensus(RANSAC) method : The redescending loss function in RANSAC is defined by the zero-one function.

$$\rho(d_i) = \begin{cases} 1 & \text{if } |d_i| \leq c \\ 0 & \text{otherwise} \end{cases} \tag{5.14}$$

This zero-one loss function often yields very poor properties since it is not continuous and does not have a unique minimum in $d = 0$. For a critical arrangement of the n points, a slight change in the value of a measurement yields an unexpected large change in the value of the estimate. Several variants that the zero-one loss function is replaced by a smooth function has been studied in [111, 121, 125]. For example, M-estimator sample consensus (MSAC) method [121] replaces the zero-one loss function with the skipped mean.

$$\rho(d_i) = \begin{cases} 1 & \text{if } |d_i| \leq c\sigma \\ 0 & \text{otherwise} \end{cases} \tag{5.15}$$

40) The median of the absolute values of random numbers sampled from the Gaussian normal distribution $N(0,1)$ is equal to $\Phi_{-1}\left(\frac{3}{4}\right) \approx \frac{1}{1.4826}$.

The MAPSAC estimator [125] uses the loss function in a maximum posterior formulation of RANSAC. A maximum likelihood motivated variant, the MLESAC [111], is developed to evaluate the likelihood of the hypothesis, representing the distribution of residuals as a Gaussian mixture model.

5.4 Anisotropic Regularization using Robust Estimator

In 3-D computer vision, the discontinuity detection is important since the discontinuity in a viewpoint image is the position which includes information separating the bundles of rays reflecting from different surfaces. Two approaches have been separately developed: edge-based and area-based approaches. Both approaches detect local changes in the intensity of the image. However, these techniques are not satisfactory for localization performance due to the sensitivity of noise.

Edges in the Gaussian scale-space : Canny [126] proposed a successful edge detector which uses the approximation by the first derivative of a Gaussian. By convolving an image with a single Gaussian-filter, the edges are detected very precisely at the maximum of the gradient magnitude in the Gaussian-smoothed image. Witkin [127] developed the Canny edge detector into an iterative method for multi-resolution processes, i.e. Gaussian scale-spaces. A coarser resolution image is iteratively generated by convolving the image with a Gaussian-filter kernel. The scale parameter of Gaussian-filter kernel is used to control the boundary strength and the direction to be detected. This parameter tuning results in strong boundaries only, which have larger values than can be detected in the specified scale [128].

$$\begin{aligned} I_\sigma(u,v) &= \mathcal{G}_\sigma * I(u,v) \\ &= (2\pi\sigma^2)^{-1} e^{-(u^2+v^2)/(2\sigma^2)} * I(u,v) \end{aligned} \tag{5.16}$$

where $I_\sigma(u,v)$ is the image intensity filtered by the local convolution kernel of the Gaussian $\mathcal{G}$ at a scale-level σ. The first derivative of Gaussian along the x-axis is given in Equation 5.16, and the other along the y-axis is obtained in a similar way. Figure 5.2(a) shows a Gaussian scale-space using a horizontal scanline u and a different scale parameter σ. Edges in the scale-space are obtained by a coarse-to-fine method as Figure 5.2(b) represents.

$$\frac{\partial I_\sigma}{\partial u} = \frac{-u}{\sqrt{2\pi\sigma^3}} e^{-u^2/2\sigma^2} = -\frac{u}{\sigma^2} I_\sigma \tag{5.17}$$

Since the 2-D gradient is a first-order operator $(\partial/\partial x, \partial/\partial y)$ defined as a vector, the direction of the gradient is defined by the Euclidean norm of gradient which consequently indicates the

(a)

(b)

Figure 5.2. Gaussian scale-space - (a) Gaussian scale-space for a horizontal scanline, (b) edges in the Gaussian scale-space

strength of the intensity change as

$$|\nabla I_\sigma(u,v)| = \left| I(u,v) * \frac{\partial}{\partial N} I_\sigma(u,v) \right| = \sqrt{\left(\frac{\partial I_\sigma(u,v)}{\partial u}\right)^2 + \left(\frac{\partial I_\sigma(u,v)}{\partial v}\right)^2} \tag{5.18}$$

where N represents a unit vector towards an arbitrary gradient direction. The gradient direction is perpendicular to the edge orientation and localizes the boundary. Thus, the equivalent function considering several directional derivatives can be easily expressed with respect to a polar coordinate system where $r(\theta) = \sqrt{u^2 + v^2}$ represents the radial distance from the origin. The function is symmetrical and independent of the orientation θ. The prediction error is defined as

$$E[I_\sigma(r), \theta] = |I_\sigma(u + s\cos\theta, v + s\sin\theta) - I_\sigma(u,v)| \tag{5.19}$$

where s is the distance of the prediction that is proportional to the scale. If we consider two possible directions for the morphology of a boundary, forward is θ and backward is given as $(\theta + \pi)$. The probability P is assigned in a proportion to their prediction errors [129].

$$P[I_\sigma(r), \theta] = \frac{E[I_\sigma(r), \theta]}{E[I_\sigma(r), \theta] + E[I_\sigma(r), (\theta + \pi)]} \tag{5.20}$$

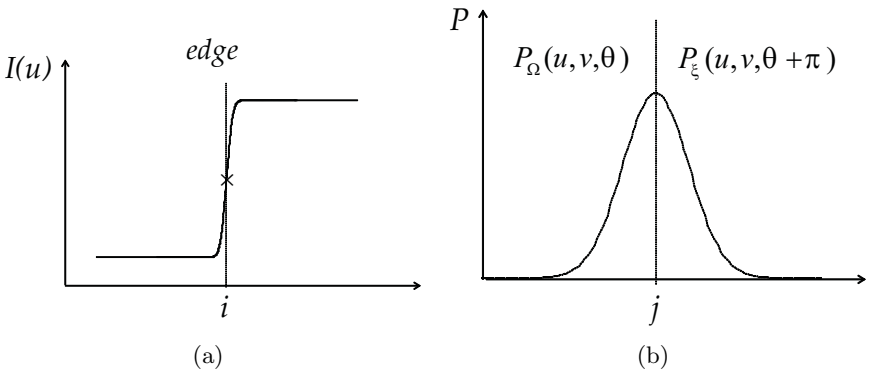


Figure 5.3. Hierarchy of edges in Gaussian scale-space - (a) an edge position i in a filtered scanline, (b) local maxima point j of intensity changes

A large prediction error in a certain direction implies a higher possibility to identify local maxima of intensity changes by analyzing the total certainties over half circles as

$$\arg\max_{\theta} \int_{\theta-\pi/2}^{\theta+\pi/2} P[I_{\sigma}(r), \theta] d\theta \tag{5.21}$$

In Figure 5.3, the optimal position of a boundary i of a scale is estimated to minimize errors by the highest possibility, thus i is located at the maxima of intensity changes j. Since the maximum of change directly implies the position of discontinuity, i is exactly located at the position of discontinuity on the scale.
The smoothness condition of Gaussian scale-space is efficient to detect an object region that consists of several piecewise weak edges, e.g. a texture object [132, 133]. On a coarse resolution, the piecewise weak edges are removed, and an object can be easily determined by regions with some strong edges. A fine resolution can be used to fit the coarsely detected edges to the position of real edges. Gaussian scale-space is a representation of a coarse-to-fine morphological hierarchy that has a trajectory of the image discontinuity. Figure 5.4 exemplifies the morphological course-to-fine hierarchy of a scale-space. The original image in Figure 5.4(a) has 5 different texture regions. In a course resolution shown in Figure 5.4(b) edges of the regions can be detected by local maxima. However, the position of edge is not exact. Although a fine resolution does not allow detecting edges of texture regions, the position of edges is exact.

Area-based approach : The simplest area-based methods to find a discontinuity is thresholding. However, a single threshold is not sufficient to divide the image areas with the similar photometric characteristic, and the analysis of semantic units is additionally needed. Region growing methods [134] employs a similarity measure between neighbored pixels to find a con-

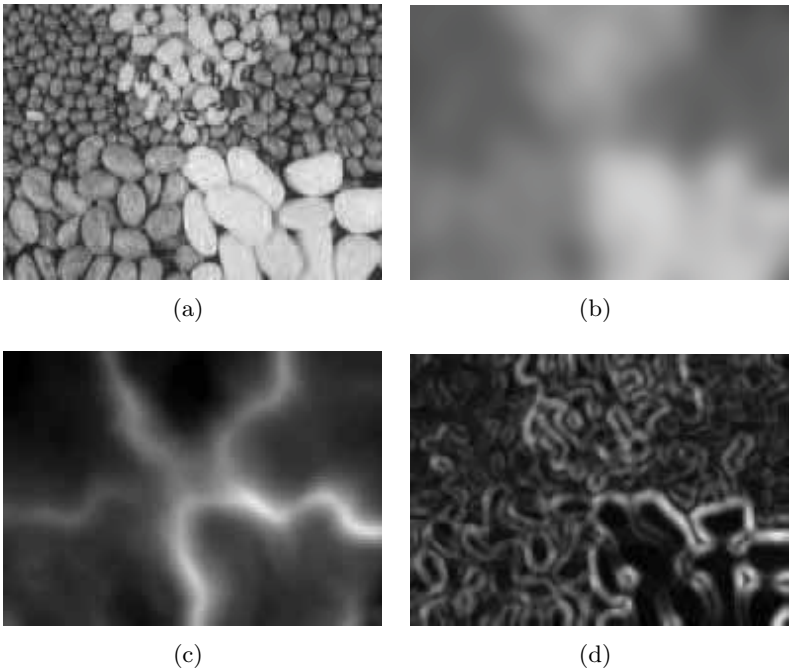


Figure 5.4. Morphological course-to-fine hierarchy in a texture image - (a)texture image with 5 regions, (b) filtered image into a coarse resolution, (c) local maxima with a coarse resolution, i.e. note that result is without thresholding of edge strength, (d) local maxima with a fine resolution

nected area. This approach can define edges being laid on a segment border by a form of closed contour. However, the method cannot determine the segment size, and thus merges some important features in one segment while other image areas are still splitted. Morphology methods [135, 136] solve these problems by adapting a flexible region into a discontinuity by using erosion and dilation. Histogram equalization can be efficiently applied to the method for the manipulation of local contrasts. However, the noise sensitivity of the gradient operation is the main problem. If the segment sizes are very small, the uniform areas may be over-splitted. A low-pass filtering the gradient image can be a solution because if the noise is reduced, the segments can easily grow. A localization problem due to the shifted positions of a discontinuity is occurred in the smoothing process[41].

Oriented Laplacian : The oriented Laplacian [130] describes a smoothly varying oriented structure. The unit vectors η and ξ are respectively defined by the gradient direction i.e. streamline direction and its orthogonal direction i.e. tangent and isophote direction for scalar

41) This problem is related to the ambiguity of blurred edge.

image.

$$\eta = \theta_+ = \frac{\nabla I_\sigma}{\|\nabla I_\sigma\|} \quad \text{and} \quad \xi = \theta_- = \frac{\nabla I_\sigma}{\|\nabla I_\sigma\|}^{\perp} \tag{5.22}$$

Two variation orientation θ_+ and θ_- are corresponded into respectively orthogonal and tangent to the vector edges. $\mathbb{N}$ is a variation norm to detect the edges. The second derivative operators $I_{\xi\xi}$ and $I_{\eta\eta}$ of I_σ is used to track the discontinuity of the vector field.

$$\frac{\partial I_\sigma}{\partial t} = c_\xi \frac{\partial^2 I_\sigma}{\partial \xi^2} + c_\eta \frac{\partial^2 I_\sigma}{\partial \eta^2} = c_\xi I_{\sigma_{\xi\xi}} + c_\eta I_{\sigma_{\eta\eta}} \tag{5.23}$$

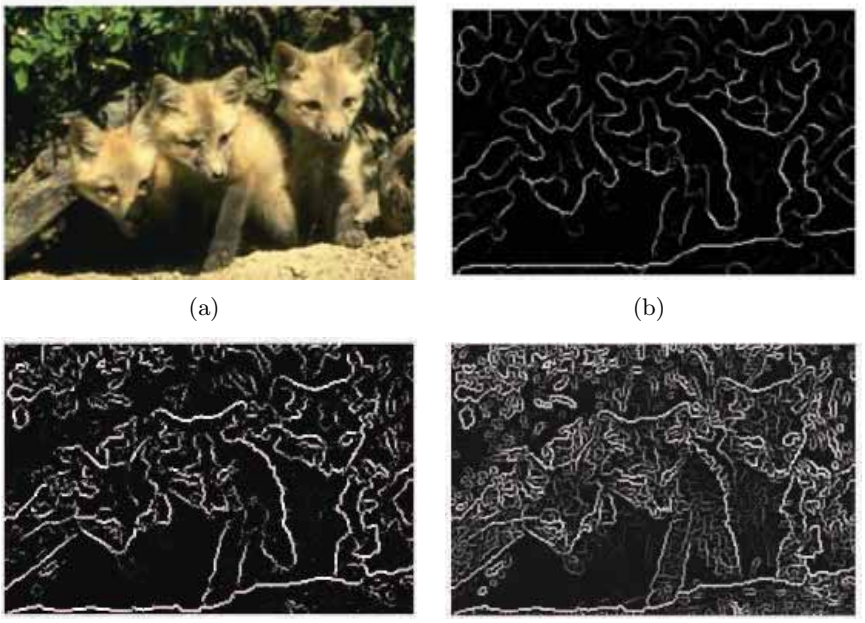
This second derivative expresses an oriented Laplacian formulation that is interpreted as two coexistent and oriented smoothing with their corresponding weights c_ξ and c_η following their perpendicular directions $\xi \perp \eta$. The edges are probabilistically marked as the location where two vectors diverge from each other in opposite directions. The probability is assigned in a proportion to their prediction errors. Figure 5.5 shows the edges detected by an oriented Laplacian with iterative isotropic Gaussian filtering. The oriented Laplacian formulation yields the basis of edge and region connection based on the diffusion theory. While an image is smoothed by an iterative filtering, a region s is progressively expanded following the direction of the flow vector $\vec{V}_\sigma(s)$. The expansion is stopped by the Laplacian characteristic $\vec{\nabla} \cdot \vec{V}_\sigma(s) = -\Delta V_\sigma(s)$. Figure 5.6 exemplifies the deformation of connected regions with some iteration. For a multi-scale representation of Gaussian scale-space, these techniques adaptively estimate the neighborhood size and the spatial scales with a parameter which specifies a desired homogeneity level within the regions. An edge flow scheme [129, 131] track meaningful discontinuity edges which exist at whole multiple scales and suppress edges which only exist at fine scales.

Anisotropic diffusion : Anisotropic diffusion methods solve this localization problem by minimizing an energy function that penalizes a high gradient. An anisotropic diffusion method is applied to the images of Gaussian scale-space.

$$\min_{I:\Omega} E(I) = \int_\Omega g\left(\|\nabla I_\sigma\|\right) d\Omega \tag{5.24}$$

where ω is the image domain, and $g : \mathbb{R} \to \mathbb{R}$ is a decreasing function vanishing on high gradient edge to stop the smoothing while the low gradient region is isotropically smoothing in Gaussian scale-space. The minimization is numerically performed by diffusive evolution with PDE equation as

$$\frac{\partial I}{\partial t} = div\left[\left(\frac{g(\|\nabla I_\sigma\|)}{\|\nabla I_\sigma\|}\right)\nabla I\right] \tag{5.25}$$



(a) (b)

(c) (d)

Figure 5.5. Oriented Laplacian of iterative blurred images - (a) a texture image, (b) with 10 iterations, (c) with 5 iterations, (d) with a single calculation

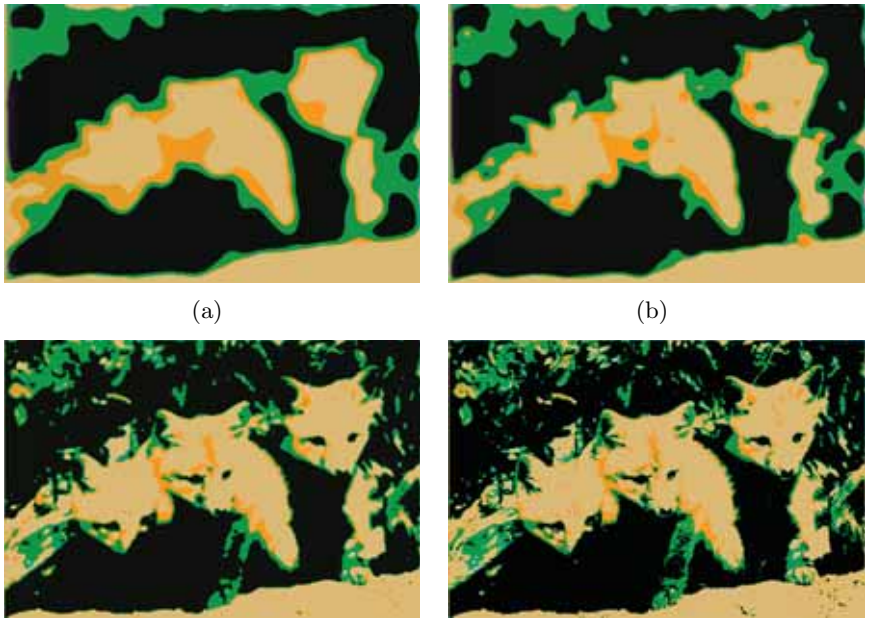
The diffusion characteristic of the g-function can be analyzed by the geometric development of a weighted gradient. The g-function is a redescending weight function of the robust estimator described in Chapter 5.3. The several kinds of weight functions were shown in Tables 5.1 and 5.2. Similarly to Equation 5.7, the weighting function g gives higher weights to the measurements with small gradients. The form $g(\|\nabla\|)\nabla$ in Equation 5.25 can be described by an influence function $\psi(\nabla) = g(\|\nabla\|)\nabla$, and Equation 5.24 defines a bounded loss function $\rho(\nabla)$.

$$\rho(\nabla) = \int_{\Omega} \psi(\nabla) d\Omega = \int_{\Omega} g(\nabla)\nabla d\Omega \tag{5.26}$$

The decomposition of the divergence derives an oriented Laplacian operation shown in Equation 5.23. The differently weighted two oriented smoothings are applied to the perpendicular directions $\xi \perp \eta$.

$$\frac{\partial I}{\partial t} = \frac{g(\|\nabla I_\sigma\|)}{\|\nabla I_\sigma\|} I_{\xi\xi} + g'(\|\nabla I_\sigma\|) I_{\eta\eta} \tag{5.27}$$

This is equivalent to a form of Equation 5.23 with $c_\xi = \frac{g(\|\nabla I_\sigma\|)}{\|\nabla I_\sigma\|}$ and $c_\eta = g'(\|\nabla I_\sigma\|)$. c_ξ and c_η respectively diffuse with different weights along the isophote direction ξ, i.e. tangent direction, and the gradient direction η. This method is similar to applying spatially



(a) (b)

(c) (d)

Figure 5.6. Morphologically detected regions - (a) with 10 iterations, (b) with 7 iterations (c) with 3 iterations, (d) with a single calculation

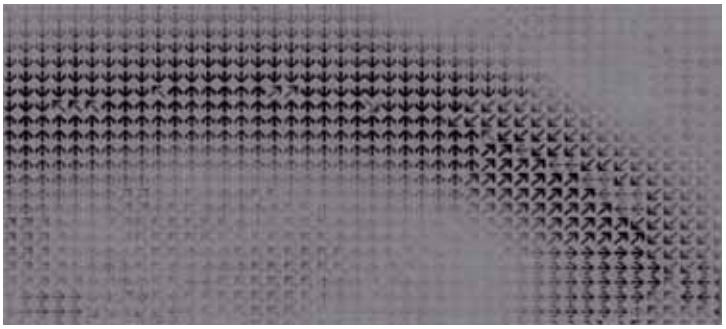


Figure 5.7. Adaptive variances in anisotropic diffusion

adaptive variances to the Gaussian smoothing. Figure 5.7 shows the adaptive variances of anisotropic diffusion. The arrow represents the diffusion direction, and the brightness is the size of variance. Directionally different weights of anisotropic diffusion prevents the important edges from over-smoothing [137, 138]. While pixels within a homogeneous region are smoothed by a Gaussian of large variance, pixels near an edge is smoothed by a small variance. This makes homogeneous image regions more homogeneous, discontinuous image edges more discontinuous.

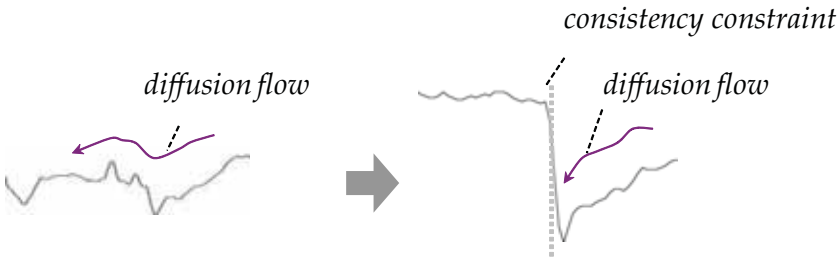


Figure 5.8. Perceptual maximum variation - (a) localization problem of diffusion on a weak edge, i.e. over-diffusion artifacts, (b) diffusion using perceptual maximum variation

Figure 5.9. Difference between the color channels - (a) colorful bird image and the subtleties of correlation between channels, (b) red channel, (c) green channel, (d) blue channel

5.5 Perceptual Maximum Variation Modeling for Over-diffusion Problem

Regularization of errors in color images is an important task for many applications. Here it is of vital importance to detect and reduce noise while preserving important image structures such as strong and weak edges and fine texture details. Diffusion is an efficient localized image regularization method based on the analysis of image structures such as direction and magnitude. However, there are two difficult problems due to the error sensitivity and channel mixture.

(1) Error sensitivity: The scalar diffusion methods employ a fluid mechanic basis which equilibrates spatial variations in concentration for regularization of images. However, regularization often results in *over-diffusion*, i.e. mislocalized flow and thus in blur of small brightness variations in images. When the important edges are formed by small brightness variations, over-diffusion causes blurring problems that remove the edges. Figure 5.8 illustrates the over-diffusion problem.

Table 5.3. Algorithm of the perceptual color modeling

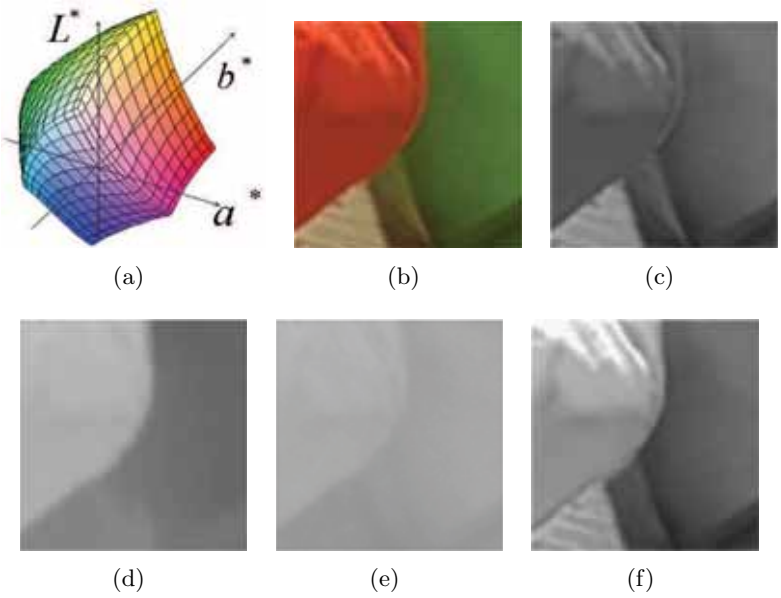
Perceptual Maximum Variation Modeling	
Step 1:	A color image is converted into a perceptually uniform color space of $CIE\text{-}L^*a^*b^*$.
Step 2:	PCA is used to find a set of orthogonal vectors lying along the direction of the maximal variation.
Step 3:	A consistency constraint is considered to enhance the localization of homogeneity between color regions by removing the diversity of global color distribution from the image.
Step 4:	The small brightness variations on a color edge are selectively modulated in luminance axis to preserve differences of chromaticity.
Step 5:	The modulated variations are projected into a perceptual dimension $\mathcal{P}$ which has high dynamic range.

(2) Channel mixture problem: Although recent diffusion methods handle the problem using tensor-valued, i.e. vector diffusivity function [139, 140], which can be adapted to local edge orientation, the methods are just suitable for a grayscale image or a single channel processing. An independent process of each color channel results color distortions due to subtleties of color correlation among the color channels [141, 142]. For example, an edge between head and body of a colorful bird shown in Figure 5.9 can be clearly distinguished in the red channel but diffusion error may occur in the other channels.

The proposed method [148] is based on the fact that the human visual system detects the boundary of objects by considering chromaticity and brightness simultaneously. If the brightness difference of neighboring objects is very small, our eyes can detect the important edges using the chromaticity difference. We estimate a higher dimensional perceptual $\mathcal{P}$-space, i.e. the size of this space is not always the same with the size of the common color space, by projecting the maximum variations of both brightness and chromaticity. The algorithm of the perceptual color modeling is shown in Table 5.3.

A RGB image is converted to $CIE\text{-}L^*a^*b^*$ color space [149] that is designed to approximate human vision and aspires to perceptual uniformity[42]. Figure 5.10(a) depicts the $CIE\text{-}L^*a^*b^*$ color space. Two regions with red and green colors in an image shown in Figure 5.10(b) can be easily distinguished by our eyes. However, the luminance values of regions are very similar as Figure 5.10(c) depicts, and the edge may be very ambiguous to be detected. The proposed

42) Perceptually uniform means that a change of the same amount in a color value should produce a change of about the same visual importance. The three coordinates of $CIE\text{-}L^*a^*b^*$ represent the lightness of the color, i.e. $L^* = 0$ yields black and $L^* = 100$ indicates white, a chromaticity axis a^* between red/magenta and green and the other chromaticity axis b^* between yellow and blue. The asterisks (*) after L, a and b are part of the full name, since they represent L^*, a^* and b^*, to distinguish them from Hunter's L, a and b.



(a) (b) (c)

(d) (e) (f)

Figure 5.10. Maximum variations in $CIE\text{-}L^*a^*b^*$ color space - (a) $CIE\text{-}L^*a^*b^*$ color space (b) original color image, (c) L^*,(d) a^*, (e) b^* of the color image, (f) $\mathcal{P}$-space resulted by the proposed modeling, i.e. note that the scale of $\mathcal{P}$-space is 255 for visualization purpose.

method solves this problem by color modeling, and the small luminance variations on a color edge is selectively modulated in a luminance axis to preserve differences of chromaticity as Figure 5.10f shows. Figures 5.10(d) and 5.10(e) respectively show the a^*- and b^*-space images.

Perceptual differences between two neighboring color pixels in $CIE\text{-}L^*a^*b^*$ can be simply measured by the Euclidean distance between the two vectorial values $c_p(\mathrm{x}) = [L_p^*, a_p^*, b_p^*]$ and $c_q(\mathrm{x}) = [L_q^*, a_q^*, b_q^*]$ with perceptual color metric δc_{pq}^* in Equation 5.28. x denotes a pel $\mathrm{x} = (x, y)$ on a $CIE - L^*a^*b^*$ color domain $I(\mathrm{x}) : \mathbb{R}_+^3 = \{c(L_p^*, a_p^*, b_p^*) \geq 0\}$.

$$\delta c_{pq}^* = \sqrt{\left(L_p^* - L_q^*\right)^2 + \left(a_p^* - a_q^*\right)^2 + \left(b_p^* - b_q^*\right)^2} \tag{5.28}$$

Colors with the same δc_{pq}^* are perceptually equal. The brightness $\omega(\mathrm{x})$ in a luminance domain L^* and the chromaticity $\upsilon(\mathrm{x})$ includes a^* and b^* respectively. x denotes an image pixel $[u, v]$. It represents the length of the color vectors and the normalized color components.

$$\omega(\mathrm{x}) = \sum_{i=1}^{n} \|I_n(\mathrm{x})\| \quad \text{and} \quad \upsilon(\mathrm{x}) = I(\mathrm{x})/\omega(\mathrm{x}) \tag{5.29}$$

We estimate visually maximum variations using the perceptual difference which combines $\omega(\mathrm{x})$ and $\upsilon(\mathrm{x})$ in color metric δc_{pq}^*. A linear transform which projects original brightness in

the luminance L_p^* and L_q^* into perceptual intensity $\mathcal{P}_p$ and $\mathcal{P}_q$ (which remains proportional to the perceptual color difference) is obtained by minimizing the following quadratic function. The constant K is chosen for the proportional weight e.g. when $K = 1$, L_p^* and L_q^* are equally modulated with the perceptual differences.

$$\mathcal{P}\left(c_{0,\cdots,p,q,\cdots,n}(\mathrm{x})\right) = \sum_{p=0}^{n-2} \sum_{q=p+1}^{n-1} \left(|\mathcal{P}_p - \mathcal{P}_q| / \delta c_{pq}^* - K\right)^2 \tag{5.30}$$

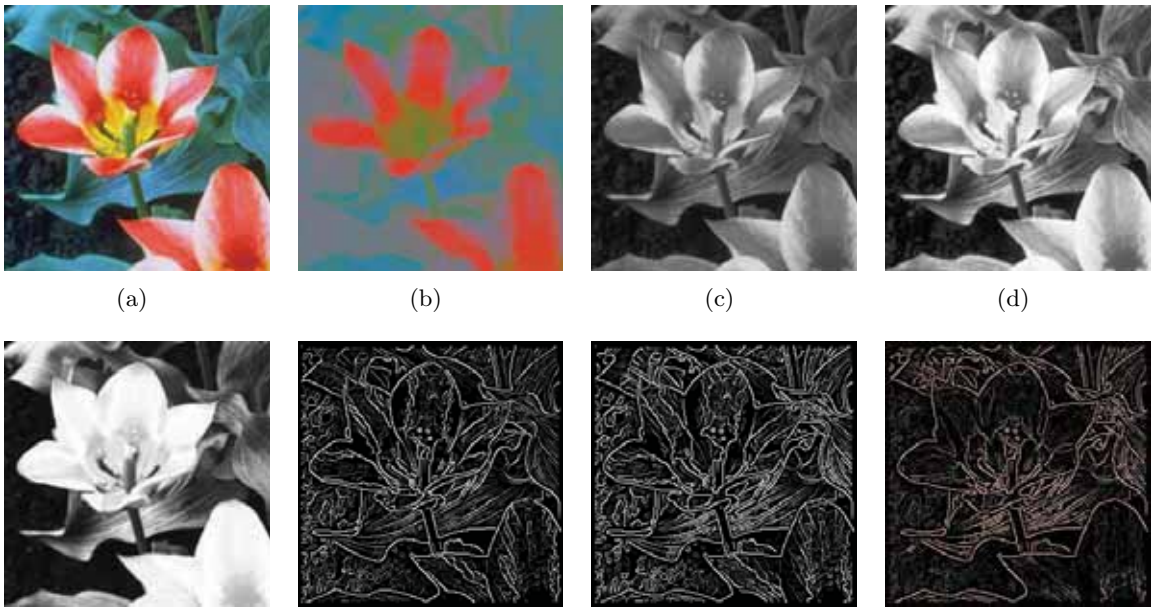
$\mathcal{P}_p$ and $\mathcal{P}_q$ generally have higher dynamic ranges than L_p^* and L_q^*, since they include a total difference of both brightness and chromaticity. $\mathcal{P}$-space has maximum variations in human visual range. Using PCA of Equation 5.30, the entire distribution of color values can be projected onto the primary L^*-axis of the ellipsoid using a linear transform. The principal components are the eigenvectors of its covariance matrices. By calculating the covariance matrix with the largest eigenvalues the perceptual maximum variation can be estimated. PCA performs very well in aligning the colors in a region which have locally-compact or globally-smooth color distributions. If we deal with whole images in a noisy condition, the color whole image distribution may not be suitable to fit local properties. Equation 5.30 is convex, but may converge to multiple global minima. We solve the problem by considering a consistency constraint. First, the chromaticity difference between two color pixels is defined by an equation similar to Equation 5.28 as

$$\delta \upsilon_{pq} = \sqrt{(a_p^* - a_q^*)^2 + (b_p^* - b_q^*)^2} \tag{5.31}$$

As shown in Figures 5.10(d) and 5.10(e), the chromaticity domain can be used as a good consistency measure because it inherently has piecewise smoothness over the image - while the brightness in L^* is affected by saturation, lightness, e.g. shadow, reflection and noise, etc. The consistency measure is decreased when the spatial distance to the pixel under consideration increases due to the coherency of objects. A consistency Λ is defined as a similarity group based on probability, using the chromaticity difference υ_{pq} and spatial distance d_{pq} with proportional constant k. γ_υ and γ_d are empirically determined.

$$\Lambda\left(c_{0,\cdots,p,q,\cdots,n}(\mathrm{x})\right) = \sum_{i=1}^{n} k \cdot exp\left(-(\delta \upsilon_{pq}/\gamma_\upsilon + d_{pq}/\gamma_d)\right) \tag{5.32}$$

Figure 5.11 exemplifies the edges in an image using perceptual maximum variation modeling. Figure 5.11(a) shows the original image that has two visually different color regions, i.e. foreground and background. The human visual system can easily distinguish the foreground flowers with red color from the background with dark green color due to the chromaticity differences. The chromaticity domain image, i.e. a^*b^*, is shown in Figure 5.11(b). However, the luminance image in Figure 5.11(c) has a small contrast between foregrounds and back-

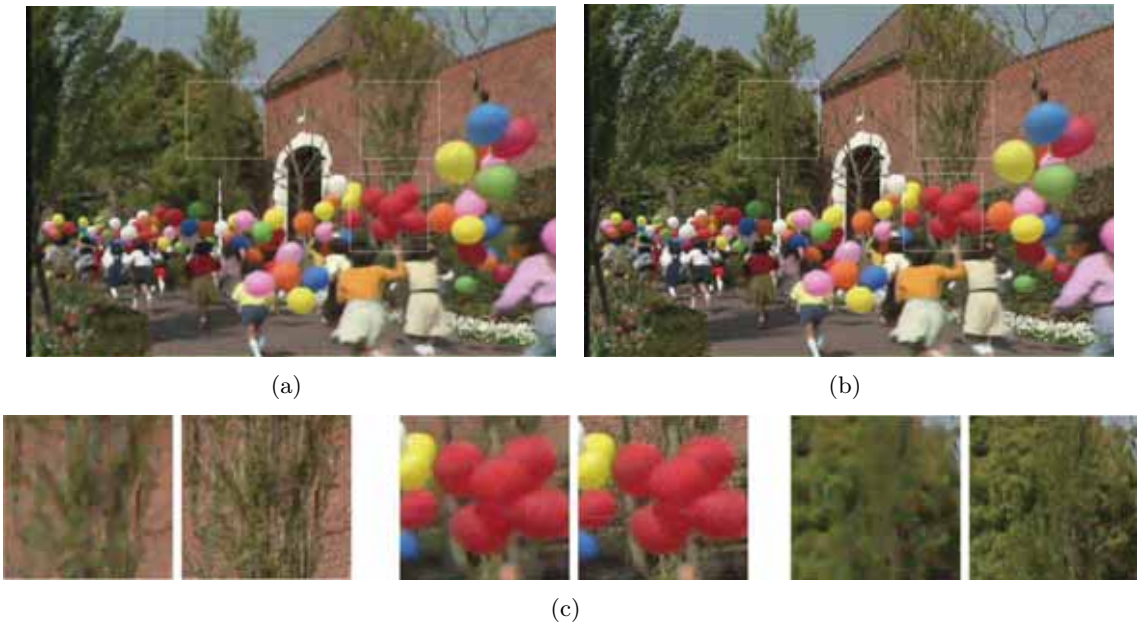


(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 5.11. Edge detection performance of perceptual maximum variation modeling - (a) original color image, (b) chromaticity image (a^*b^*), (c) luminance image (L^*) (d) histogram equalization of luminance image, (e) maximum variation modeling result, (f) edge image of Figure b, (g) edge image of Figure d, (h) edge image of Figure e

grounds that are both dark. Although the local contrasts can be increased by a well-known histogram equalization method that distribute them on the histogram, the equalization result shown in Figure 5.11(d) is not sufficient to represent perceptual differences in the image. Since histogram equalization enhances a lower local contrast to gain a higher contrast without affecting the global contrast, the edge contrast between color regions cannot be increased. The proposed method solves the problem by a variation modulation method projecting all color variation into a perceptual dimension with a high dynamic range. Figure 5.11(e) depicts a maximum variation modeling result. PCA is used to find a set of orthogonal vectors lying along the direction of the maximal variation. Figures 5.11(f) and 5.11(g) respectively represent edges of the original image and edges of the histogram equalized image. Histogram equalization does not enhance real edges between foreground and background regions but small edges in the regions. The edge image with the perceptual maxima variations depicts real detection of visual image differences in Figure 5.11(h). The size of $\mathcal{P}$-space is 255 for visualization purpose although the real size is much larger than 255.



(a) (b)

(c)

Figure 5.12. Anisotropic diffusion with perceptual color modeling - (a) Perona and Malik color diffusion for 10% Gaussian noise image, PSNR=26.199dB (b) Proposed method for 10% Gaussian noise image, PSNR=33.518dB (c) details with over-diffusion on weak edges i.e. left: case Figure a, right: Figure b

5.6 Robust Anisotropic Color Image Regularization with Perceptual Maximum Variation Modeling

The localization at weak features which have small brightness variations is fundamental problem of diffusion regularization, and this often results in removal of weak features. We address this problem with perceptual maximum variation modeling described in the previous chapter. In Figure 5.12(a), we exemplify the over-diffusion problem of weak features using Perona and Malik color diffusion [143]. The proposed method using perceptual maximum variation overcomes the problem as Figure 5.12(b) shows. Figure 5.12(c) represents some detailed images on weak edges.

If a noisy color image is considered as a noisy color vector space $I(\mathrm{x}) : \mathbb{R}^2 \rightarrow \mathbb{R}^n$, the space can be separated into brightness $\omega(\mathrm{x}) : \mathbb{R}^2 \rightarrow \mathbb{R}^+$ and chromaticity $\upsilon(\mathrm{x}) : \mathbb{R}^2 \rightarrow S^{n-1}$. $\mathrm{x} = [\mathrm{u}, \mathrm{v}]$ denotes image pixels. We deal with the brightness and chromaticity components separately for diffusion. First, the diffusion of brightness is calculated by tensor-valued diffusivity function

$$\partial_t \omega(\mathrm{x}) - div(\mathbf{D} \cdot \nabla \omega(\mathrm{x})) = \mathbf{0} \tag{5.33}$$

$\mathbf{D}$ denotes a positive definite symmetric matrix called *diffusion tensor* and ∂_t is the derivation with respect to the time. Instead of evaluating the gradient of initial brightness, we propose

to obtain a perceptual diffusion flow tensor using the Cartesian product of the gradient vector $(Q_x, Q_y)^T$ in the $\wp$-space resulting from Equation 5.30 with itself. While the small variations $\nabla\omega(\mathrm{x})$ of brightness are well filtered, important structures in the variations are still preserved by localized diffusion flow in the perceptual maximum.

$$\mathbf{D}\left(\nabla_{\wp}\right) = \begin{pmatrix} Q_x^2 & Q_x Q_y \\ Q_x Q_y & Q_y^2 \end{pmatrix} \tag{5.34}$$

$Q_x = \mathcal{P}(\mathrm{x}) * \mathcal{G}_{x,\sigma}$ and $Q_y = \mathcal{P}(\mathrm{x}) * \mathcal{G}_{y,\sigma}$ are obtained by x- and y-directional derivatives of 2-D Gaussian kernel $\mathcal{G}(\mathrm{x}) = (2\pi\sigma^2)^{-1} \exp\left(-(x^2+y^2)/2\sigma^2\right)$ with a standard deviation σ. $\mathcal{P}$ is a space with perceptual maximum variations which combines the small differences in both brightness and chromaticity using a least squares optimization with PCA[43]. A consistency constraint is employed to avoid influence from global color distributions and to enhance homogeneous color regions. With a scale σ of successively smoothed concentrations, its eigenvectors describe the direction of highest and lowest contrast. These contrasts are given by corresponding two positive eigenvalues. Anisotropic diffusion of brightness is performed by weighting of the eigenvalues using a redescending weight function [44] that higher weights are given to the measurements with small values.
Let the chromaticity $\upsilon_i(\mathrm{x}):\mathbb{R}^2 \rightarrow \mathbb{R}$ describe each of the n-components of $\upsilon(\mathrm{x})$. The gradients of the components $\nabla\upsilon_{\mathrm{i}} = (\partial\upsilon_i/\partial x)\vec{x} + (\partial\upsilon_i/\partial y)\vec{y}$ can be defined with unit vectors $\vec{x}$ and $\vec{y}$ which have the values of the component gradients

$$\|\nabla\upsilon_{\mathrm{i}}\| = \left((\partial\upsilon_i/\partial x)^2 + (\partial\upsilon_i/\partial y)^2\right)^{1/2} \tag{5.35}$$

in the x- and y-directions. We solve a constrained minimization problem called *harmonic map* [141] with a constraint $\|\nabla\upsilon\| = 1$ as

$$\partial_t\upsilon_i - div(\|\nabla\upsilon\|^{p-2} \cdot \nabla\upsilon_i) + \upsilon_i \|\nabla\upsilon\|^p = 0, \quad 1 \leq i \leq n \tag{5.36}$$

Here, $\|\nabla\upsilon\| = \sum_{i=1}^{n} \left((\partial\upsilon_i/\partial x + \partial\upsilon_i/\partial y\right)^{1/2}$ is the absolute value of the image gradient, i.e. total component gradient. When $p = 2$, this equation equals to the isotropic diffusion $\partial_t\upsilon - \Delta\upsilon_i + \upsilon_i \|\nabla\upsilon\|^p = 0$, which substitutes the divergence term into a component Laplacian $\Delta\upsilon_i = \left(\partial^2\upsilon_i/\partial x^2 + \partial^2\upsilon_i/\partial y^2\right)^{1/2}$. Anisotropic diffusion in chromaticity results in the range of $1 \leq i \leq 2$ given in the weighting function.

43) see Chapter 5.5
44) see Table 5.1 and 5.2.

5.7 Experimental Results

Robust fundamental matrix estimation : Feature correspondences are used to estimate the global ray projection from the focal point to the camera center, and to establish the collineation in the viewpoints. If the image points are corrupted by identically distributed Gaussian noise, the accuracy of the parameter estimation may be decreased. A robust estimation criterion is derived from the noise distribution of the image points, and we can estimate the transformed fundamental matrix by minimizing the following weighted sum of squares

$$\int_{i=1,\cdots,n} \rho_\sigma \left(f_i/\sigma_i\right)^2 \tag{5.37}$$

where ρ is a bounded loss function given in Equation 5.3 and $\sigma_{f_i}^2 = (\partial f_i/\partial \tilde{\mathbf{m}}_i)^T$ is the variance of f_i. We use Tukey's biweight function shown in Equation 5.8 that is calculated in the ρ. The relationship between the transformed fundamental matrix and image points are

$$f_i : \tilde{\mathbf{m}}_i'^T \mathbf{P}'^T \mathbf{F} \mathbf{P} \tilde{\mathbf{m}}_i = 0 \tag{5.38}$$

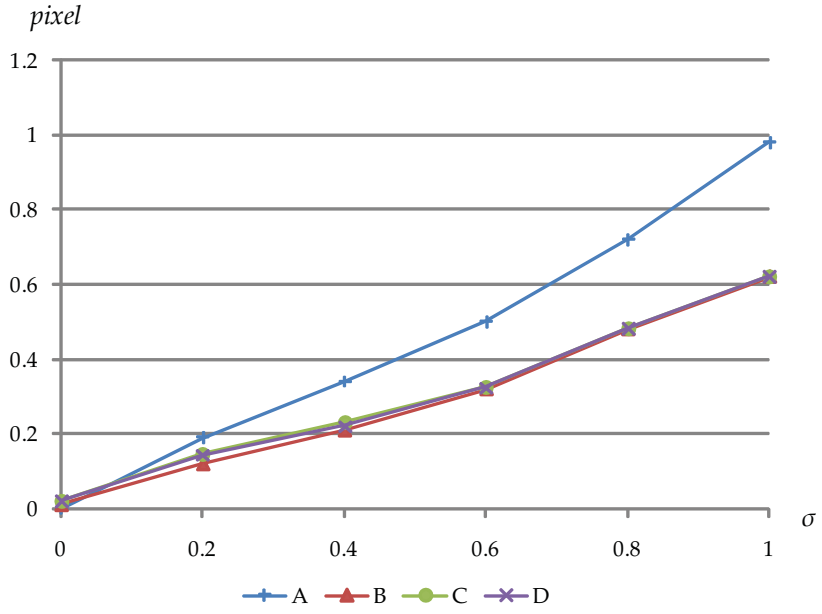
The estimated $\mathbf{F}$ matrix should be a rank-2 matrix in order to model the epipolar geometry with all the epipolar lines intersecting in a unique epipole. The $\mathbf{F}$ is decomposed by a singular value decomposition (SVD) as Equation 4.15 depicts.

$$\mathbf{F}{=}\mathbf{U}\mathbf{D}\mathbf{V}^T \tag{5.39}$$

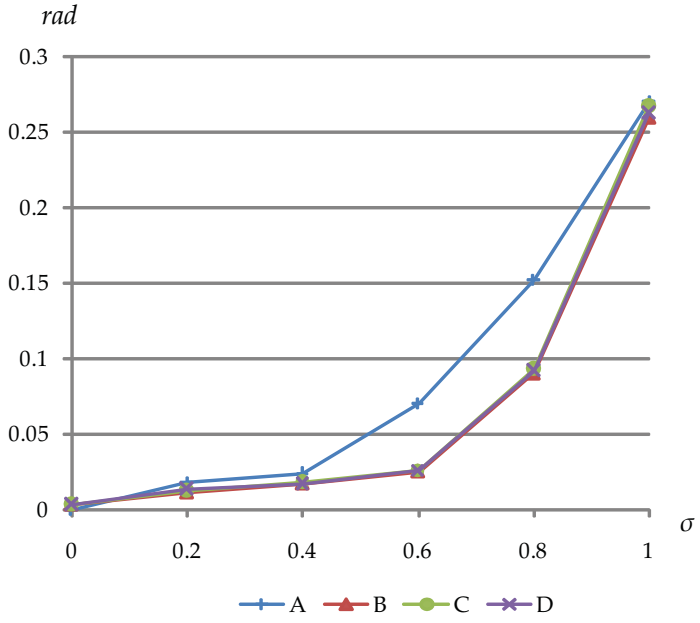
where $\mathbf{D} = diag\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \sqrt{\lambda_3}\right)$. The component with the smallest weight is removed to obtain $\hat{\mathbf{D}} = diag\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, 0\right)$, and then $\mathbf{F}$ is recalculated as $\mathbf{F} = \mathbf{U}\hat{\mathbf{D}}\mathbf{V}^{\mathbf{T}}$.
We analyze the first and second frames of the corridor sequence provided in [103] with increasing Gaussian noise level for evaluating robustness, and the following sets of parameters are evaluated:

A: Initial estimate.
B: All 12 entries of $\mathbf{P}'$.
C: 5 parameters of $[\mathbf{t}]_\times \mathbf{R}$.
D: 5 parameters of $\mathbf{P}'\mathbf{F}\mathbf{P}$.

Ground-truths for 2-D and 3-D geometry are available for this sequence. It is important to test the reprojection errors and angular errors of epipoles since the locations of the epipoles are often not accurate in most studies. Figures 5.13(a) and 5.13(b) respectively represent the reprojection errors and angular errors of epipoles for the standard deviations of noise σ. Our method has less than 1 pixel reprojection error and maximum 0.27 rad. angular errors even for a large noise level. Two frame results with noise in Figure 5.14 shows all epipolar lines



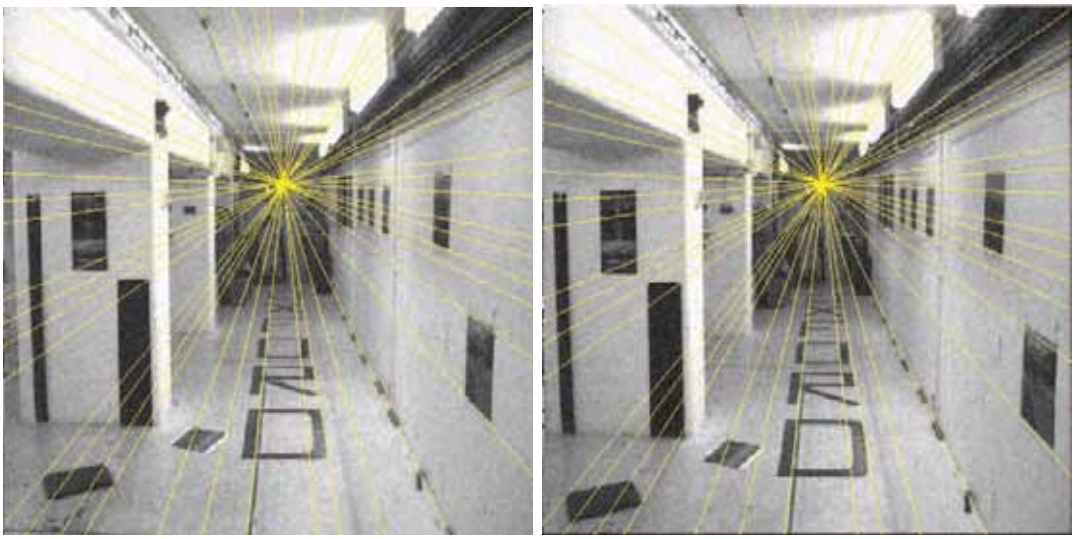
(a)



(b)

Figure 5.13. Evaluation of robust fundamental matrix estimation - (a) reprojection errors of epipoles, (b) angular errors.

including detected features and the epipole. Average distance errors of several robust methods for fundamental matrix estimation under different noise levels are shown in Figure 5.15.



(a) (b)

Figure 5.14. Epipolar geometry estimation for two frames of the corridor sequence with Gaussian noise - (a) the 1st frame (b) the 2nd frame

The least-squares (LS) methods does not have accurate result when the eigenvalues may get a rank-3 matrix. However, robust methods with eigen analysis removes the outliers more efficiently so that the accurate F-matrix can be obtained. RANSAC [121], MLESAC [111] and MAPSAC [125] show similar results.

Super-sampled ray-space : The robustness of epipolar geometry against noise can be applied to a super-sampling using ray-space. Super-sampling is a technique that in some way decrease the aliasing of an imaging system. Most camera sensors in a dark environment need high sensitivity, and some aliasing actually happens. An average value of large amount of pixel samples taken at several instances can eliminate the aliasing. In the view of ray-space, the pixel samples form a unit ray from a 3-D surface to the camera center.
We add 7% Gaussian noise to the original image frames. Figures 5.16(a) and 5.16(b) respectively show the original image and the error image due to the Gaussian noise. Epipolar geometry characterizes the uncalibrated stereo using a set of fundamental matrix since there are different views of geometry caused by the monocular camera motion. However, the noise problem in epipolar geometry estimation is heteroscedastic[45] because each measurement is affected by noise of different parameters. Hartley [161] solved this problem by separately normalizing points in two image planes. However, the normalization process does not consider the nature of heteroscedastic noise. During triangulation, each 3-D point is localized with a

45) The data vectors are supplied by different sources, e.g. different cameras, and are inherently heteroscedastic.

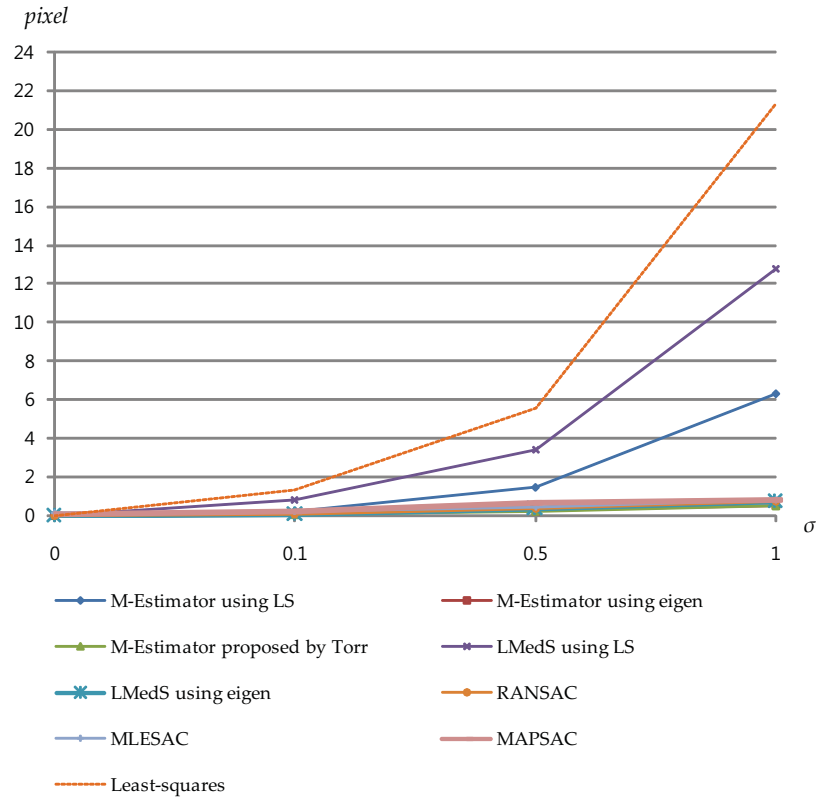
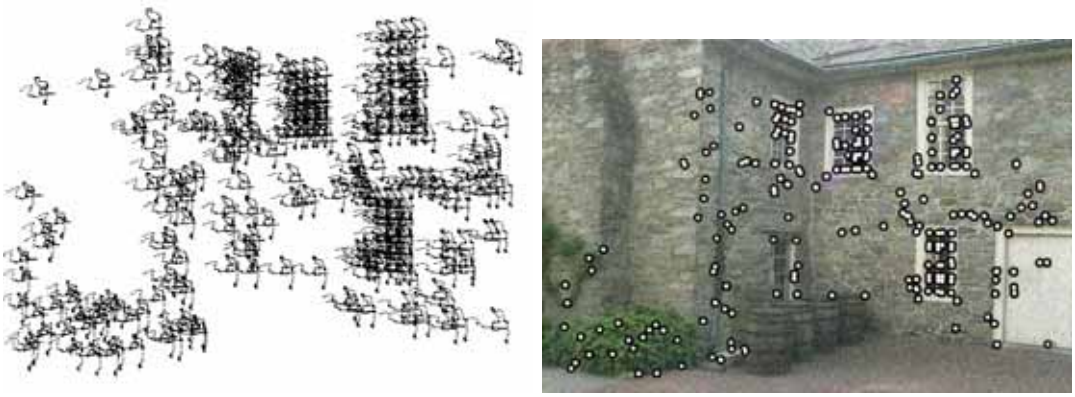


Figure 5.15. Comparison of several robust fundamental matrix estimation methods

different covariance. The measurements and parameters are closely related by a constraint equation, and a noise model characterizes the noise affecting the measurement. While the noise is affecting the true values, the robust estimator provides a noise model measuring the error scale with covariance matrix, and discriminate inliers from outliers. While the features are detected and tracked as Figure 5.16(c) the most stable features shown in Figure 5.16(d) are remained in many noisy feature tracks and used to estimate the 3-D point cloud and relative camera positions. Figure 5.17 visualizes the 3-D point cloud and camera positions using X-Y X-Z and Y-Z axes. The optical ray-space can be reconstructed, and noisy pixels of viewpoints are super-sampled as Figure 5.18(b) shows. A number of parts in the noisy images and super-sampled images are compared in Figure 5.18, and the proposed method can prominently decrease the Gaussian noises in a simple hand-held video.

Structure preserving noise regularization : Error regularization is an important task for many computer vision applications. Here it is of vital importance to reduce noise while preserving important image structures such as strong and weak edges and fine texture details. Diffusion is an efficient localized image regularization method based on the analysis of image structures such as direction and magnitude. However, the localization at weak features which

(a) (b)



(c) (d)

Figure 5.16. Feature tracking in noisy image sequences - (a) original first frame image, (b) difference between original image and noisy image, (c) detected features, (d) trajectory of tracked features

Table 5.4. Comparison using Lena image with 4, 10% color impulsive noise

Algorithms	4%	10%
None	17.983	11.702
Arithmetic mean filter 3x3(AMF)	25.971	14.802
Perona and Malik color (PM) [143]	28.146	17.508
Vector directional filter (VDF) [144]	30.466	17.716
Vector median filter (VMF) [146]	31.427	22.342
Fuzzy vector directional filter (FVDF) [145]	30.827	20.089
Modified vector median filter (MVMF) [147]	38.446	26.411
Proposed method	39.841	27.834

have small brightness variations is fundamentally difficult. The over-diffusion, i.e. a value mixture problem, often results in the removal of weak features. We solved this problem with perceptual maximum variation modeling in Chapters 5.5 and 5.6.

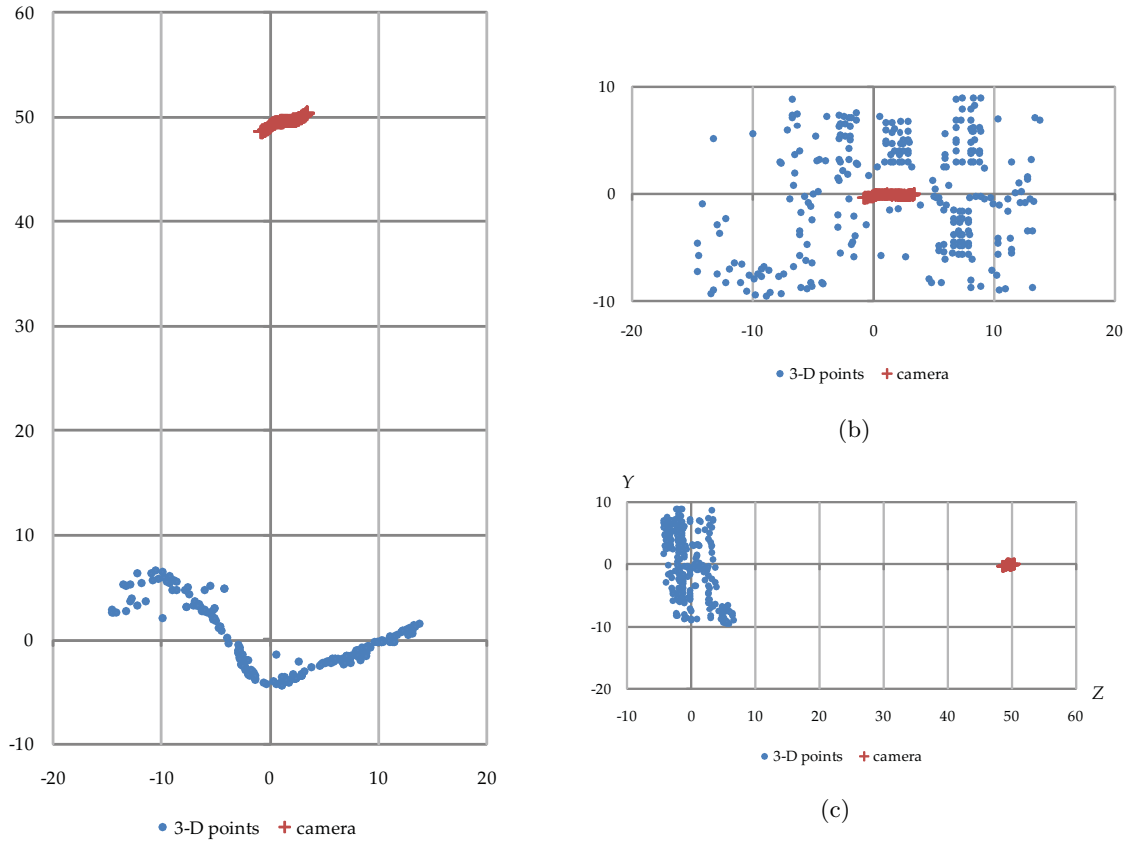


Figure 5.17. Relative 3-D position of cameras and tracked points - (a) plot with X-Z axes, (b) plot with X-Y axes, (c) plot with Y-Z axes

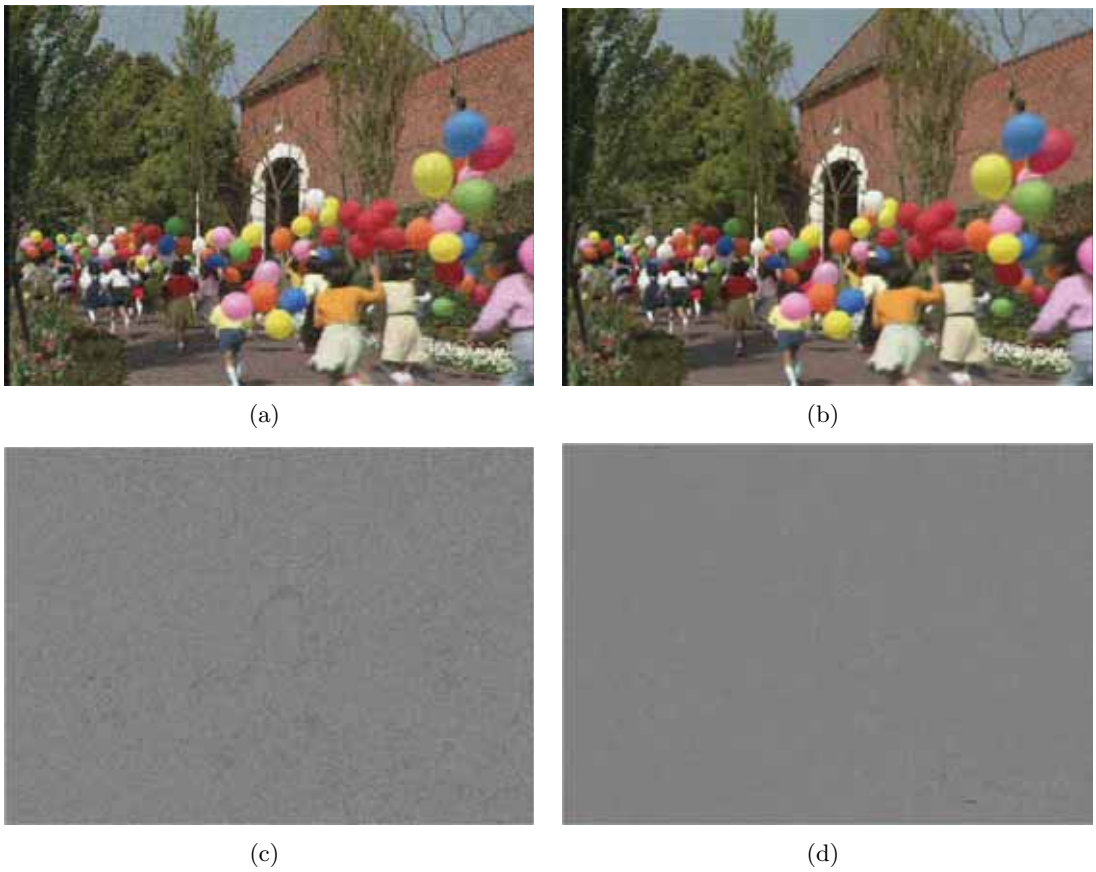
Some studies of color filtering deal with the value mixture problem using color vectorial methods. An efficient method is to extend the medians of the scalar space to the color vectorial data. Vector median filters (VMF) are obtained by considering L_1, L_2 norms for ordering image vectors according to their relative magnitude differences [146, 147]. Other methods consider vector directional filters (VDF) [144, 145]. These works are related with a heuristic approach which makes homogeneity directed correlation among the color channels. However, the localization performance on ambiguous edges with weak variations is still unstable. The human visual system perceives small brightness variations using a knowledge-based analogy from color components. Perceptually motivated color spaces are used to evaluate mutual coherency and geometrical continuation. Although these methods achieve good results for preserving colors, they do not preserve small brightness variations in the color images.

We investigate the efficiency of the proposed method on zero-mean Gaussian color noise and

Figure 5.18. Super-sampling in ray-space - (a) one frame of noisy video sequence, (b) super-sampling result of optical-ray, (c) left-bottom part of the noisy image, (d) left-bottom part of the super-sampled result, (e) right-center part of the noisy image, (f) right-center part of the super-sampled result, (g) center-bottom part of the noisy image, (h) center-bottom part of the super-sampled result, (i) center part of noisy image, (j) center part of the super-sampled result

zero-mean impulsive color noise[46]. Figure 5.19 shows results of our method using the Balloon image with 10% Gaussian color noise for each color channel in original color image. In Figure 5.20, the noise regularization performance using the Parrot image with 10% and 15% Gaussian color noises is visually evaluated. Perceptual modeling estimates maximum variations in color consistency to enhance convergence. Then, the diffusion process is less sensitive

46) Impulse color noise had random amplitude and spectral content with a large perturbation of the color values.



(a) (b)

(c) (d)

Figure 5.19. Regularization of an image with 10% Gaussian color noise - (a) Balloons image with 10% Gaussian noise (b) denoising result of 5a using the proposed method, PSNR=33.518dB (c) error image of Figure a (d) error image of Figure b

to momentary variations, e.g. noise but more sensitive to discontinuous variations such as edges between two regions. In the geometric continuation, robust denoising is achieved even in image regions containing weak edges and small brightness variations by iterative diffusion with a scale. PSNR[dB] is evaluated for color in three channels as

$$10 \log \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} (255 \cdot 3N_1N_2 / \left\| I(i,j) - I'(i,j) \right\|^2) \tag{5.40}$$

Table 5.4 depicts the efficiency of the proposed method compared with heuristic correlated color filtering approaches i.e. AMF, VMF, FVMF, VDF and MVDF for 4% and 10% percentages of color impulsive random noise. The comparisons verify the performance of spatio adaptation. The proposed method with perceptual maximum variation modeling achieved significant improvements over the other techniques, ranging from 1.5 dB to 13.9 dB in PSNR. Figure 5.21 is provided for visual comparison of results in the table. Figure 5.21(k) shows

(a) (b)

(c) (d)

Figure 5.20. Regularization of images with 10% and 15% Gaussian color noises - (a) Parrot image with 10% Gaussian noise (b) denoising from Figure a, PSNR=38.146dB., (c) Parrot image with 15% Gaussian noise, (b) denoising from Figure c,

that our method results in sharper and more detailed structures after denoising.

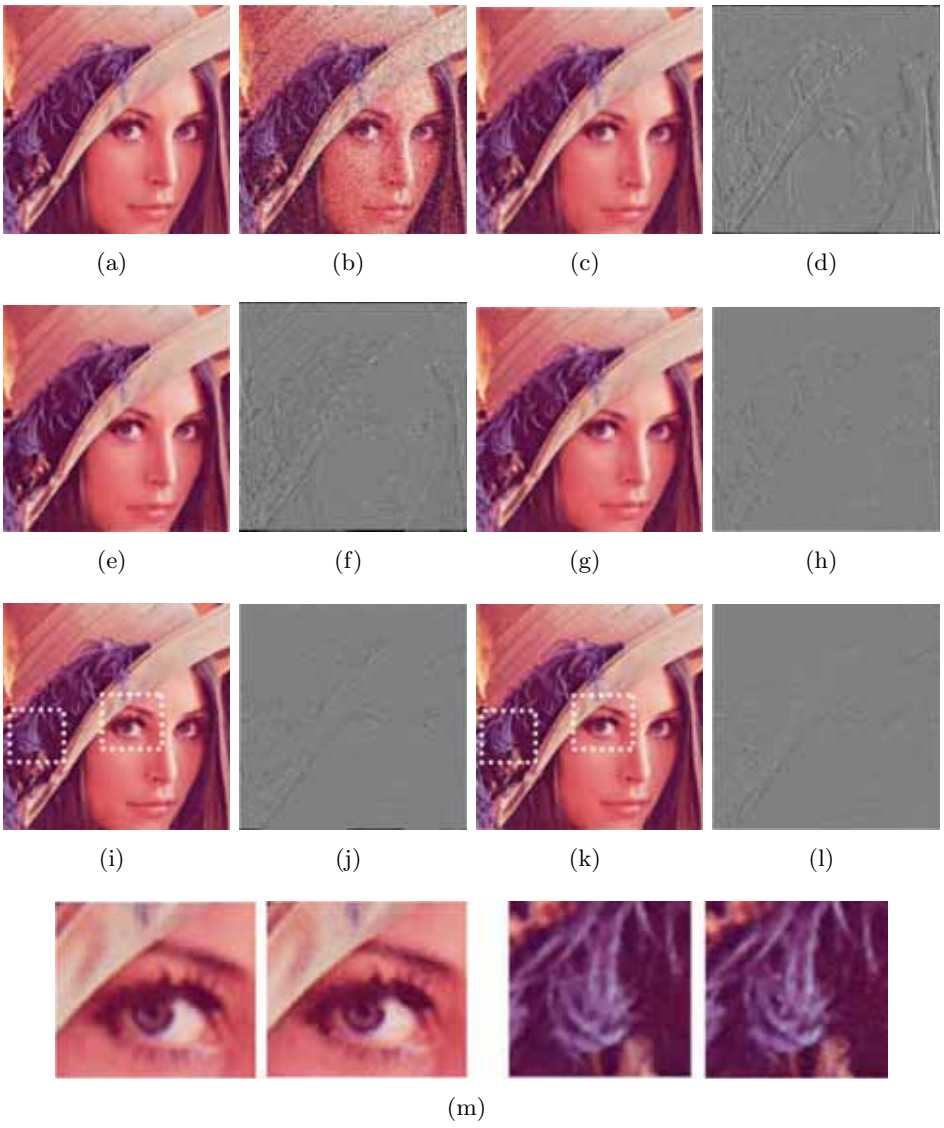


Figure 5.21. Comparison of denoising effects - (a) original image, (b) distorted image with 4% impulse noise (c) denosing using AMF method (d) error image of Figure c, (e) denosing using PM method, (f) error image of Figure e, (g) denosing using VMF method, (h) error image of Figure g, (i) denosing using MVMF method, (j) error image of Figure i, (k) denosing using the proposed method, (l) error image of Figure k, (m) Differences between Figure i and k

Chapter 6

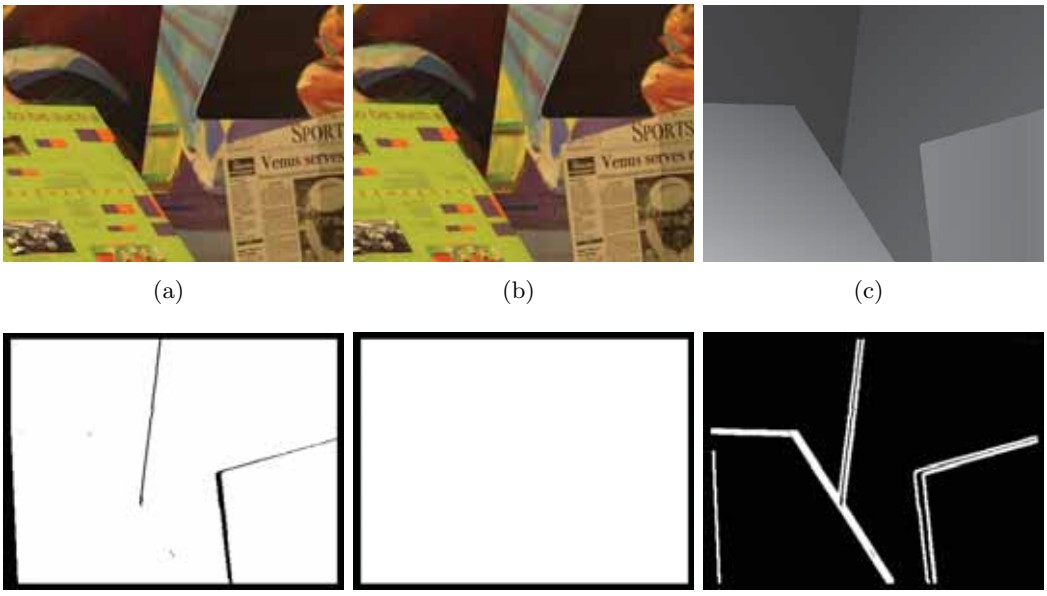
Robust 3-D Sceneflow Estimation

Video-based modeling and rendering methods needs accurate sampling position in the 3-D space since a linear approximation of the optical axis is not enough to fully cover the 3-D scene geometry. During the image formation process of the camera, explicit 3-D information about the scene is lost. Thus, 3-D sampling needs to infer the scene geometry from the 2-D images, and disparity and motion parallax are the cues to recover the scene geometry. The 3-D sceneflow method is a multi-stereo approach which fuses both stereo and visual motion cues for understanding structure. Motion tracking is relatively easier to establish correspondences than stereo matching, but stereo provides better structure estimates. In this chapter, robust estimations for disparity and optical flow is introduced separately, and a spatio-temporal consistency constraint for correspondences is added to combine the two estimation results. Finally we can obtain the reliable spatio-temporal correspondences.

6.1 Robust Anisotropic Disparity Estimation

A dense disparity map estimated in the epipolar lines between two viewpoints[47] is an important basis for the image-based modeling and rendering. For the two images taken simultaneously with a pair of cameras, the goal of dense disparity estimation is to match each pixel in one image to its corresponding pixel in the other image. A stereo image pair is shown in Figure 6.1(a) and 6.1(b). The ground-truth disparity map which is generated by a structured-light method [35] is represented in Figure 6.1(c). It is very difficult to obtain a similar disparity map to the ground-truth using an image matching because the problem of dense disparity estimation is the ambiguity[48] of local image structure due to image noise, unbalanced brightness, similar texture and occlusion, etc. In Figure 6.1(d), the white color region are non-occluded regions, and the black color regions are occluded and border regions. Regions marked by a white color in Figure 6.1(e) are all regions including half-occluded, and

47) When the images are rectified, the epipolar lines are image scanline.
48) If pixels in an image look alike, it is difficult to find corresponding pixels from another viewpoint image.



(a) (b) (c)

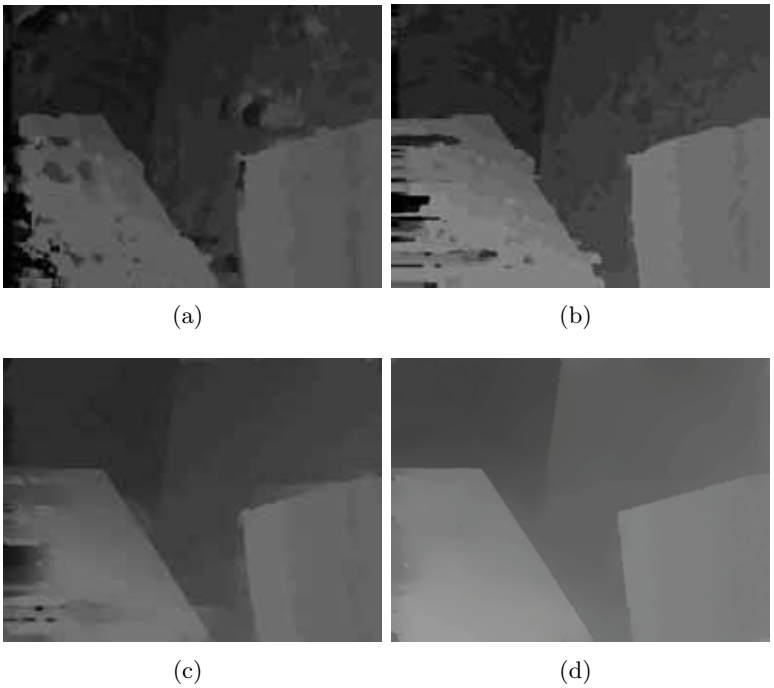
(d) (e) (f)

Figure 6.1. Dense disparity estimation problem - (a) the left image, (b) the right image, (c) ground-truth map generated by a structured-light method, (d) occlusion region, (e) boundary region, (f) depth discontinuity

Figure 6.1(f) shows regions near depth discontinuities using white color.
Recent methods employ local appearance matching with a landmark of the coherency of objects or boundary constraints between features, edges and disparity discontinuity etc. to obtain more reliable performance [151, 152]. Regions divided by edge are used for matching, and a fixed size window causes some poor localization of matching as Figure 6.2a shows. Adaptive window matching methods [72–74], that adjusts the window size using gradient, solves the edge localization problem. However, there are some regions without matches as Figure 6.2b exemplifies. Energy-based methods incorporate a regularization function that attempts to filter off mislocalization errors in the iteratively error minimization.

(a) Isotropic regularization [153] : Convolution carries out filtering based on variances. However, the scale of linear transformations in the convolution leads to undesired smoothing of important discontinuities.

(b) Anisotropic regularization [154, 155] : This method prevents important structure from blurring. The basic idea is to modify the filter scale at the discontinuities with steep intensity gradients. When this method is applied to disparity estimation, homogeneity enhancement and discontinuity preservation can be obtained in some images. However, ill-posed local minima during solving the partial differential equations (PDEs) are the serious drawback. Figure 6.2c shows the result of an anisotropic disparity estimation



(a) (b)

(c) (d)

Figure 6.2. Examples of several dense disparity matching - (a) region-divided matching method, (b) adaptive window matching method, (c) adaptive window matching with a pure anisotropic diffusion regularization, (d) adaptive window matching with a combination of isotropic and anisotropic regularizations

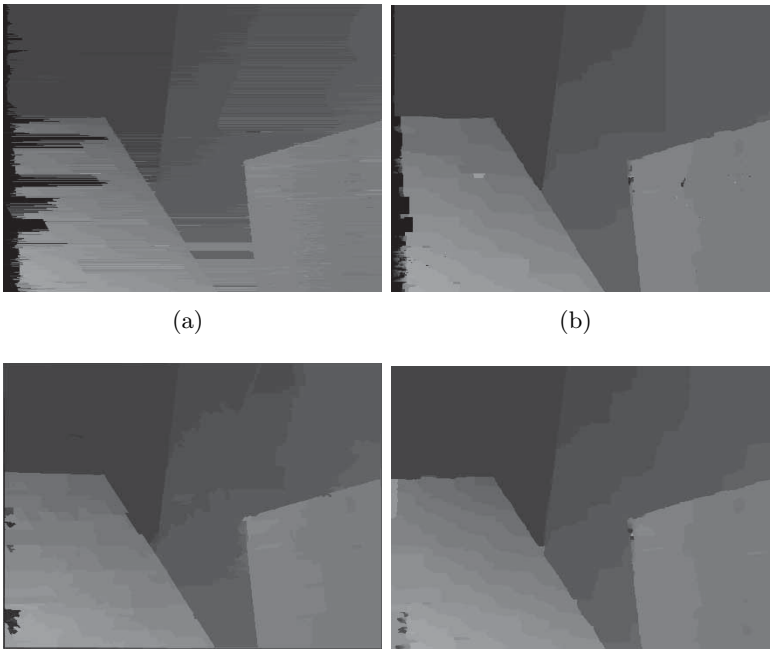
method without consideration of local minima. Since strong edges are not always an object boundary, strong texture edges causes regularization artifacts, e.g. backward diffusion and over-splitting.

(c) A combination of isotropic and anisotropic regularization [159, 160]. This method can solve the local minima problem in a pure anisotropic regularization since texture edges can be isotropic ally filtered while real object edges still remain. The proposed method uses isotropic pre-filtered scale-space and anisotropic diffusion derived in the smoothed space. Figure 6.2(d) shows a result of the proposed method, and several disparity maps generated by different methods are shown in Figure 6.3.

Local appearance matching in scale-space regions : A good measure of homogeneity is needed to restrict matching within each coherent object. A matching region can be established by analyzing the smoothly varying orientation structure[49] in the Gaussian scale-space as

$$\eta = \theta_+ = \frac{\nabla \mathrm{I}_\sigma}{\|\nabla \mathrm{I}_\sigma\|} \quad \text{and} \quad \xi = \theta_- = \frac{\nabla \mathrm{I}_\sigma^\perp}{\|\nabla \mathrm{I}_\sigma\|} \tag{6.1}$$

49) see Equation 5.22.



(a) (b)

(c) (d)

Figure 6.3. Comparison of dense disparity maps generated by several different methods - (a) dynamic programming method, (b) graph-cut method [156], (c) layered method [157], (d) hierarchical belief propagation [158]

where $I_\sigma = \left(2\pi\sigma^2\right)^{-1} e^{-\mathrm{x}2/2\sigma 2} * I\left(\mathrm{x}\right)$ is the pre-filtered scale-space by a Gaussian kernel. The unit vectors η and ξ respectively defines the gradient direction of filtered image and its orthogonal direction (i.e. isophote direction). Two varying orientations θ_+ and θ_- correspond to the gradient and the isophote direction. The scale parameter σ of Gaussian-filter kernel is used to control the boundary strength. The coarse-to-fine structure of scale-space provides the best trade-off between detection and localization performance. In the *Helmholtz* theory, any vector field $\overrightarrow{F}$ can be represented as a sum of a conservative and solenoidal vector field with vector potential $\overrightarrow{A}$.

$$\vec{F} = \vec{F}_{con} + \vec{F}_{sol} = -\vec{\nabla}V + \vec{\nabla} \times \vec{A} \tag{6.2}$$

Taking the divergence of both sides as

$$\vec{\nabla} \cdot \vec{F} = -\Delta V + \vec{\nabla} \cdot (\vec{\nabla} \times \vec{A}) \tag{6.3}$$

where Δ is the oriented Laplacian shown in Chapter 5.4. Since the second term $\overrightarrow{\nabla} \cdot \left(\overrightarrow{\nabla} \times \overrightarrow{A}\right)$ is zero, the boundary function V can be solved by the Poisson equation. Let the image be a continuous function that is only divided by visual edges into $n+1$ regions $\{R_0, \ldots, R_i, R_j, \ldots, R_n\}$. The supporting region $\mathcal{W}_\sigma$ is calculated by combining similar pixels enclosed by boundary

[131].

$$\mathcal{W}_\sigma(V) = \int_{R_i} w_\sigma \left(I_p(\mathrm{x}), I_q(\mathrm{x})\right) d\mathrm{x} + \int_{R_i}\int_{R_j} w_\sigma \left(I_p(\mathrm{x}), I_q(\mathrm{x})\right) d\mathrm{x} \tag{6.4}$$

where x = [u, v], and $w_\sigma\left(I_p, I_q\right)$ is a dissimilarity function with a scale between the neighboring pixels $\{I_p\left[(\theta_+, \theta_-), \sigma\right], I_q\left[(\theta_+, \theta_-), \sigma\right]\} \in R_N$ on the smoothly varying oriented structure of Equation 5.22. The first and second terms respectively purpose the grouping of similar pixels in a region, e.g. R_i and R_j.
Dense disparity vectors $\mathbf{d}$ are locally estimated in the supporting region $\mathcal{W}_\sigma$ by matching the perceptual maximum intensity and then the vectors are refined by using a coarse-to-fine scheme. The energy function is

$$E_\Omega(\mathbf{d}) = \int_\Omega \rho\left[I_l(u,v) - I_r(u + \mathbf{d}_{l\to r}(u) \in \mathcal{W}_\sigma(V), v)\right]^2 d\mathrm{x} + \lambda \int_\Omega e_\sigma(\mathbf{d}) d\mathrm{x} \tag{6.5}$$

Ω is the image domain, the subscripts denote the matching direction, e.g. $l \to r$ for left-to-right direction and e_σ is a regularization term with Lagrange multiplier λ. This method results in an accurate local disparity estimation since matching errors are restricted in $\mathcal{W}_\sigma$. ρ is the determinant of a potential function that derives boundary flow to restrict the outliers of matching, and it is a kind of bounded loss function[50] for the robust estimation introduced in Chapter 5.1.

Anisotropic regularization : Although the bounded loss function can eliminate outliers due to mislocalization in the stereo matching, it does not always guarantee the homogeneous proposition. Traditional morphological operations frequently consider an isotropic smoothing to enhance the local homogeneity in an image. However, isotropic smoothing over disparity boundaries often causes conspicuous blurring errors. We regularize a disparity vector field in Equation 6.5 by an edge-preserving anisotropic diffusion term e_σ as

$$\begin{aligned} e_\sigma &= \psi[\nabla I_{l,\sigma}(u,v), \nabla \mathbf{d}_{l\to r}(u,v)] du dv \\ &= g(\|\nabla I_{l,\sigma}\|) \nabla \mathbf{d}_{l\to r} \end{aligned} \tag{6.6}$$

$\psi(\nabla)$ is a modified version for disparity of the discrete Perona and Malik diffusion model [137, 143], which has a form as $\psi(\nabla) = g(\nabla)\nabla'$ in the influence function. $g(\nabla)$ is a redescending weight function that acts as a *edge-stopping function* of diffusion. This function modifies the diffusion coefficient at edges and to derive discontinuity. It is possible to choose a error weight g of robust estimator shown in Table 5.1, and our choice of g is the weighting function of the Perona and Malik model as

$$g(\nabla) = e^{-(\nabla^2/K^2)} \tag{6.7}$$

50) see Equation 5.7

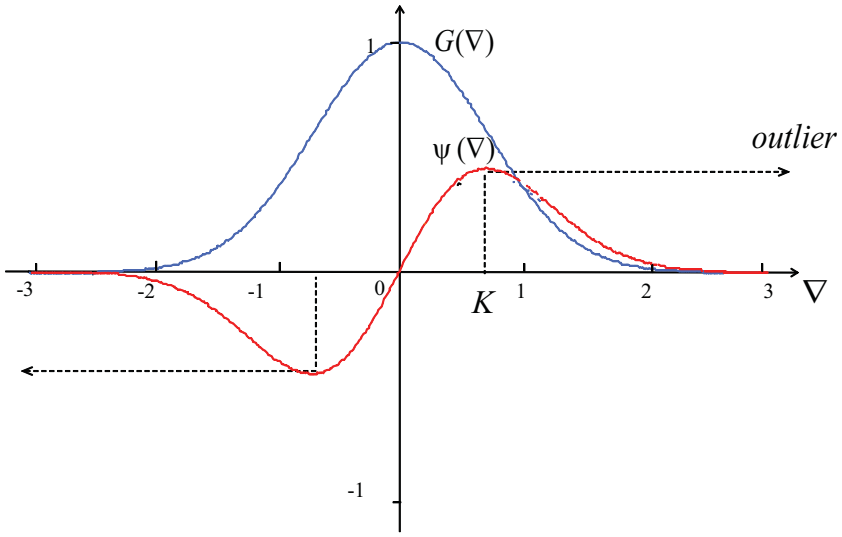


Figure 6.4. Anisotropic weighting and influence functions in the Perona and Malik model

where a positive constant K that controls the level of contrast of edges to affect the diffusion process. Figure 6.4 depicts the anisotropic edge-stopping function $g(\nabla)$ and influence function $\psi(\nabla)$. The influence function starts reducing the diffusion when the gradient magnitude increases beyond a fixed point that is determined by a scale parameter. Thus, errors in the range ($K \sim \infty$) can be removed as outliers. Since the stereo images are pre-filtered using different Gaussian scales for local appearance matching, the scale-space is isotropic. The influence function modifies the isotropic diffusion in the scale-space into anisotropic diffusion near discontinuities by decreasing on the range ($0 \sim K$). The anisotropic diffusion enhances a smooth disparity of a homogeneous region and simultaneously preserves important disparity discontinuities.

Numerical solution : Diffusion is a process that equilibrates spatial variations with concentration. Thus, we iteratively solve the energy minimization problem of Equation 6.5 using the Euler-Lagrange equation and the asymptotic analysis of *parabolic PDEs* as

$$\begin{aligned} \partial_t \mathbf{d}_{l\to r}(u,v,t) &= \lambda \mathrm{div}\left[g\left(\|\nabla I_{l,\sigma}(u,v)\|\right)\nabla \mathbf{d}_{l\to r}(u,v)\right] \\ +\frac{\partial I_r(u+\mathbf{d}_{l\to r}(u),v)}{\partial u} &\left[I_l(u,v) - I_r(u+\mathbf{d}_{l\to r}(u),v) \in \mathcal{W}_{l,\sigma}(V)\cdot\vec{N}\right] \end{aligned} \tag{6.8}$$

A matching region $\mathcal{W}$ can be considered as the regional boundary[51] that has the range of global disparity vectors. The range gives us a regional constraint to restrict matching errors within the region $\mathcal{W}$. Windows for every pixel can be established by a level-set function.

51) The morphological characteristic of regional diffusion can be simply written as $a_t = La$ for $a(u,v,t)$ where L is an elliptic operator.

The dense matches in the regions of a stereo image pair are performed locally, and the regularization is globally performed by the iterative solution of PDEs shown in 6.8. The PDEs minimize the total energy of Equations 6.5 and 6.6. A solution of the parabolic problem can be discretized using finite differences and coincides with the image filtering. An inhomogeneous time diffusion process with discrete sampling solves the problem as follows:

$$\begin{aligned}\frac{\mathbf{d}^{t+1}_{l\to r}-\mathbf{d}^{t}_{l\to r}}{\tau_u} &= \lambda \mathrm{div}\left[g\left(\|\nabla I_{l,\sigma}(u,v)\|\right)\nabla \mathbf{d}^{t}_{l\to r}(u,v)\right]+\\ \frac{\partial I_r(u+\mathbf{d}^{t}_{l\to r}(u),v)}{\partial u}&\left[\left[I_l(u,v)-I_r(u+\mathbf{d}_{l\to r}(u),v)\in \mathcal{W}_{l,\sigma}(V)\cdot\vec{N}\right]\right.\\ &\left.+\frac{\partial I_r(u+\mathbf{d}^{t}_{l\to r}(u),v)}{\partial u}\left(\mathbf{d}^{t+1}_{l\to r}-\mathbf{d}^{t}_{l\to r}\right)\right]\end{aligned} \tag{6.9}$$

The inhomogeneous time process, each pixel is diffused with a different time scale related to the pixel confidence. High confidence pixels are diffused much slower than low confidence pixels. By increasing the time step τ_σ relatively to the filter scale, outliers e.g. mismatches, mislocalization error and occlusion hole etc., in a partially homogeneous disparity field can be removed, and the boundaries are simultaneously preserved by the edge-stopping function g that suppresses diffusion in an area of high gradient. One of the most steepest weighting function $g(\nabla)=e^{-(\nabla^2/K^2)}$ is considered for the strong boundary preservation.
The numerical solution includes two steps: the gradient and the divergence. We do not use central difference approximation for the gradient and the divergence because it may result in an unconditionally unstable scheme. Instead, we can use forward differences for gradient and backward differences for the divergence as

$$\frac{\partial I(u,v)}{\partial u}\approx \delta^{(+)}_u I(u,v) \quad \text{and} \quad \frac{\partial I(u,v)}{\partial v}\approx \delta^{(+)}_v I(u,v) \tag{6.10}$$

and

$$div\left[\mathrm{g}\left(\nabla\right)\nabla'\right]\approx \delta^{(-)}_u\left[\mathrm{g}\left(\nabla\right)\nabla'\right]+\delta^{(-)}_v\left[\mathrm{g}\left(\nabla\right)\nabla'\right] \tag{6.11}$$

where $(+)$ and $(-)$ respectively denote the forward direction and the backward direction of a gradient operator. In a 2-D image, a gradient operator yields a 2×2 matrix, and the divergence operator collapses this into a 2×1 vector. The edge stopping function $g(\nabla)$ and the influence function $\psi(\nabla)=g(\nabla)\nabla'$ are calculated for the u and v axis. The columns of the matrix are independently computed, and the backward differences of the matrix are used to compute the divergence operation. Finally we can obtain the final numerical form of

Equation 6.9 as

$$\begin{aligned}\frac{\mathbf{d}_{l\to r}^{t+1}-\mathbf{d}_{l\to r}^{t}}{\tau_u} &= \lambda\Big[\delta_u^{(-)}\Big(e^{-[\delta_u^{(+)}\left(I_{l,\sigma}(u,v)\right)^2/K^2]}\delta_u^{(+)}\mathbf{d}_{l\to r}^{t}(u,v)\Big)\\ &\quad+\delta_v^{(-)}\Big(e^{-[\delta_v^{(+)}\left(I_{l,\sigma}(u,v)\right)^2/K^2]}\delta_y^{(+)}\mathbf{d}_{l\to r}^{t}(u,v)\Big)\Big]\\ &\quad+\delta_u^{(+)}\mathbf{d}_{l\to r}(x)\Big[\Big(I_l(u,v)-I_r(u+\mathbf{d}_{l\to r}^{t}(u),v)\in\mathcal{W}_{l,\sigma}(V)\cdot\vec{N}\Big)\\ &\quad+\delta_u^{(+)}I_r(u,v)(\mathbf{d}_{l\to r}^{t+1}-\mathbf{d}_{l\to r}^{t})\Big]\end{aligned} \tag{6.12}$$

6.2 Anisotropic Disparity Estimation with Perceptual Maximum Variation

In the previous chapter, multiple resolutions of a scale-space provided the best trade-off between the feature detection and the localization performances. Global disparity could be estimated with the constraint of strong and meaningful boundaries in coarse resolution, and then it is iteratively refined into finer resolutions by using PDEs. While the disparity field is diffusively smoothed by isotropic propagation following streamlines in the Gaussian scale-space, anisotropic diffusion suppresses the length of propagation only for the orthogonal direction of edges. However, the anisotropic regularization often causes over-diffusion artifacts during evaluating small brightness variations in the images. If important object boundaries are formed by small brightness variations, mislocalized flow causes over-diffusion problems, smoothing the disparity discontinuities of objects. We solve this small variation problem by the perceptual maximum variation modeling.

A $\mathcal{P}$-space with perceptual maximum variations in a high dynamic range is obtained by the method described in Chapter 5.5. Figure 6.5(a) shows $\mathcal{P}$-space obtained from a well-known stereo test image. The stereo image pair has a low dynamic range and contains complex objects at different depths generating several occlusions, as well as poorly-textured regions in the backgrounds. Most luminance matching methods, even high complex algorithms cannot generate an accurate dense disparity map as Figure 6.6 exemplifies, although our eyes can distinguish a lamp with red color and a white statue from the backgrounds due to the chromaticity difference as Figure 6.5(b) represents. The lamp cannot be found in the luminance image shown in Figure 6.5(c), and the histogram equalization of a luminance image cannot enhance the region as much as visual differences as Figure 6.5(d) depicts. Figure 6.5(e) shows the $\mathcal{P}$-space using maximum modulation of perceptual variations projecting all color variation into the $\mathcal{P}$-space with a high dynamic range. Comparison of detected edges among Figures 6.5(f), 6.5(g) and 6.5(h) exemplify that the visual difference can be well enhanced in the $\mathcal{P}$-space.

Outliers of disparity estimation are removed by a modified anisotropic diffusion function that

is derived in the $\mathcal{P}$-space:

$$e_\sigma = g(\|\nabla \mathcal{P}_{l,\sigma}\|)\nabla d_{l\to r} \tag{6.13}$$

where $\nabla \mathcal{P}_\sigma$ is the perceptual maximum variation with a scale in Equation 5.30, and $g(\nabla)$ is an anisotropic diffusion function in Equation 6.7 and 6.6. This modifies the diffusion coefficient at the edges in the $\mathcal{P}$-space and derives discontinuities in the robust influence function $\psi(\nabla) = g(\nabla)\nabla'$. The energy function is very similar to Equation 6.5 where the image I is modified to space $\mathcal{P}$.

$$\begin{aligned} E = \int_\Omega \rho[\mathcal{P}_l(u,v) - \mathcal{P}_r(u + \mathbf{d}_{l\to r}(u) \in \mathcal{W}_{l,\sigma}(V), v)]^2 d\mathrm{x} \\ +\lambda \int_\Omega e_\sigma(\mathbf{d}_{l\to r}) d\mathrm{x} \end{aligned} \tag{6.14}$$

where $\mathrm{x} = [\mathrm{u}, \mathrm{v}]$ is the 2-D image pixel, and $\mathcal{W}_\sigma$ is the window shown in Equation 6.4. We solve the energy-minimization problem by discretizing the following numerical equation using finite differences.

$$\begin{aligned} &\frac{\mathbf{d}^{t+1}_{l\to r} - \mathbf{d}^t_{l\to r}}{\tau_\sigma} = \lambda div \left[\left(\frac{g(\|\nabla \mathcal{P}_{l,\sigma}\|)}{\|\nabla \mathcal{P}_{l,\sigma}\|} \right) \nabla \mathbf{d}^t_{l\to r} \right] + \\ &\frac{\partial \mathcal{P}_{r,\sigma}(u + \mathbf{d}^t_{l\to r}, v)}{\partial \mathrm{x}} \left[\left(\mathcal{P}_{l,\sigma}(u,v) - \mathcal{P}_{r,\sigma}(u + \mathbf{d}_{l\to r}, v) \in \mathcal{W}_\sigma(V) \cdot \vec{N} \right) \right. \\ &\qquad \left. + \frac{\partial \mathcal{P}_{\sigma,r}(u + \mathbf{d}^t_{l\to r}(u), v)}{\partial \mathrm{x}} (\mathbf{d}^{t+1}_{l\to r} - \mathbf{d}^t_{l\to r}) \right] \end{aligned} \tag{6.15}$$

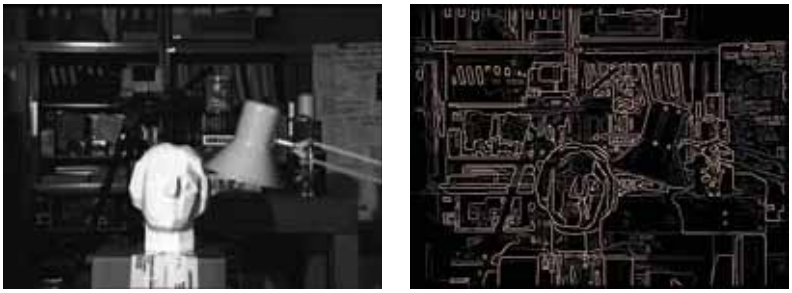
To remove outliers of matching, a local matching region with the ranges is fitted using normal vector $\vec{N}$ of the edge following streamlines of scale-space. The divergence term *div* performs a global diffusion of the local disparity map to remove outliers in each scale. The other parts of the equation are very similar to Equation 6.9. Figure and respectively represent a result of the proposed method and the error image for a given ground-truth in [35].

6.3 Robust Hybrid Recursive Optical Flow Estimation

The motion estimation from a sequence of dynamic images is one of the most important tasks in computer vision. Two different approaches, i.e. discrete motion estimation and optical flow estimation, have been separately developed. The discrete motion estimation establishes the correspondences by measuring similarity using blocks or masks. The advantages are simplicity and reliability for discrete large motion. However, detailed motion of a deformable-body cannot be recovered because block and mask are inherently rigid with translational motion. Blocking artifacts and poor motion prediction along the moving boundaries are serious drawbacks. The hybrid recursive method was first proposed in [163] to solve the problems. The method recursively refines block motion vectors to arrive at a dense motion vector field using a gradient technique. It can be hierarchically applied using multi-resolution representation

(a) (b)

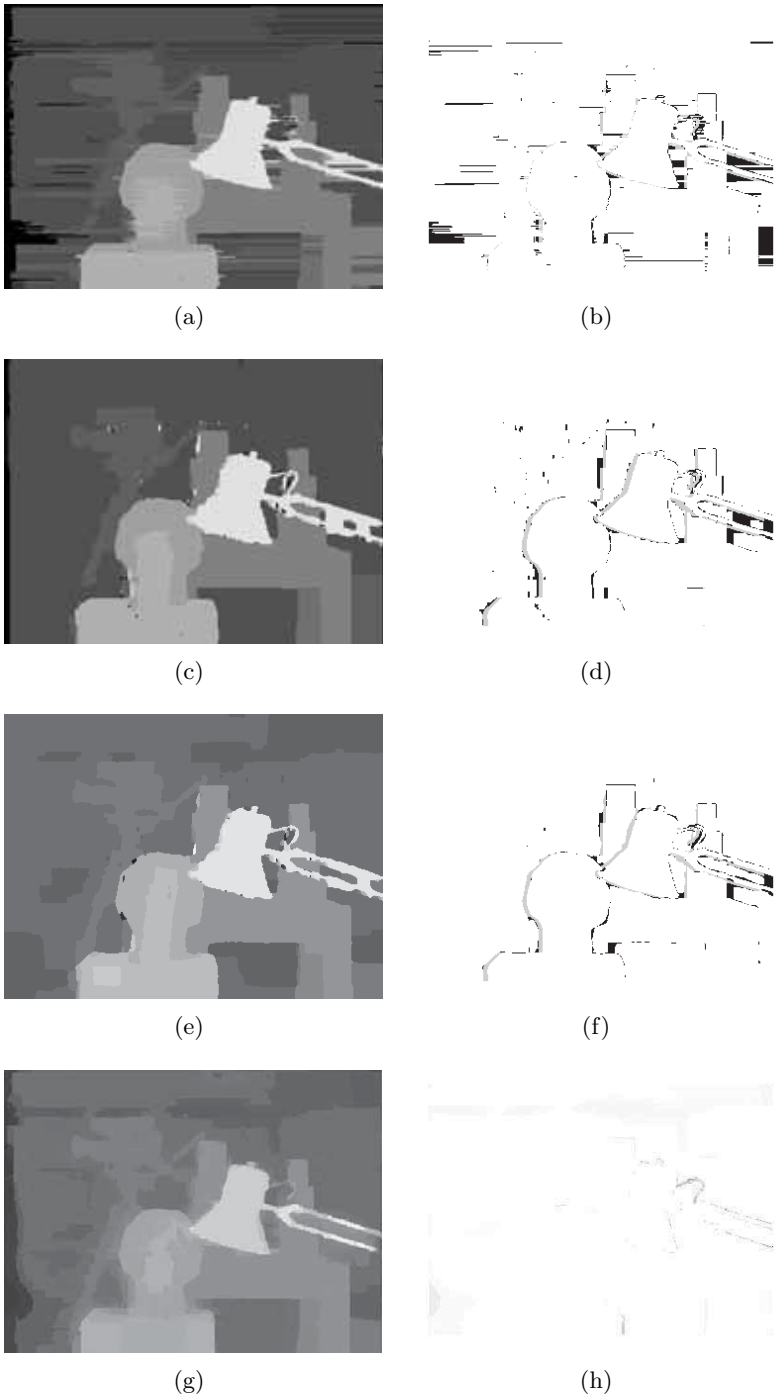
(c) (d)



(e) (f)

(g) (h)

Figure 6.5. Discontinuities in perceptual maximum variations - (a) original color image, (b) the chromaticity image (a^*b^*), (c) the luminance image (L^*), (d) histogram equalization on the luminance space, (e) the maximum variation modeling result, (f) edge image of Figure c, (g) edge image of Figure d, (h) edge image of Figure e



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 6.6. Evaluation using ground-truth - (a) dynamic programming method, (b) error image of dynamic programming method, (c) graph-cut method, (d) error image of graph-cut method, (e) hierarchical belief propagation method, (f) error image of hierarchical belief propagation method, (g) proposed method, (h) error image of proposed method

for the accuracy. On the other hand, optical flow estimation [164–171] aims to obtain a delicate velocity field using the computation of spatial and temporal image derivatives. Most methods define the displacement of pixels based on brightness-conserving assumptions and gradient constancy assumptions or both.

(a) Brightness-conserving assumption : the brightness of objects in subsequent frames does not change by small movements.
(b) Gradient consistency assumption : It determines the displacement with a criterion that is invariant against the change of brightness.

Using the partial derivatives, the optical flow efficiently handles the piecewise and detailed variation of displacement. However, the discontinuity from a large motion causes inappropriate matching and ill-posed local minima. In order to overcome these defects, hierarchical algorithms have been proposed based on a multi-resolution representation [164, 165]. A coarse but robust estimation of the motion field is obtained at a low level. Then, it is iteratively refined at higher levels. A motion trajectory can be also considered instead of using low levels [170]. From a motion trajectory, a dense flow field is estimated as a process of interpolation. Some researches [167–170] have exploited robust parameter estimation to remove outliers. The robust estimation is less sensitive to the outliers using the localization of the motion field. The localization methods are classified into two approaches. The first approach concentrates on removing outliers of a local energy model to determine the best flow within a region. For example, the robust quadratic estimation which uses least squares is solved by a regression [168, 169]. The second approach corresponds to the design of a global energy model with a regularization function which preserves the discontinuity [164–167]. The outliers are removed by smoothing each homogeneous motion area. For example, one model considering a robust M-estimator instead of the quadratic estimator is proposed in BlackAnandan1996. The most recent approach is the combination of the local energy model and global energy model [167].
In this Chapter, we combine the advantage of the discrete motion estimation and the optical flow estimation in an efficient common scheme, leading to a hybrid recursive energy-based method. The proposed method [162] is described with two frameworks: incremental updating framework and robust energy minimization framework.

Incremental updating framework : It unites the discrete motion estimation and the optical flow estimation in the concept of *displaced frame difference* (DFD). When the discrete motion estimation recursively calculates the *displaced block difference* (DBD) from a new frame, the optical flow iteratively refines using the DBDs as the initial value. Incremental detailed motion values from the new frame with the previous optical flow are recursively updated using the *displaced pixel difference* (DPD). Incrementally integrated motion has a lot of advantages, i.e. the motion can be accessed at any time. Thus, motion is temporally

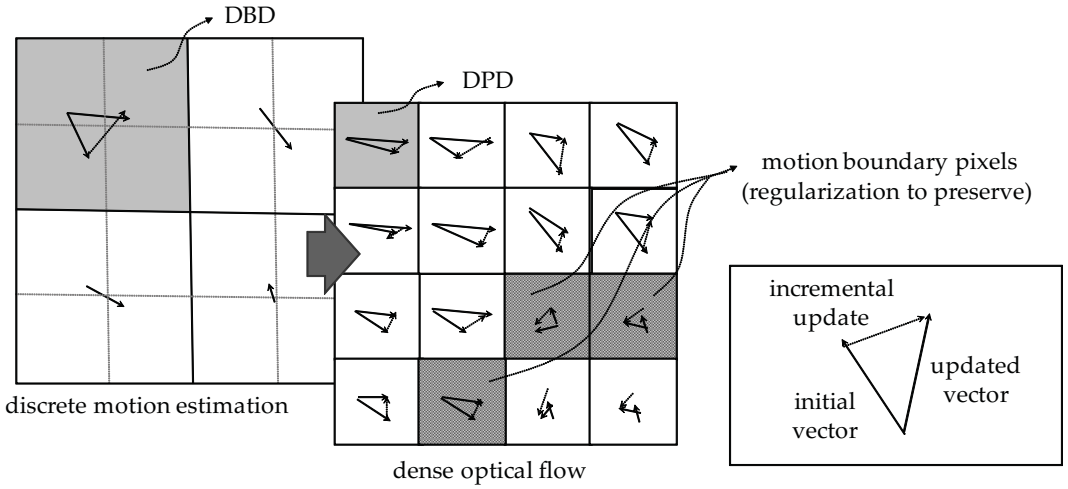


Figure 6.7. Incremental updating framework

refined. Only pairs of frames need to be analyzed, therefore the amount of computation is reduced. Moreover, it is adaptive to the changes of motion and luminance over long time. If we assume large motion (du, dv), integration of incremental motion is defined as translational displacement mapping.

$$I(u, v, t) = I(u + du, v + dv, t + dt) \tag{6.16}$$

The best correspondence of intensity $I(u, v)$ is found by matching over the horizontal, vertical and temporal increments (du, dv, dt). The matching algorithm usually searches a candidate set of motion vector (du, dv) which minimizes some functions of displaced frame difference (DFD) as

$$DFD(\mathrm{x}, \mathbf{d}) = \mathrm{I_t(x)} - \mathrm{I_{t+1}(x + \mathbf{d})} \tag{6.17}$$

where $\mathbf{d}$ is the total displacement on the image domain x = [u,v]. By taking a Taylor series, Equation 6.17 is linearized as

$$I_t(\mathrm{x}) = I_{t+1}(\mathrm{x}) - \mathbf{d}^T \nabla I_{t+1}(\mathrm{x}) - e_{t+1}(\mathrm{x}) \tag{6.18}$$

∇ is the multidimensional gradient operator and $e_{t+1}(\mathrm{x})$ represents the higher order terms of the expansion to be set up at each pixel in a block. Using 6.17 and 6.18, the expression can be rearranged involving the (DFD) with zero displacement as

$$DFD(\mathrm{x}, 0) = I_t(\mathrm{x}) - I_{t+1}(\mathrm{x}) = \mathbf{d}^{\mathrm{T}} \nabla I_{t+1}(\mathrm{x}) + e_{t+1}(\mathrm{x}) \tag{6.19}$$

d is estimated by calculating the sum of displaced block difference (DFD) and displaced pixel difference (DPD) for a pair of frames.

$$\mathbf{d} = DBD\left(U_B, V_B\right) + DPD\left(u_P, v_P\right) \tag{6.20}$$

The subscripts $_B$ and $_P$ denotes a calculation respectively using blocks and pixel.

Robust energy minimization framework : The optical flow should be robustly estimated inside the homogeneous region of a moving object. It should be recovered without smoothing across the motion discontinuity. Robust error norm of anisotropic diffusion and adaptive spatial-temporal anisotropic regularization are integratively applied in this framework to preserve the motion boundary. The local energy model and the global energy model are used to remove outliers. DBD and DPD are defined by a large increment (U_B, V_B) and a small increment (u_P, v_P) between any pair of frames. First, the images are pre-filtered, then down-sampled to remove noise and to reduce the system cost. The final DBD is hierarchically refined from the minimization of the correspondence energy E_B with the multi-dimensional blocks sizes M and N.

$$E_B(U,V) = \sum_{u=0}^{M}\sum_{v=0}^{N} \left|I_t(u,v) - I_{t+1}(u+U_B, v+V_B)\right|^2 \tag{6.21}$$

When $DBDs$ are calculated in the block recursive stage, $DPDs$ are estimated by iteratively refining the $DBDs$ as the initial value of the pixel recursive stage in the gradient solution in 6.18. The final output vector flow is obtained by recursively updating the increments between the $DBDs$ and $DPDs$. The method achieves highly reliable and per-pixel total displacements to prevent local minima for large motion. Figure 6.7 illustrates the method.

Robust statistics that was described in Chapter 5.3 have been applied to remove outliers in many computer vision problems. It yields robust results for smoothing homogeneous areas and preserving the discontinuity boundaries. A diffusion function $g(s) = \left[1 + (s/\varepsilon)^2\right]^{-1}$ which is called the *edge-stopping function* suppresses diffusion in areas of high gradients. A constant ε controls the level of contrast of edges to affect the smoothing process. The discrete Perona and Malik diffusion is obtained by integrating $(s) \cdot s$ into a non-convex potential energy as

$$\rho_g = \int \left[g(s) \cdot s\right] ds = \int \left[\psi(s)\right] ds \; = \sigma^2 \log\left[1 + 1/2\left(s^2/\sigma^2\right)\right] = \sigma^2 \rho_L \tag{6.22}$$

where the influence function $\psi(s)$ encloses $g(s) \cdot s$. The derivative $div[g(s) \cdot s]$ modifies the diffusion coefficient at the boundary. Equation 6.22 shows that ρ_g for $\varepsilon^2 = 2\sigma^2$ can be equivalently treated with Lorentzian error norm ρ_L. Fig. 2 illustrates the relationship of $g(s)$, $\psi(s)$ and ρ_g. In this chapter, we integrate the robust error norm into the energy

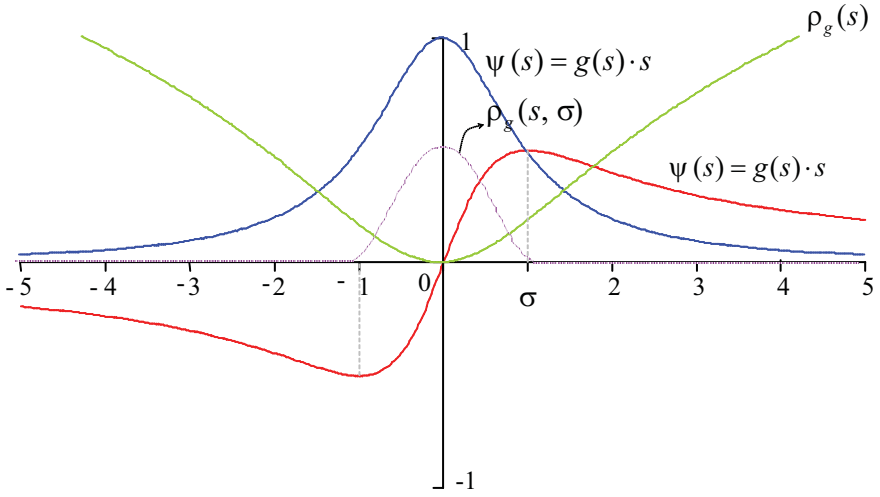


Figure 6.8. Robust anisotropic function

minimization to estimate the optical flow. The edge-preserving robust statistic considers the motion boundaries as those points to be *outliers* between piecewise smooth flow regions. Considering the brightness-conserving assumption and the gradient constancy assumption for 6.16, we derive a robust energy function that penalizes deviations in Equation 6.23

$$E_D(\mathbf{d}) = \int_\Omega \rho_g \left(|I(\mathrm{x}) - I\,(\mathrm{x}+\mathbf{d})|^{\mathbf{2}} + \kappa\, |\nabla I(\mathrm{x}) - \nabla I\,(\mathrm{x}+\mathbf{d}))|^2 \right) d\mathrm{x} \tag{6.23}$$

κ is a proportional weight coefficient between the two terms. Now, the goal of robust estimation of ρ_G is to find the threshold σ to remove outliers. Black [137] achieved the appropriate σ considering Tukey's biweight norm[52] as

$$\rho_g(s,\sigma) = \begin{cases} 1/2\left[1-(s/\sigma)^2\right]^2 & \text{if } |s| \le \sigma \\ 0 \quad \text{(i.e.outlier)} & \text{otherwise} \end{cases} \tag{6.24}$$

The robust local flow can be achieved by removing the outliers. Now, we consider a global approach using a regularization function preserving the motion discontinuity. An anisotropic diffusive regularization in 6.25 is applied to a global energy.

$$E_R(\mathbf{d}) = \int_\Omega \psi\left(\nabla I(\mathrm{x}), \mathbf{d}(\mathrm{x})\right) d\mathrm{x} \tag{6.25}$$

$\psi\left(\nabla I(\mathrm{x}), \nabla\mathbf{d}(\mathrm{x})\right) = g\left(\|\nabla_3\mathbf{d}\|\right)\nabla\mathbf{d}$ is a modified version of discrete Perona and Marik diffusion equation $g(s)\cdot s$ and $\nabla_3 = (\partial_u, \partial_v, \partial_t)$ denotes the spatial-temporal gradient [171]. The diffusion direction is sensitivity-adaptively controlled by using different step-sizes for the

52) see Equation 5.9 and 5.8.

spatial gradient $\nabla I(\mathrm{x})$ and the temporal gradient $\nabla \mathbf{d}(\mathrm{x})$ in a discrete sense. Finally, the total energy term which combines $E_D(\mathbf{d})$ of Equation 6.23 with $E_R(\mathbf{d})$ in Equation 6.25 using the Lagrange multiplier λ is obtained in Equation 6.26.

$$E(\mathbf{d}) = E_D(\mathbf{d}) + \lambda E_R(\mathbf{d}) \tag{6.26}$$

Numerical Solution : We iteratively solve the minimization problem of 6.26 by the associated Euler-Lagrange equations as

$$\begin{aligned} &\rho_g \left(I_t^2 + \kappa(I_{ut}^2 + I_{vt}^2) \cdot (I_u I_t + \kappa(I_{uu} I_{ut} + I_{uv} I_{vt})\right) \\ &-\lambda div \left(g \left(\|\nabla_3 \mathbf{d}\|\right) \cdot \nabla u\right) = 0 \end{aligned} \tag{6.27}$$

and

$$\begin{aligned} &\rho_g \left[I_t^2 + \kappa(I_{ut}^2 + I_{vt}^2) \cdot (I_v I_t + \kappa(I_{vv} I_{vt} + I_{uv} I_{ut})\right] \\ &-\lambda div \left(g \left(\|\nabla_3 \mathbf{d}\|\right) \cdot \nabla v\right) = 0 \end{aligned} \tag{6.28}$$

where we define the abbreviations of spatial-temporal derivations as

$$\begin{aligned} &I_u = \partial_u I(\mathrm{x} + \mathbf{d}), \quad \mathrm{I_v} = \partial_\mathrm{v} \mathrm{I}(\mathrm{x} + \mathbf{d}), \quad \mathrm{I_t} = \mathrm{I}(\mathrm{x}) - \mathrm{I}(\mathrm{x} + \mathbf{d}), \\ &I_{uu} = \partial_{uu} I(\mathrm{x} + \mathbf{d}), \quad \mathrm{I_{uv}} = \partial_\mathrm{uv} \mathrm{I}(\mathrm{x} + \mathbf{d}), \quad \mathrm{I_{vv}} = \partial_\mathrm{vv} \mathrm{I}(\mathrm{x} + \mathbf{d}), \\ &I_{ut} = \partial_u I(\mathrm{x} + \mathbf{d}) - \partial_\mathrm{u} \mathrm{I}(\mathrm{x}), \quad \mathrm{I_{vt}} = \partial_\mathrm{v} \mathrm{I}(\mathrm{x} + \mathbf{d}) - \partial_\mathrm{v} \mathrm{I}(\mathrm{x}) \end{aligned} \tag{6.29}$$

The asymptotic analysis of a parabolic PDE with natural boundary conditions approximates $\mathrm{x}^{\mathrm{k}+1}$ using an iteration variable x^{k} in Equations 6.30 and 6.31. High confidence values are calculated much slower than low confidence values.

$$\begin{aligned} &\rho_g \left((I_t^{k+1})^2 + \kappa \left((I_{ut}^{k+1})^2 + (I_{vt}^{k+1})^2\right) \cdot (I_u^k I_t^{k+1} + \kappa(I_{uu}^k I_{ut}^{k+1} + I_{uv}^k I_{vt}^{k+1})\right) \\ &-\lambda div \left(g \left(\left\|\nabla_3 \mathbf{d}^{k+1}\right\|\right) \cdot \nabla u^{k+1}\right) = 0 \end{aligned} \tag{6.30}$$

and

$$\begin{aligned} &\rho_g \left((I_t^{k+1})^2 + \kappa \left((I_{ut}^{k+1})^2 + (I_{vt}^{k+1})^2\right) \cdot (I_v^k I_t^{k+1} + \kappa(I_{vv}^k I_{vt}^{k+1} + I_{uv}^k I_{ut}^{k+1})\right) \\ &-\lambda div \left(g \left(\left\|\nabla_3 \mathbf{d}^{k+1}\right\|\right) \cdot \nabla v^{k+1}\right) = 0 \end{aligned} \tag{6.31}$$

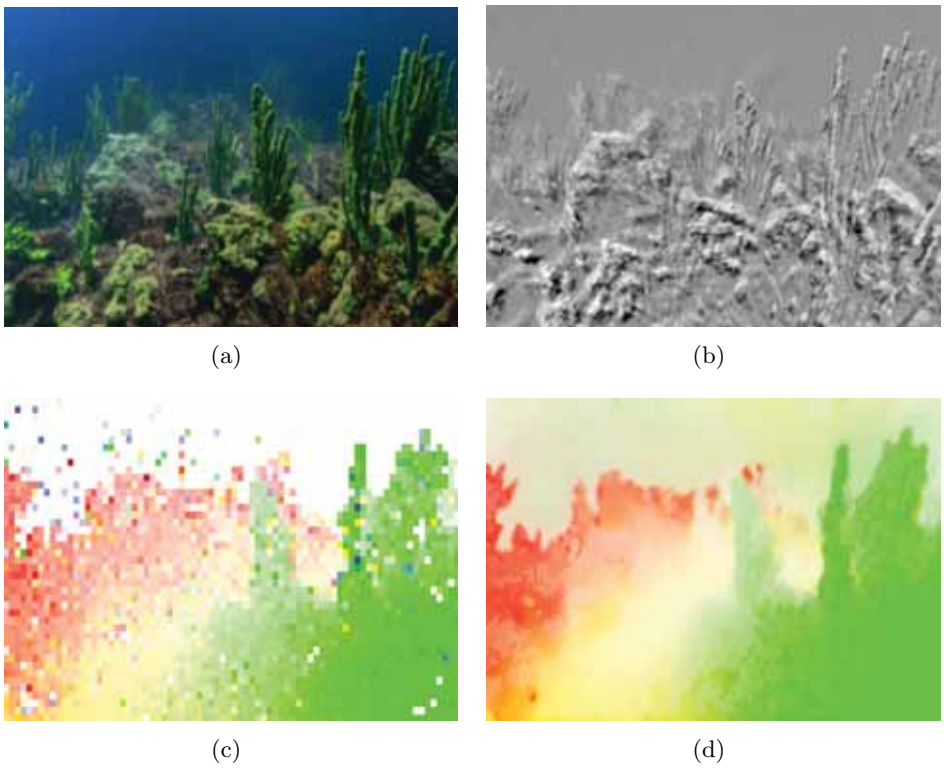
The computational complexity of the nonlinear system is removed retaining the first-order terms of the Taylor approximation in implicit discretization,

$$\begin{aligned} I_t^{k+1} &\approx I_t^k + (I_u^k du^k + I_v^k dv^k), \\ I_{ut}^{k+1} &\approx I_{ut}^k + (I_{uu}^k du^k + I_{uv}^k dv^k), \\ I_{vt}^{k+1} &\approx I_{vt}^k + (I_{uv}^k du^k + I_{vv}^k dv^k) \end{aligned} \tag{6.32}$$

where $u^{k+1} = u^k + du^k$ and $v^{k+1} = v^k + dv^k$. u^{k+1} and v^{k+1} are easily calculated using the previous iteration steps u^k, v^k and the incremental updates du^k, dv^k. Equation 6.30 can be rewritten into Equation 6.33. The case of Equation 6.31 can be solved in a similar way.

$$\begin{aligned} &\rho_g \left[\left(I_t^k + (I_u^k du^k + I_v^k dv^k)\right)^2 + \kappa \left\{ \left(I_{ut}^k + (I_{uu}^k du^k + I_{uv}^k dv^k)\right)^2 + \right.\right. \\ &\left.\left. \left(I_{vt}^k + (I_{uv}^k du^k + I_{vv}^k dv^k)\right)^2 \right\} \right]^\tau \cdot I_u^k \left(I_t^k + (I_u^k du^{k,\tau+1} + I_v^k dv^{k,\tau+1})\right) \\ &+\kappa\rho_g \left[\left(I_t^k + (I_u^k du^k + I_v^k dv^k)\right)^2 + \kappa \left\{ \left(I_{ut}^k + (I_{uu}^k du^k + I_{uv}^k dv^k)\right)^2 + \right.\right. \\ &\left.\left. \left(I_{vt}^k + (I_{uv}^k du^k + I_{vv}^k dv^k)\right)^2 \right\} \right]^\tau \cdot \left[I_{uu}^k \left(I_{ut}^k + (I_{uu}^k du^{k,\tau+1} + I_{uv}^k dv^{k,\tau+1})\right) + \right. \\ &\left. I_{uv}^k \left(I_{vt}^k + (I_{uv}^k du^{k,\tau+1} + I_{vv}^k dv^{k,\tau+1})\right) \right] \\ &-\lambda div \left(g \left(\left\| \nabla_3 (u^k + du^k) \right\| \right)^\tau \cdot \nabla (u^k + du^{k,\tau+1}) \right) = 0 \end{aligned} \tag{6.33}$$

The non-convex robust estimation problem converges into a local minimum. A solution is to utilize the previous step as a good initial estimate of the flow and to obtain the incremental value. The motion u^k and v^k are recursively estimated as $(U_B + u_P)^k$ and $(V_B + v_P)^k$. Whenever the updated values U_B and V_B from the block recursion are initialized, the new increments du^k and dv^k are recursively updated in the same way. Consequently, an additional iterative minimization strategy with a few steps τ converges quickly into a global minimum. Figure 6.9 exemplifies the initial estimate and refined optical flow result. The 1st frame image and difference image between the 1st and 2nd frames are shown in Figures 6.9(a) and 6.9(b). The image size is 640×360, and the block initial motions are obtained by a discrete motion estimation with anisotropic block regularization as Figure 6.9(c) represents. The incremental optical flow motions for image pixels are recursively updated in the block motion-compensated images. The final optical flow estimates shown in Figure 6.9(d) have the sum of discrete block motions, i.e. U_B and V_B, and incremental pixel motions, i.e. U_P and V_P. The recursive block-to-pixel estimation supports the good localization performance. The color code map shown in Figure 6.10 is efficient to visualize the orientation and the magnitude of a dense flow vector. The color represents the orientation of the vector and brightness stands for its magnitude. In Figure 6.11, the convergence of the proposed method is compared to the well-known Lukas-Kanade optical flow method. The Lukas-Kanade method cannot estimate motions in homogeneous regions, e.g. sky, ground and shadow, and some texture regions, e.g. the walls of a castle and trees, and results in the low density motion field. Figures 6.11(a)



(a) (b)

(c) (d)

Figure 6.9. Optical flow estimation result - (a) the 6th frame image, (b) difference image between the 6th and 12th frames, (c) discrete motion estimation result, (d) refined optical flow result

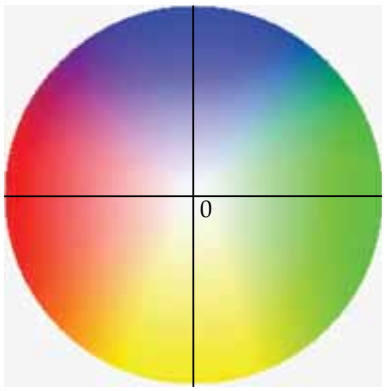
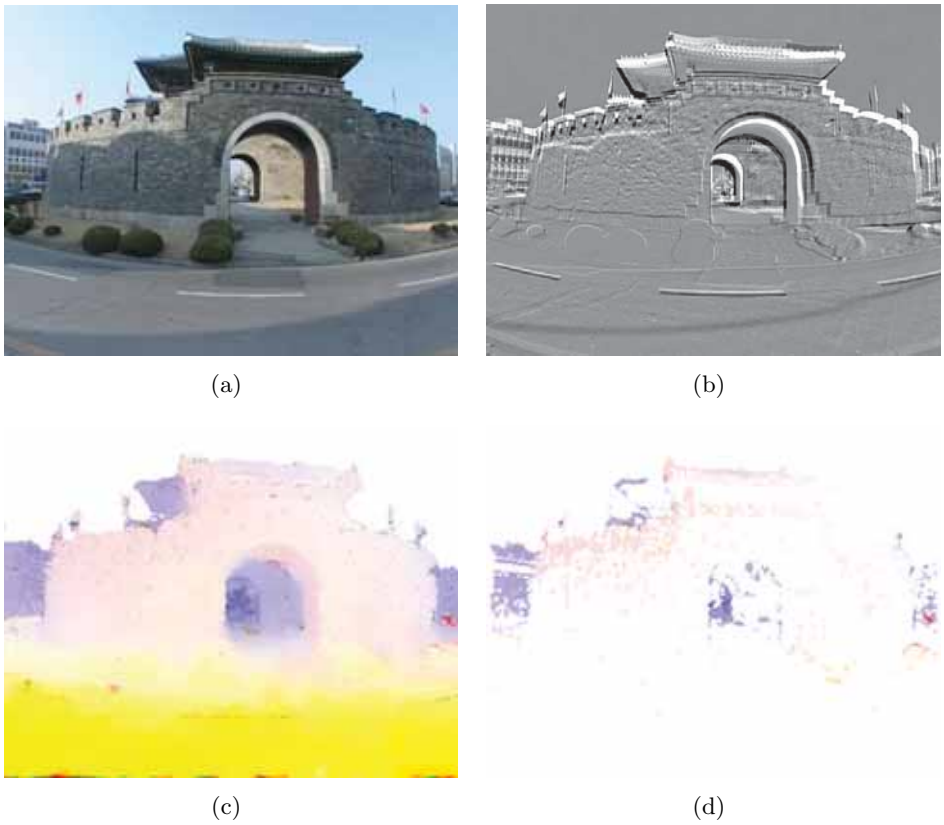


Figure 6.10. Color encoding of flow vectors

and 6.11(b) respectively show the 1st frame image and the difference image to the 2nd frame image. In comparing Figures 6.11(c) and 6.11(d), the proposed method has higher density motion field with correct flows than the Lukas-Kanade method.



(a) (b)

(c) (d)

Figure 6.11. Optical flow estimation result - (a) the 1st frame image, (b) difference image between the 1st and 2nd frames, (c) estimated motion field using the proposed method (d) motion field using Lukas-Kanade method

6.4 3-D Sceneflow Estimation

In both multi-stereo and structure-from-motion techniques, image correspondences established between a single pair of stereo or temporal images and the structural information derived from any combination of spatio-temporal images should be consistent. Chebaro et. al. [104] first use traditional matching methods to find spatio-temporal four frame sets of the feature correspondences: two stereo and two temporal, based on line segments and planar regions. Using the spatio-temporal 4-frame model, the consistency of matches can be checked. For any inconsistency, temporal matches are remained and the conflicting stereo matches are rejected. Another method [105] explicitly uses motion information to reject false stereo matches from a number of candidates or simply to confirm the validity of a stereo match, or vice versa.

Assuming a parallel stereo camera configuration, a point, $\tilde{\mathbf{M}}(X, Y, Z)$ in 3-D space projects to $I_L(u_L, v_L, f)$ in the first camera and $I_R(u_R, v_R, f)$ in the second one. Simple stereo geometry

allows us to recover the coordinates of $\tilde{\mathbf{M}}$ as

$$X = \frac{ub}{\mathbf{d}}, \quad Y = \frac{vb}{\mathbf{d}} \quad \text{and} \quad Z = \frac{bf}{\mathbf{d}} \tag{6.34}$$

where f is focal length, and b is baseline. The projection in each image plane $I_L(u_L, v_L, f)$ and $I_R(u_R, v_R, f)$, and the optical centers O_L and O_R define two similar triangles, so that we can write the relationship as

$$\frac{Z}{b} = \frac{Z - f}{b - (u_R - u_L)} \tag{6.35}$$

The image velocities of corresponding points in the two cameras are denoted by $v_L(u_L, v_L)$ and $v_R(u_R + \mathbf{d}, v_R)$ where $\mathbf{d}$ is the disparity vector and the difference is relative flow as

$$\Delta v_L(u_L, v_L, \mathbf{d}) = v_R(u_R + \mathbf{d}, v_R) - v_L(u_L, v_L) \tag{6.36}$$

The temporal derivative of Equation 6.34 is

$$\mathbf{V}_Z = -\frac{bf}{(u^R - u^L)^2} \times (v_u^R - v_u^L) = -\frac{Z^2}{bf}\Delta v_u \tag{6.37}$$

The 3-D velocity is given by $\mathbf{V} = \varphi \times \mathbf{P} + \mathbf{t}$, and the third component of the 3-D velocity of a point along the Z-direction is $\mathbf{V}_Z = t_Z + \varphi_X Y - \varphi_Y X$. The 3-D velocity in the Z-direction i.e. $\mathbf{V}_Z$ can be expressed as a linear equation on three of the six parameters that can be substituted in the equation of binocular flow.

$$\mathbf{t}_Z + \phi_X Y - \phi_Y X = -\frac{Z^2}{bf}\Delta v_u \tag{6.38}$$

and

$$\frac{\Delta v_u}{bf} = \begin{bmatrix} \frac{-1}{Z^2} & \frac{-v}{fZ} & \frac{u}{fZ} \end{bmatrix} \begin{bmatrix} \mathbf{t}_Z \\ \phi_X \\ \phi_Y \end{bmatrix} \tag{6.39}$$

where $\mathbf{t} = \begin{bmatrix} t_X & t_Y & t_Z \end{bmatrix}$ and $\varphi = \begin{bmatrix} \varphi_X & \varphi_Y & \varphi_Z \end{bmatrix}$ are the translational and rotational components of rigid-body motion, and the 3-D point coordinates X and Y are replaced by its inverse perspective projection equations.

Non-uniformly distributed 3-D samples : For non-uniformly distributed 3-D samples $\tilde{\mathbf{M}}$ on an object surface $S(\tilde{M})$, sampling positions are calculated by the bundle adjustment

method[53] minimizing the distances $\mathbf{D}$ between the sample X and the reprojected pixel $\tilde{m}_k$ on the k-th camera's image plane as

$$S(X) = \min_{\mathrm{X}} \sum_{k=1}^{n} \left(D(\tilde{\mathbf{m}}_k, \mathbf{P}_k^{-1} \tilde{\mathbf{M}})^2 \right) \tag{6.40}$$

where $\mathbf{P_k}$ denotes a projection from the k-th camera and $\mathbf{P} = (\mathbf{R}|\mathbf{t})$ is the projection matrix with rotational and translational components. Projecting samples in 3-D space to 2-D image plane leads to depth discontinuities in a staircase-like structure. Preserving the discontinuities is very important to avoid sampling artifacts. 3-D scene geometry is calculated by a joint method of global projection and local sceneflow correspondences. Pixels of image planes that are projected from multiple cameras can be connected in a chain by a disparity-link of each pair of planes. For an image triplet I_{k-1}, I_k, I_{k+1}, two image pairs (I_{k-1}, I_k) and (I_k, I_{k+1}) form two stereo image pairs. Two links of correspondence are calculated as

$$\tilde{\mathbf{m}}_{k+1} = \left(\mathbf{R}_{k+1}^{k} \right)^{-1} \mathbf{d}_{(k,k+1)} \left[\left(\mathbf{R}_{k}^{k+1} \right) \tilde{\mathbf{m}}_k \right] \tag{6.41}$$

and

$$\tilde{\mathbf{m}}_{k-1} = \left(\mathbf{R}_{k-1}^{k} \right)^{-1} \mathbf{d}_{(k,k-1)} \left[\left(\mathbf{R}_{k}^{k-1} \right) \tilde{\mathbf{m}}_k \right] \tag{6.42}$$

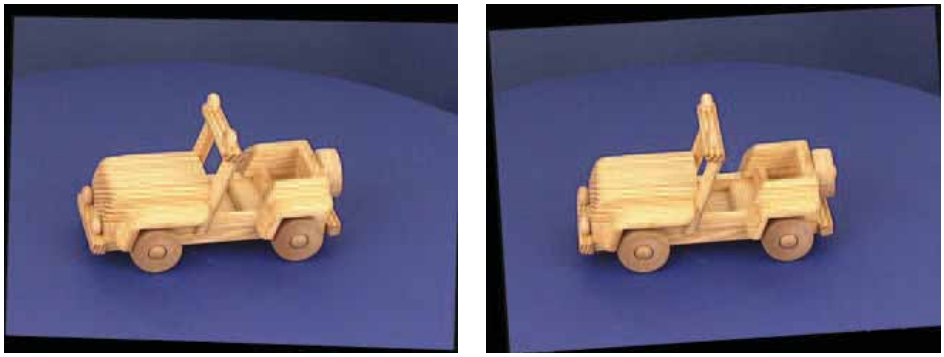
I_k^{k+1} and I_k^{k-1} respectively denote upward and downward rectified images by the projection matrices $\mathbf{P}_k^{k+1}$ and $\mathbf{P}_k^{k-1}$ transformed by rotational matrices $\mathbf{R}_k^{k+1}$ and $\mathbf{R}_k^{k-1}$. Two dense correspondence maps $\mathbf{d}_(k, k-1)$ and $\mathbf{d}_(k, k+1)$, respectively hold the downward correspondences from I_k^{k-1} to I_{k-1}^k and from I_k^{k+1} to I_{k+1}^k. This linking process is repeated along the whole n-input images to create long correspondence-chain connections. This method is very efficient to initialize dense and reliably 3-D samples in the scene geometry since the geometric density can be increased by overlapping distributed samples within small depth ranges. Higher depth ranges are obtained in long correspondence chains, and most occlusions caused by large baselines can be eliminated.

6.5 Experimental Results

Anisotropic dense disparity estimation : Image and video-based applications should be robust for indoor and outdoor environments to analyze natural images and videos. The robustness of dense disparity estimation can be tested for some difficult lighting conditions and camera setups such as mutual reflection, shadow, unbalanced brightness and unknown epipolar geometry. The first evaluation is performed using outdoor and indoor stereo image

53) see Appendix A.7.

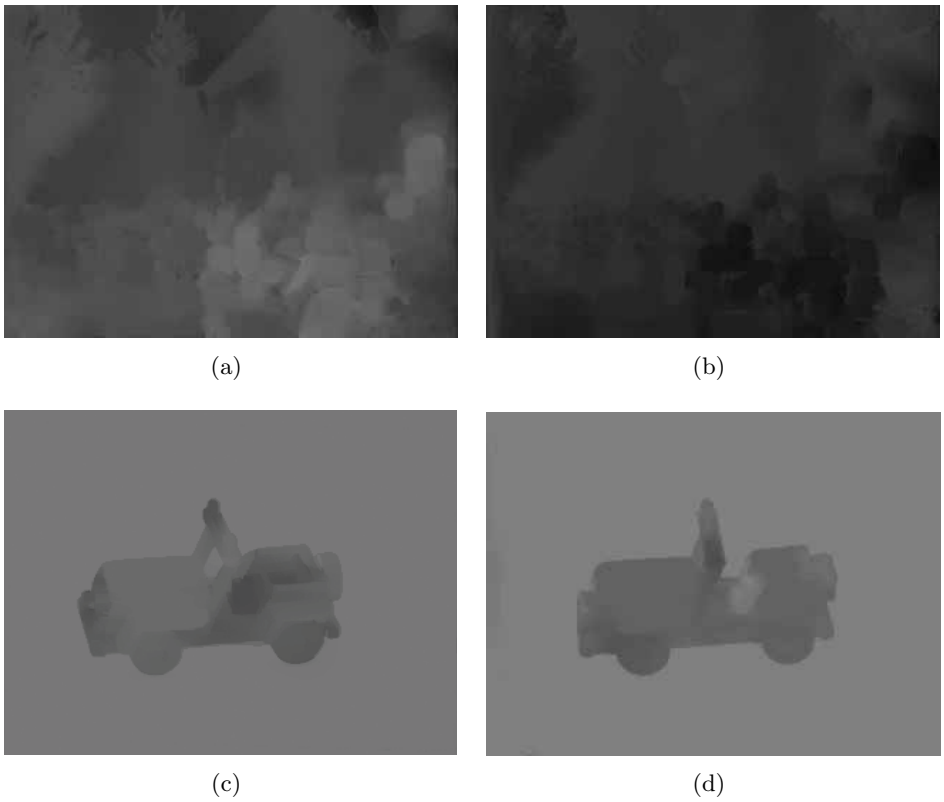
(a) (b)



(c) (d)

Figure 6.12. Test images with unbalanced lighting condition and homogeneous regions - (a) the left Balloon image, (b) the right Balloon image, (c) the left Car Toy image, (d) the right Car Toy image

pairs shown in Figure 6.12. A color stereo image pair with balloons, i.e. Balloon image, shown in Figures 6.12(a) and 6.12(b) have a size of 720×480 pixels and 24bits/pixel and contains difficult lighting conditions with imbalance between left and right images, i.e. left image is darker than the right image. Additionally, the image has a complex image structure, i.e. several objects with different disparities. Another stereo image pair, Car Toy captured in a studio has more stable lighting conditions and accurately rectified as Figures 6.12(c) and 6.12(d) represent. The image size and color resolution are the same to the Balloon images. However, the disparity estimation using toy images is difficult because the images have a lot of untextured regions with similar color and intensity values in foregrounds and backgrounds, and such homogeneous regions easily cause ambiguous matching problems.
The proposed algorithm yields excellent dense disparity maps as Figure 6.13 shows. Figures 6.13(a) and 6.13(c) are the left-to-right disparity maps for the Balloon and Car Toy images where brighter values represent large disparity vectors. Darker values in the right-to-left disparity maps shown in Figure 6.13(b) and 6.13(d) represent small disparities. While the energy function is iteratively solved by PDEs, the dense disparity vectors are locally es-



(a) (b)

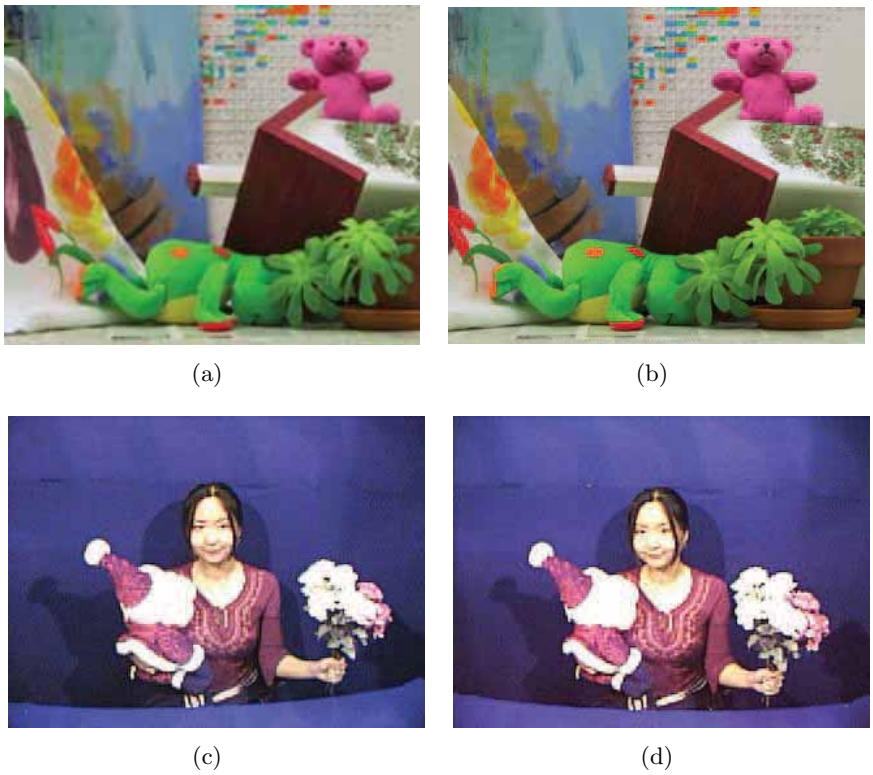
(c) (d)

Figure 6.13. Results of anisotropic disparity estimation - (a) the left-to-right disparity map of Balloon images, (b) the right-to-left disparity map of Balloon images, (c) the left-to-right disparity map of Car Toy images, (d) the right-to-left disparity map of Car Toy images

timated and globally regularized with anisotropic diffusion. An edge-stopping function that strongly weights for a large gradient provides an excellent localization matching and structure preserving regularization. Detected occlusions can be diffusively filled in the regularization process.

We compare the proposed method to some well-known disparity estimation methods by using large-baseline matching tests. A large baseline disparity estimation is a challenging task in computer vision, since it is usually difficult to establish correspondences across images. If the baseline is large, epipolar geometry must be calculated before disparity matching. Two large-baseline stereo sequence data sets shown in Figure 6.14 are used for the evaluation. The first test image set with a green doll, i.e. Teddy image, shown in Figures 6.14(a) and 6.14(b) has 450×375 resolution, and the maximum horizontal difference[54] is 80 pixels. The 2nd image set with a woman, i.e. Kate image, shown in Figures 6.14(c) and 6.14(d) has 320×240 pixels. The Kate images are captured by a stereo camera with a toed-in setup and not rectified.

54) This images are captured by a stereo camera with parallel camera configuration that aligns the epipolar lines parallel to the horizontal scanlines.



(a) (b)

(c) (d)

Figure 6.14. Large baseline test images - (a) the left Teddy image, (b) the right Teddy image, (c) the left Kate image, (d) the right Kate image

Figure 6.15 represents the results of well-known SSD, graph-cut methods and the proposed algorithm. SSD is a pixel-based matching algorithm using a sliding window to calculate as error criteria the sum of squared differences. Figures 6.15(a) and 6.15(b) depicts that SSD produces a very large matching error since a fixed window increases the likelihood of a mismatch per pixel. A large baseline matching with an unknown epipolar geometry causes serious errors even for an optimization method as Figure 6.15(d) exemplifies. The graph-cut method is one of the most reliable optimization algorithms that can obtain links-and-cuts of multiple possible values by repeatedly minimizing an energy function involving only binary variables. Although the graph-cut method requires high computational power, the ambiguity of unparallel wide baseline matching causes a lot of wrong links. When the epipolar geometry is known, the graph-cut method generates a good result for very large-baseline images shown in Figure 6.15(c). However, the method needs post-processing to occlusion holes filling. On the contrary, the proposed method represents the most reliable and detailed disparity maps as Figures 6.15(e) and 6.15(f) represent. The performance comparison proves the robustness of the proposed algorithm.

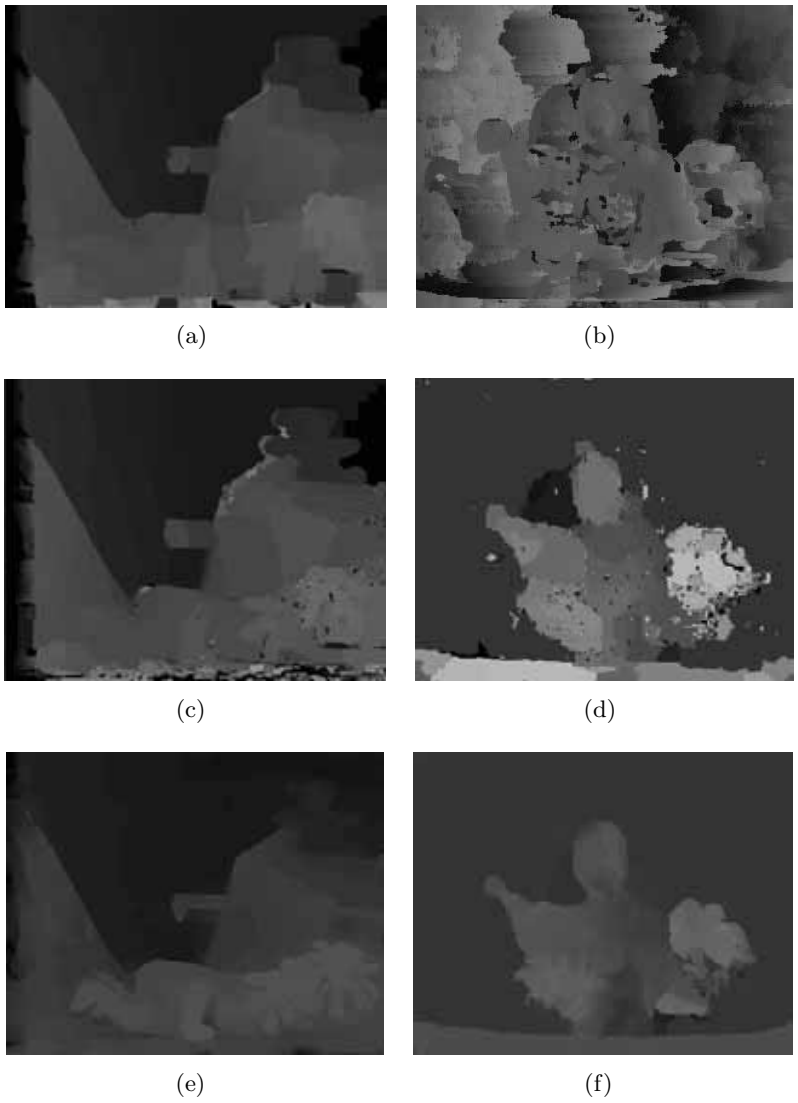
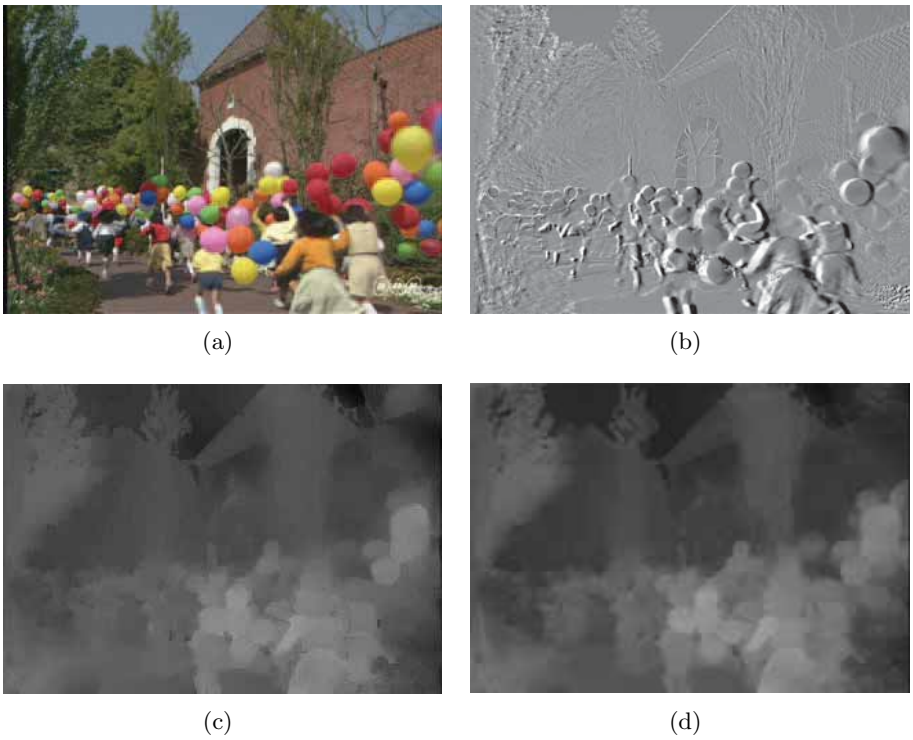


Figure 6.15. Comparison of large baseline dense disparity estimations - (a) SSD result for Teddy images, (b) SSD result for Kate images, (c) graph-cut result for Teddy images, (d) graph-cut result for Kate images, (e) anisotropic disparity estimation result for Teddy images, (f) anisotropic disparity estimation for Kate images

Anisotropic disparity estimation with perceptual maximum variation modeling : in an image region with strong texture gradients, anisotropic disparity estimation may fall into local minima and cause an over-splitting problem. The isotropic property of the Gaussian scale-space is employed in the proposed method to smooth local edges in a texture region, and a robust weighting modifies the scale-space to anisotropic diffusion. This method is more efficient than the pure anisotropic diffusion method since exact gradients are very sensitive to noises. However, if some visually important edges are weak in the luminance channel,



(a) (b)

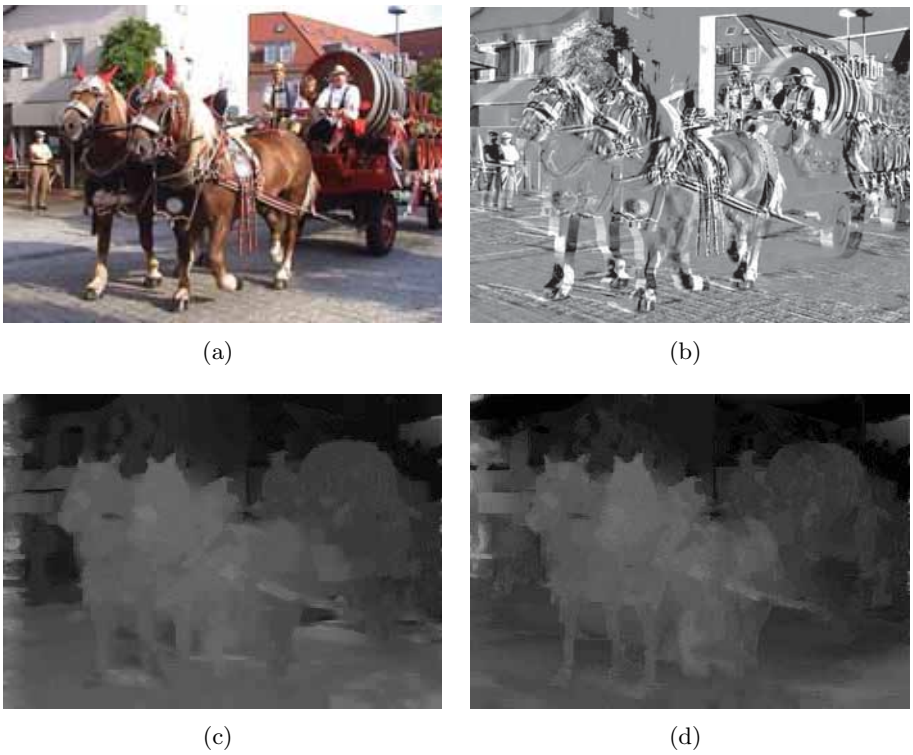
(c) (d)

Figure 6.16. Anisotropic disparity estimation with perceptual maximum variations for Balloon image - (a) The left Balloon image, (b) difference between left and right Balloon images, (c) anisotropic disparity estimation of the Balloon images, i.e. the method of Chapter 6.1, (d) anisotropic disparity estimation with perceptual maximum variations of Balloon images, i.e. the method of Chapter 6.2

isotropic over-diffusion and back-diffusion make the edges ambiguous since the weighting process is based on the brightness variations. The proposed method in Chapter 6.2 solves the over-diffusion and back-diffusion problem by modeling perceptual maximum variations using a least squares optimization applying PCA. Thus, anisotropic disparity estimation can be accurately performed in perceptual continuation coherency of a scene.

In Figure 6.16, the improvement of the previous method is exemplified for the Balloon images, i.e. shown in Figures 6.12(a) and 6.12(b). Edge-preserving disparity estimation can be robustly achieved in image regions containing important structures with strong and weak brightness variations because the maximum variation modeling improves the convergences into the global minima. Figure 6.17 shows another result for wagon stereo images obtained by a similar way. The resolution of wagon images is 720×576.

Figure 6.18 shows the absolute disparity error compared to the ground-truth of Tsukuba and Cones images of [35] using the percentage of error pixels. We ignore 18 pixels of the border for the Tsukuba image when computing the statistics due to the absence of data in the



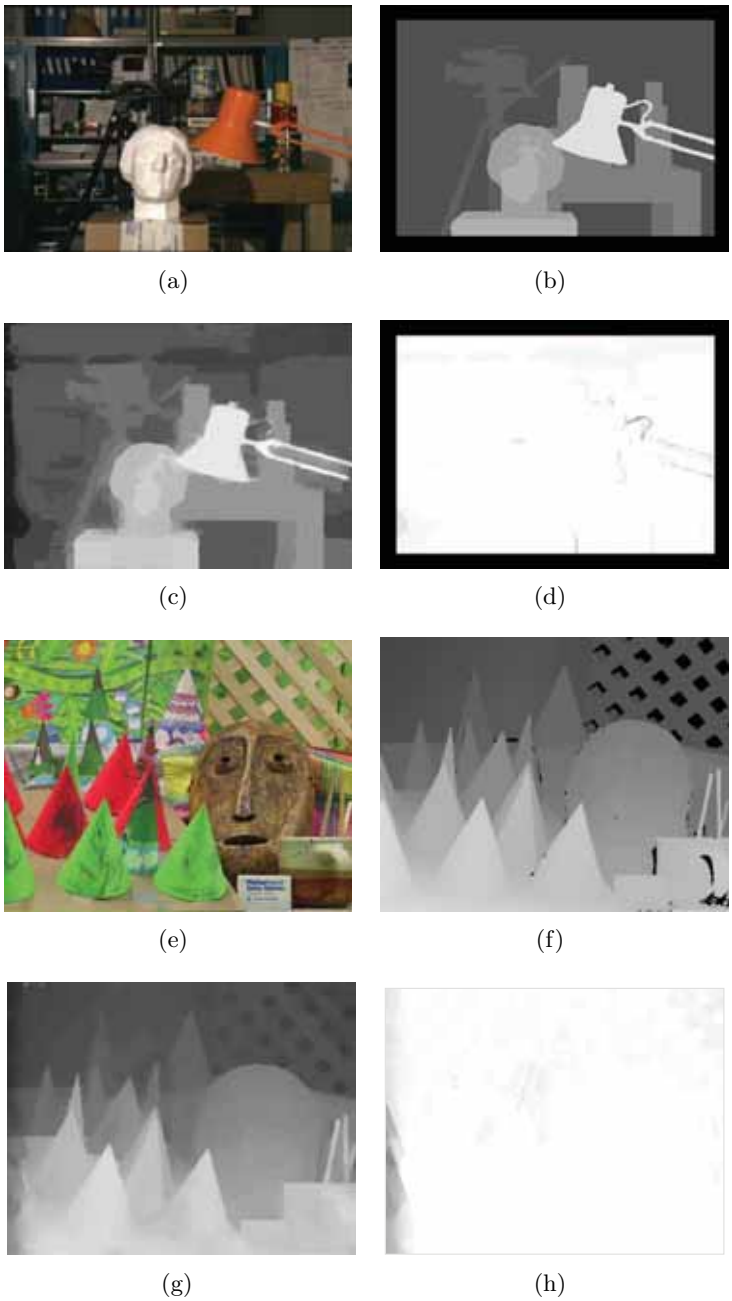
(a) (b)

(c) (d)

Figure 6.17. Anisotropic disparity estimation with perceptual maximum variations for the Wagon images - (a) the left Wagon image, (b) difference between left and right Wagon images, (c) anisotropic disparity estimation of wagon images, (d) anisotropic disparity estimation with perceptual maximum variations of Wagon images

ground-truth. Figures 6.18(b) and 6.18(f) represent the ground-truth of Tsukuba and Cones images. Figures 6.18(c) and 6.18(g) show the robust performance of the proposed method for discontinuities and homogeneous regions. Figures 6.18(d) and 6.18(h) are the error images with absolute difference over 1.0 between the ground-truth and estimates. Table 6.1 depicts the performance comparison between well-known reliable optimization methods, e.g. layered stereo [157] and graph-cut [156], requiring much heavier computation costs than the proposed method. The errors are calculated using alpha maps which decide the evaluating regions. Only the white regions of the alpha maps are used for calculating the absolute difference. The proposed method achieves a high accuracy performance in texture and discontinuity over other techniques.

Robust optical flow on large motion fields : Optical flow estimation on large motion fields is fundamentally difficult since pixels between two discrete pixels increase the possibility converging into local minima. A new reliable hybrid recursive method for optical flow estimation was proposed in Chapter 6.3. The method efficiently combines the advantage of discrete

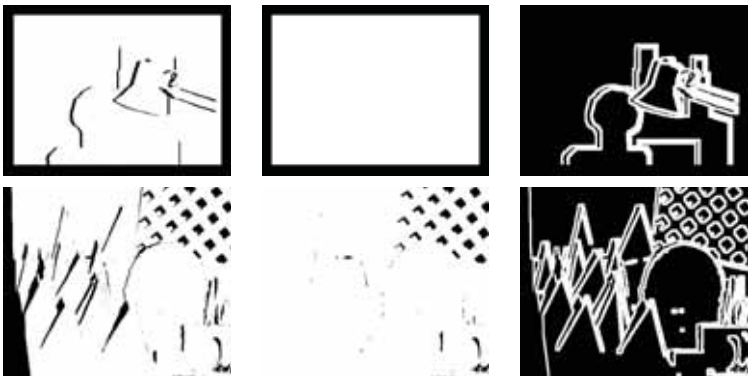


(a) (b) (c) (d) (e) (f) (g) (h)

Figure 6.18. Ground-truth evaluation of anisotropic disparity estimation with perceptual maximum variations - (a) the left Tsukuba image, (b) ground-truth for Tsukuba images, (c) simulation result for Tsukuba images, (d) error pixels (i.e., darker pixels have larger error) that have absolute difference over 1.0 for Tsukuba result, (e) the left Cones image, (f) ground-truth for Corns images, (g) simulation result for Corns images, (h) error pixels over 1.0 for Corns result

Table 6.1. in the white regions of alpha maps; *noocc.*: non-occluded regions; *all*: all regions; *disc.*: areas of depth discontinuities

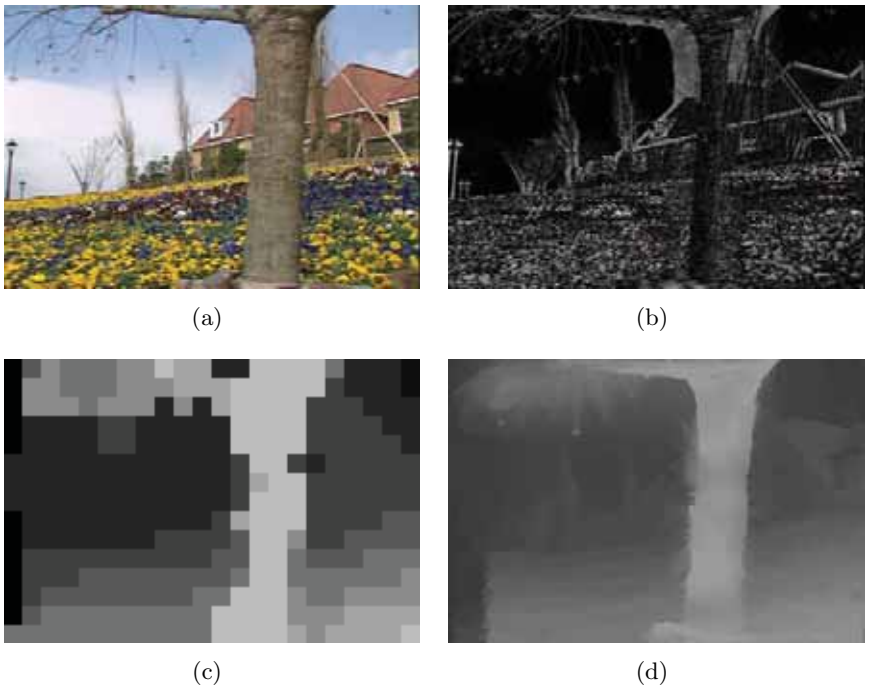
Method	Tsukuba			Cones		
	noocc.	all	disc.	noocc.	all	disc.
SSD+min-filter	5.23	7.07	24.1	10.6	19.8	26.3
layered stereo	1.3	1.57	6.92	6.59	14.7	14.4
graph-cut	1.94	4.12	9.39	7.7	18.2	15.3
proposed method	1.06	1.21	4.47	3.41	8.55	5.02



noocc. *all* *disc.*

motion estimation and optical flow estimation in a recursive block-to-pixel estimation scheme. Integrated local and global approaches using the robust statistic of anisotropic diffusion removes outliers from the estimated motion field. We separately describe the process with two frameworks i.e. an incremental updating framework and a robust energy minimization framework. With robust error norms of anisotropic diffusion, the formulation usually leads to non-convex optimization problems. Thus, the solution has many local minima, and convergence to the global minima is not guaranteed. Our hybrid recursive energy-based method employs a hierarchical block-to-pixel estimation concept to prevent this problem.

First, the performance is evaluated for a very large motion of a dynamic scene. The Flower Garden sequence of size 320×240 pixels and 150 frames is shown in Figure 6.19. It consists of a tree in the foreground, with a background increasingly vanishing away by the panning of the camera from left to right. The sequence has a very large motion in excess of 6 pixels/frame and also has an abundance of fine moving textures. Discrete motion is estimated using 8 blocks as Figure 6.19(c), and optical flow is incrementally updated based on the block motion compensated image. The results in Figure 6.19(d) show the dense motion field between the 1st frame and 6th frame using the magnitude. It proves the robustness and efficiency of our method for a very large motion. Comparative results per frame for the whole sequence are shown with some traditional methods in Table 6.2 using the result in [170]. Our method by far outperforms the above methods. The performance of the proposed method is next evaluated



(a) (b)

(c) (d)

Figure 6.19. Optical flow estimation for discrete frames of Flower Garden image sequence - (a) the 3rd frame image, (b) the 7th frame image, (c) difference image between the 3rd and 7th frames, (d) block initial vector (8×8)) of the proposed method, (e) final motion field of the proposed method, (f) Lukas-Kanade method

Table 6.2. Performance comparison for discrete motion

Algorithms	Mean error	Standard deviation	Density (%)
Horn and Schunck (modified)	16.09	13.64	100
Lucas and Kanade	15.75	13.3	100
Nagel	17.11	14.55	100
Anandan	12.25	10.03	100
Odobez and Bouthemy	13.21	11.12	97.77
Proposed method	4.41	3.13	100

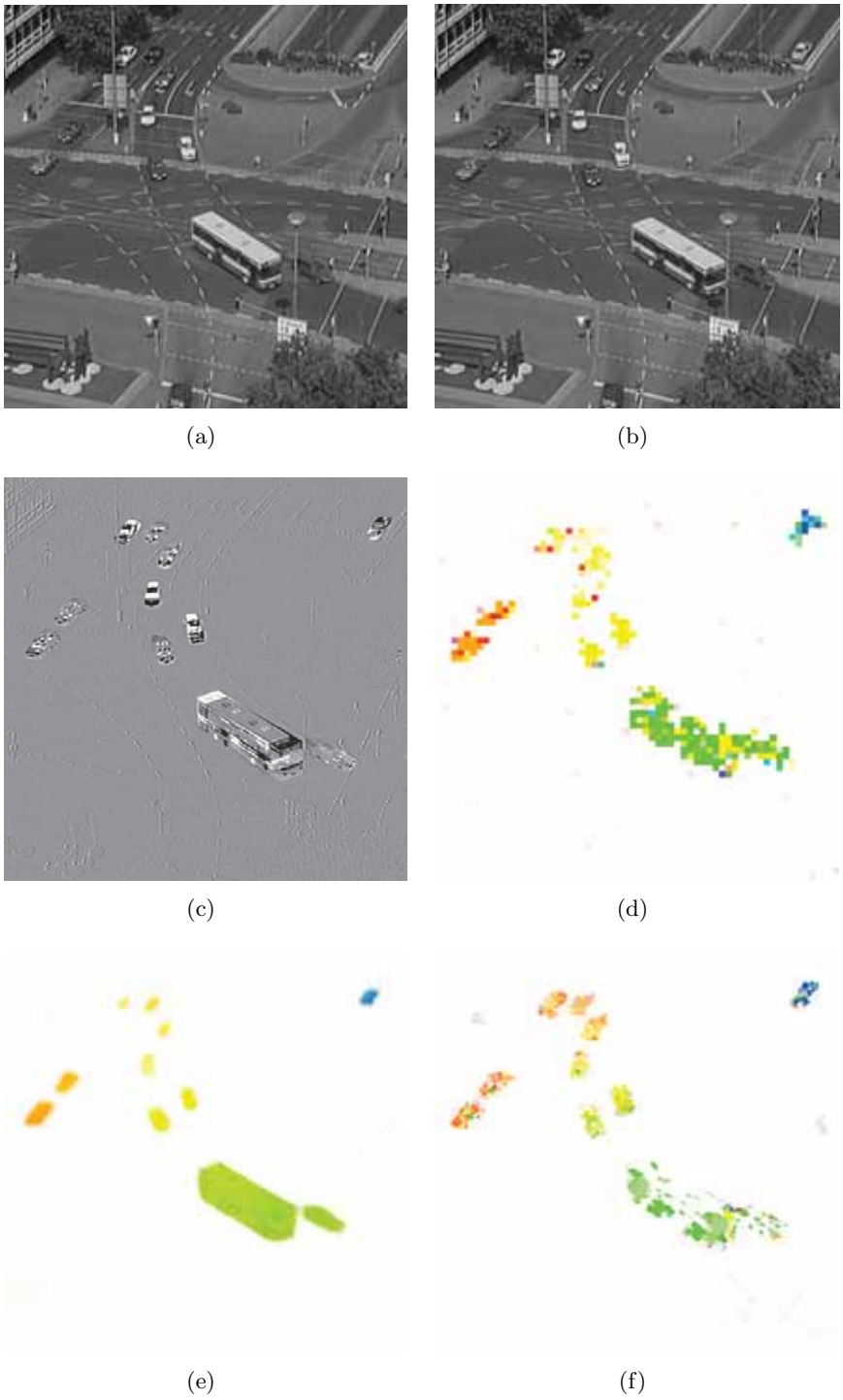
with two natural image sequences captured from static cameras, i.e. the Ettlinger Tor traffic sequence and the Hamburg taxi sequence [172]. Figure 6.20 shows the 3rd frame and the difference with the 7th frame of Ettlinger Tor traffic sequence which has the size of 512×512 pixels. Figures 6.20(d) and 6.20(e) respectively represent the results of block recursion and pixel recursion using color encoding of flow vectors shown in Figure 6.10. Both the block initial vector and final dense motion field are adjusted to be more visible with the same scale. The result shows the robustness of the estimated vectors and their good localization into the discontinuity. The discrete motion estimation problem is exemplified in Figure 6.20(f). The

Lukas-Kanade optical flow method cannot estimate fast object motions and results in low density estimates. Figure 6.21 represents the Hamburg taxi sequence which has the size of 256×190 pixels. The result also represents the excellent performance of our method. The sharp motion boundaries are well preserved.

Spatio-temporal sceneflow estimation : Stereo matching and tracking estimates the 3-D displacement field for points in the 3-D world using multi-camera video data. Such methods can be cooperated with motion estimation. Chapter 6 has described to a novel formulation to combine dense disparity estimation and dense optical flow that provides reliable results using only two cameras by fusing binocular sequences into coherent 3-D points. The cameras are self-calibrated by a reliable feature tracking method and the view geometry is also known. The method chooses the minimum error sampling between optical flow and disparity with a robust estimation framework. Outliers in estimates are globally restricted by a local appearance region method shown in Chapter 6.1 and robustly regularized by spatio-temporally anisotropic diffusion introduced in Chapter 6.3. To avoid the local minima problem inherent in solving the PDEs, the scale-space method is used within the solution. The spatio-temporally sceneflow is fitted to the bundle of rays through a global minimization of the reprojection error as Chapter 6.4 described. Two videos, Pillars and Under Sea shown in Figure 6.22 are captured using a monocular camera motion, and the camera parameters and the sparse scene structure is relatively recovered by a state-of-art SfM method. The relations between the views, i.e. known as epipolar geometry, can be used to warp all original viewpoints in the scene into a slightly different viewpoint near the real camera path. When the dense 3-D sceneflow is estimated in the synthesized viewpoints using a warping, the 3-D information is 3-D cloud points, i.e. a large set of 3-D sampling positions, which can be efficiently used for image-based modeling and rendering.

Figures 6.22(a) and 6.22(b) are the 1st and 2nd frames of the Pillars image. Dense optical flow between two frames is estimated as Figure 6.22(c) shows. The color denotes the rotation of flow vectors and the brightness is the velocity as Figure 6.10 represents. Figure 6.22(d) is the dense depth map of Pillars images.

For the Under Sea images, the 6th and 12th frames shown in Figures 6.22(e) and 6.22(f) are used to estimate the sceneflow. Figures 6.22(g) and 6.22(h) respectively represent the spatial and temporal component of the sceneflow.

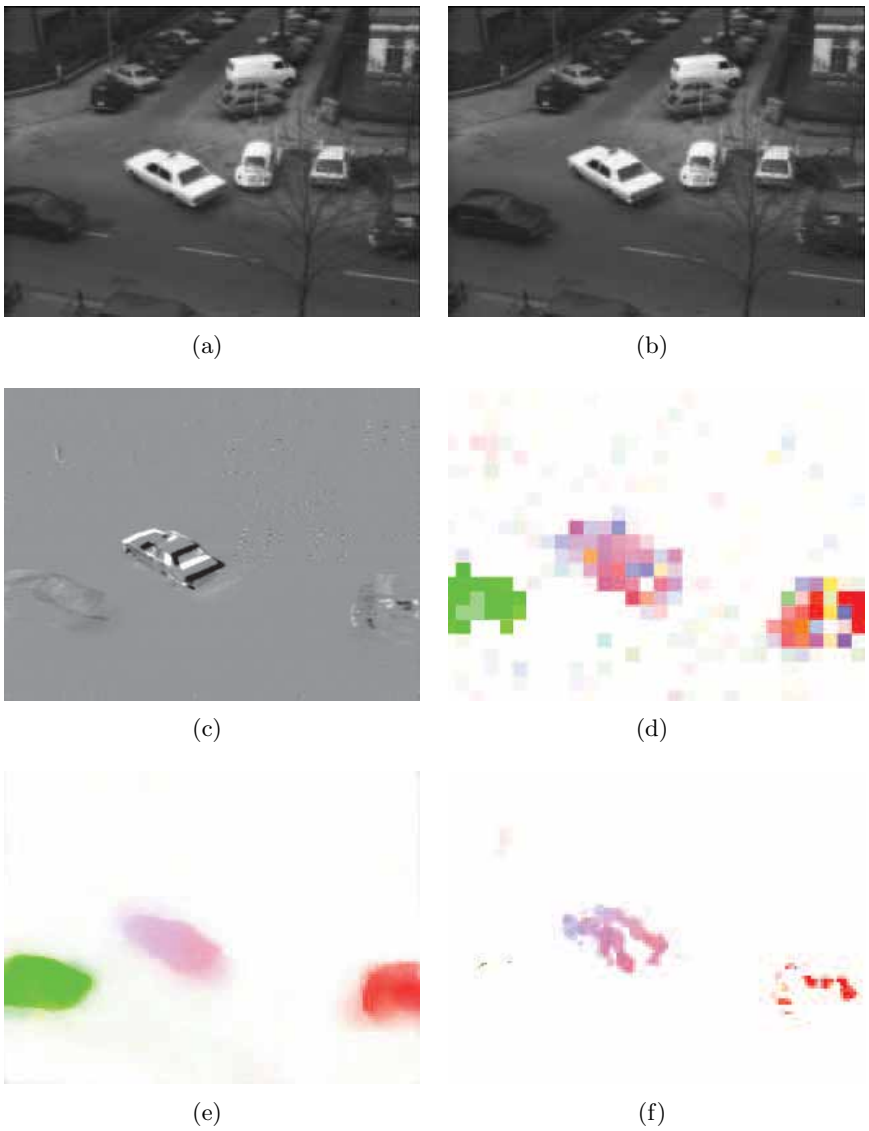


(a) (b)

(c) (d)

(e) (f)

Figure 6.20. Optical flow estimation for discrete frames of Ettingertor image sequence - (a) the 3rd frame image, (b) the 7th frame image, (c) difference image between the 3rd and 7th frames, (d) block initial vector (8×8)) of the proposed method, (e) final motion field of the proposed method, (f) Lukas-Kanade method

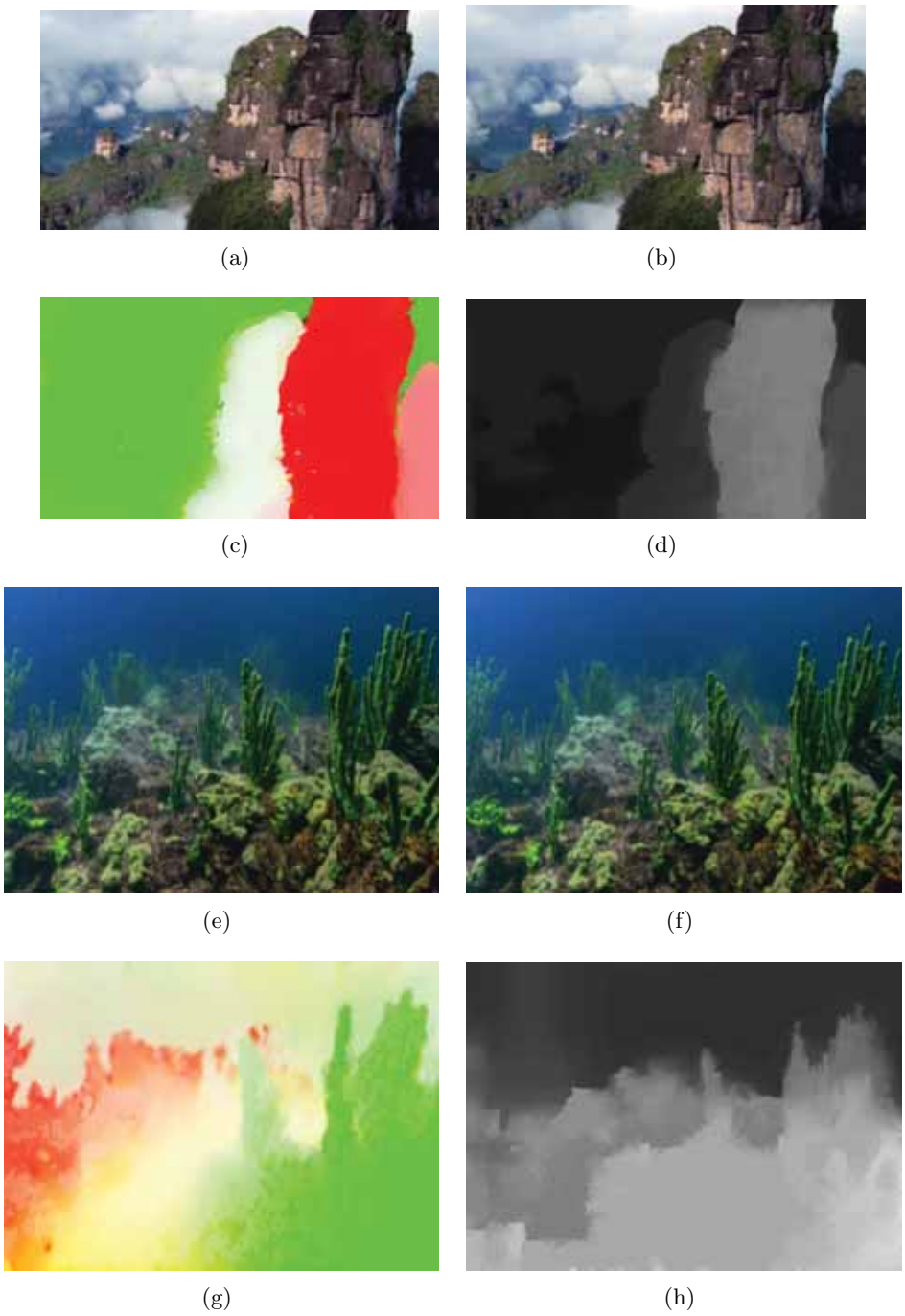


(a) (b)

(c) (d)

(e) (f)

Figure 6.21. Optical flow estimation for discrete frames of Taxi image sequence - (a) the 3rd frame image, (b) the 7th frame image, (c) difference image between the 3rd and 7th frames, (d) block initial vector (8×8)) of the proposed method, (e) final motion field of the proposed method, (f) Lukas-Kanade method



(a) (b) (c) (d) (e) (f) (g) (h)

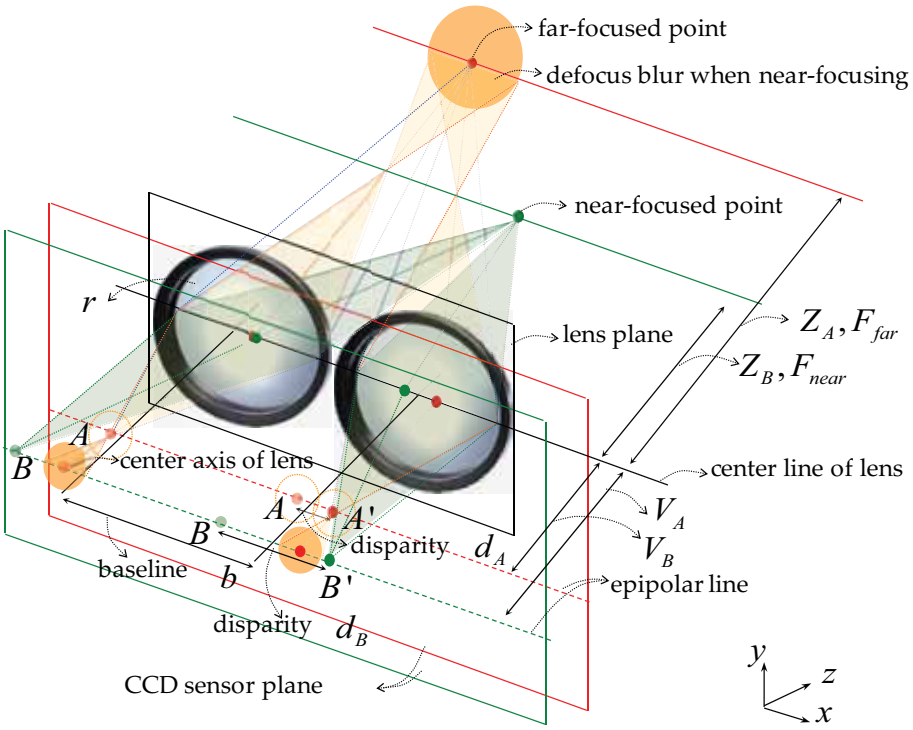
Figure 6.22. Spatio-temporal scene flow estimation result - (a) the 1st left frame of Pillars image, (b) the 2th left frame of Pillars image, (c) optical flow motion between the 1st and 2nd left frames of Pillars image, (d) disparity between the left and right Pillars images for the 1st frame, (e) the 6th left frame Under Sea image (f) the 12th left frame of Under Sea image, (g) optical flow motion between the 6th and 12th left frames of Under Sea image, (h) disparity between the left and right Pillars images for the 6th frame

Chapter 7

Confocal Stereo with Pinhole Image Recovery

7.1 Problem of Pinhole Camera Assumption

Recent years have seen a lot of advances in the problem to reconstruct a complex 3-D scene from a series of multiple images. Most algorithms choose a human visual feature perceiving depth e.g. disparity and depth of field (DoF). By establishing the correspondences in a series, objects are distinguished by depth information related to the respective position. However, the ill-posed correspondence problem cannot be perfectly solved by only one series.
For example, disparity estimation is difficult to solve the ambiguity problem in local image structures due to image noise, unbalanced brightness, similar texture and occlusion, etc. To achieve a more reliable estimation performance, local appearance matching [72–74] with boundary constraints between features, edges and disparity discontinuity etc. is employed. However, the performance is not satisfactory due to the error localization problem. The latest researches incorporate isotropic and anisotropic regularization terms which attempt to filter off the mislocalized error. Although these methods are efficient to preserve local structure, semantic information is additionally needed to combine divided regions in a real object. Practical camera systems use a *real aperture camera model* which yields focus-related blurring (i.e. DoF). Depth from focus (or defocus) methods [173, 174] exploits the variation of the blurring in a number of images captured at different focus settings. When the camera is focused on an object at a depth, the other objects in a different depth are blurred. The relative blurring between the defocused images can be utilized as a stereo cue. In this chapter, a novel anisotropic disparity estimation embedding the defocus cue, i.e. confocal constraint, is proposed for a real-aperture 3-D camera system shown in Figure 7.1.



(a)

Figure 7.1. Real-aperture 3-D camera system

7.2 Real-Aperture Stereo Camera Analysis

Blurring of the defocus by optic can be approximated by linear convolution between the focused image and the blurring function i.e. known as *point spread function* (PSF). Blurring between near- and far-focus images can be estimated by the second central moments of the blur circle because the PSF is a *circularly symmetrical function*. Subbarao and Surya [173] proposed the S-transform of Laplacian as a focus operator. A defocus function acts as a low-pass filtering, e.g. 2-D Gaussian and a focus operator which performs the inverse. The difference of the standard deviations, i.e. the spread parameter between near- and far-focused images can be mapped to the respective depths. However, the operator is suitable only for equifocal surfaces since the operator is isotropic. In [174], Favaro et al. employs an anisotropic diffusion to solve the problem. If a scene is highly textured, the method is sufficient to estimate a reliable depth. However, regions with weak textures are still ambiguous to distinguish from blurred regions.

For the solution, fusions with stereo [175–177] were proposed. In [175], a probabilistic model of focus and stereo merges the depths by weighted averaging the local variances that are estimated by Cramer-Rao inequality in the unbiased estimator. However, this method cannot

guarantee the accuracy of the estimated variances. A Markov random field method [176] integrates the depths by smoothness prior in an energy functional which is minimized by simulated annealing. Another method [177] employs graph-cut for a focus measure. However, these methods do not consider the local features to avoid the convergence to local minima. In Chapters 5.4 and 6.1, the multiple-resolutions of the scale-space provide the best trade-off between the detection and the localization performance of the features. The range of disparity matching is propagated by isotropic gradient weights following the streamline of the scale-space and localized in the spatially invariant sparse Laplacian kernel by the Poisson solver. The dense disparity estimation is locally performed by a scheme of local appearance matching. Anisotropic diffusion globally regularizes the disparity map by suppressing the length of the isotropic propagations to the orthogonal direction of edges. The method gave us good dense disparity maps preserving the important local structures. However, it cannot combine divided regions in the main object.
In this chapter, we try to combine the regional structure in a focal object in defocused images. A focus image offers semantic information such as the sharp boundary of a focusing object which is easily distinguished by the defocus blur. The defocus cue serves a constraint to localize the disparity estimation in a confocal boundary. However, the variances in PSF are isotropic and the propagation near the edges of unequifocal surfaces may not be accurate. We regularize the propagation in the dense disparity map by anisotropic filtering.

7.3 Confocal Disparity Estimation

If a focused point belongs to an object surface, the diameter σ of blurring in a series of defocus images is given by a lens law [176]

$$\sigma_n = \kappa r_n V_n \left(1/F_n - 1/Z_n - 1/V_n\right) \tag{7.1}$$

n is the number of focus images, r is the lens aperture, F is the focal length, $V$is the sensor plane-to-lens distance and Z is the distance between the object surface and the lens, i.e. the object distance. We use a parameter κ from a pre-calibration to fit the focal depth plane into the depth from disparity.
Since conventional stereo systems use a pin-hole camera model, we assume that an ideal focus (i.e. $F_n = f$) enables the infinite depth of field. The ideal focuses F_{far} and F_{near} on the surfaces which are located at respective depths Z_a and Z_b accompany defocus blurring as Figure 7.1 illustrates. The confocal disparity can be defined by an epipolar constraint.

$$\mathbf{d} = bf/z \tag{7.2}$$

z is the associated depth to the disparity $\mathbf{d}$ and b is the baseline. In the pin-hole model, the f in Equation 7.3 should be the same with the V_n in Equation 7.3. The relationship between the disparity and the focus of a stereo image pair is achieved by the confocal condition which should estimate a same depth (i.e. $z \approx Z$).

$$\mathbf{d}_n = b\left(V_n/F_n - \sigma_n/\kappa r_n - 1\right) \tag{7.3}$$

Since all parameters except σ_n are given by camera setting, if the focal length of the stereo camera is just changed, the stereo matching can be localized by the confocal constraint within the range of d_n which is estimated by defocus blurring.

7.4 Confocal Constraint of Defocus

For an object surface $\mathbf{S}$ with a function $s : \mathbb{R}^2 \rightarrow [0, \infty]$ which assigns a depth value to each pixel coordinate, the irradiance $I : \Omega \subset \mathbb{R}^2 \rightarrow [0, \infty]$ with another function $R : \mathbb{R}^2 \rightarrow [0, \infty]$ on the surface is observed as the radius σ of the defocus blur. The defocus image $J(x)$ which is dependent upon the camera optics is defined as

$$J(\mathrm{x}) = \int h\left(I(\mathrm{x}), \sigma(s, F_n)\right) R(\mathrm{x})\mathbf{d}\mathrm{x} \tag{7.4}$$

where $h : \Omega \times \mathbb{R}^2 \rightarrow [0, \infty]$ is the PSF which is defined as the impulse response by Green's function [174]. The radius σ is related on the surface s and the focus parameter F. Let a focus image be the irradiance map I_0 of the focal surface $s(F_0)$ of a initial focus F_0. The depth $Z_n = s(F_0) \sim s'(F_n)$ is estimated by observing the variances of a defocus image I_n is taken with a different focus F_n at a time t. The energy function is

$$E(\sigma) = \int_{\Omega} \left(h\left(I_0(\mathrm{x}+\sigma(F_0), 0) - h\left(I_n(\mathrm{x}+\sigma(F_n), t)\right)^2 d\mathrm{x}\right.\right. \tag{7.5}$$

If we approximate the PSF by a generalized 2-D Gaussian as

$$\mathcal{G}_{\mathrm{V}_n}(\mathrm{x}) = \left(\sqrt{2\pi\left|\mathrm{V}_n\right|^{1/2}}\right)^{-1} e^{-(\mathrm{x}_n^T \mathrm{V}_n^{-1} \mathrm{x}_n/2)} \tag{7.6}$$

where V is the symmetric 2×2 variance matrix with the determinant $|\mathrm{V}|$, the eigenvectors determine the principal axes η and ξ from the partial derivatives of a Gaussian and the eigenvalues (μ_+, μ_-) determine the scalar variance along these axes.

$$\mathrm{V}(r) = \mu_+ \eta\eta^T + \mu_- \xi\xi^T \text{ and } \eta \perp \xi \tag{7.7}$$

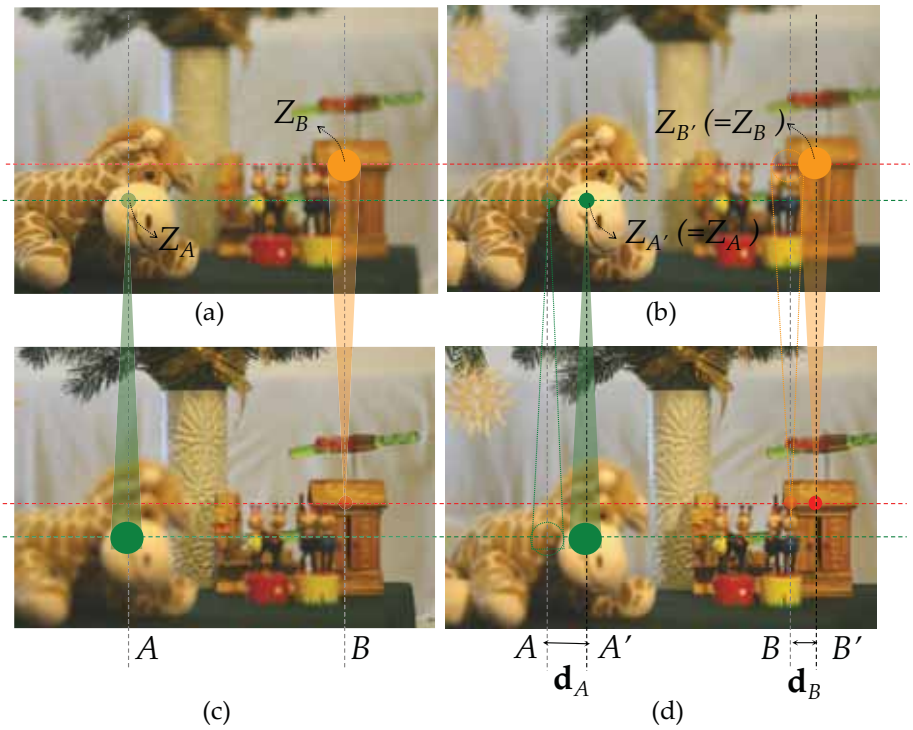


Figure 7.2. Confocal disparity estimation - (a) definition of the parameters in the system, (b) the left viewpoint-image in a near-focusing setup ($f = F_{near}$), (c) the right viewpoint-image at F_{near} , (d) the left viewpoint-image in a far-focusing setup ($f = F_{far}$), (e) the right viewpoint-image at F_{far}

where r is the displacement in a polar coordinate $r(\theta) = \sqrt{x^2 + y^2}$. This derives an ellipsoid which has a combination of value-weighted orthogonal orientations $\mu_+\theta_+\theta_+^T + \mu_-\theta_-\theta_-^T$ with $\theta_+ \perp \theta_-$ centered at a point. This is more general to represent a partial homogeneous region in defocus images by convolution $I_n(\mathrm{x}) = I_0(\mathrm{x}) \otimes \mathcal{G}_{V_n}(\mathrm{x})$. In the n defocus images, the smoothly varying structure is defined by a positive definite tensor $\mathbf{D}$ that denotes a gradient flux.

$$\mathrm{x}(r|t, \mathbf{D}) = \left(\sqrt{2\pi^n |2t\mathbf{D}|}\right)^{-1} e^{-r^T \mathbf{D}^{-1} r/4t} \tag{7.8}$$

The variances in the PSF is observed by the displacement of tensors at each image point x. It represents a difference in propagations to the direction $(\theta_+\theta_-)$. If the space of the functions is continuous with the partial derivatives in $\mathbb{R}^2$, Equation 7.4 can be minimized in the PDEs by a harmonic solution $I(\mathrm{x}) \rightarrow J : \mathbb{R}^2 \times [0, \infty] \rightarrow \mathbb{R}$

$$\begin{cases} J(I(\mathrm{x}), 0) = \sigma & {}^{\forall}\mathrm{x} \in \Omega \\ J(I(\mathrm{x}), t) = div\,(\mathbf{D}(\mathrm{x})\nabla(I_n(\mathrm{x}), t)) & t > 0 \end{cases} \tag{7.9}$$

where $J(\mathrm{x}) = I(\mathrm{x}) + \varepsilon$ where ε denotes a correction gradient symbol related to the amount of blurring and the divergence term *div* is implemented by a sparse Laplacian kernel in the Poisson solution $\mathbf{D}(\mathrm{x}) \nabla (I_n(\mathrm{x}), t) \cdot n = 4\pi\rho$ with $\rho = 0$.The unit vector n is orthogonal to $\partial\Omega$.

7.5 Anisotropic Disparity Estimation with Confocal Constraint and Recovery of Pinhole Image

With the confocal constraint the d_n from Equation 7.3 for a focusing stereo object, we obtain the maximum window size $\mathbb{W}$ for the disparity matching. Figure 7.2 illustrates the confocal constraint for disparity matching.

$$\max_{\mathbf{D}} \mathcal{W}_{l,n}(\mathrm{x}) \leq \sum\nolimits_{i=0}^{d_n} I_{r,n}(\mathrm{x} + i) \tag{7.10}$$

where the subscripts l and r respectively denote the left and right image. The sampling of pixels in the focal plane in defocus stereo images recovers a non-blurring image for all objects i.e. pinhole image. The confocal disparity is achieved by iteratively updating the estimated disparities in the window $\mathbb{W}$ as

$$E_\Omega(d) = \int_\Omega (I_l(\mathrm{x}) - I_r(\mathrm{x} + d_{l\to r}(\mathrm{x} + \sigma) \in \mathcal{W})^2 d\mathrm{x} + \lambda \int_\Omega e_\sigma(d) d\mathrm{x} \tag{7.11}$$

The subscripts denote the matching direction, e.g.$l \to r$ is left-to-right direction, and e_σ is a global regularization term which minimizes the matching error with Lagrange multiplier λ. The range of matching is restricted within the confocal depth range. The window using isotropic PSFs is suitable to avoid the splitting problem in a homogeneous equifocal surface. However, the edges of unequifocal surfaces may not be exactly localized. Hence, we globally regularize the dense disparity map by an anisotropic diffusion.

$$e_\sigma = g(\|\mathbf{D}_l(\mathrm{x})\|) \nabla d_{l\to r}(\mathrm{x}, t) \tag{7.12}$$

$g(\nabla)$ is an anisotropic diffusion weight which suppresses the propagation of only the edge direction θ_-.

$$g(\mu_+, \mu_-) = (e^{-(\mu_+^2/\gamma^2)}, \mu_-) \tag{7.13}$$

where a positive constant γ controls the level of contrast of edges affecting the regularization process that was described in Equation 6.7. The graph of the g-function was shown in Figure 6.4. It enhances the discontinuities in θ_- by the flux function $\psi(\nabla) = g(\mu_+, \mu_-) \nabla$. The detail numerical solution of energy functional in Equation 7.5 is described in Chapter

(a) (b)

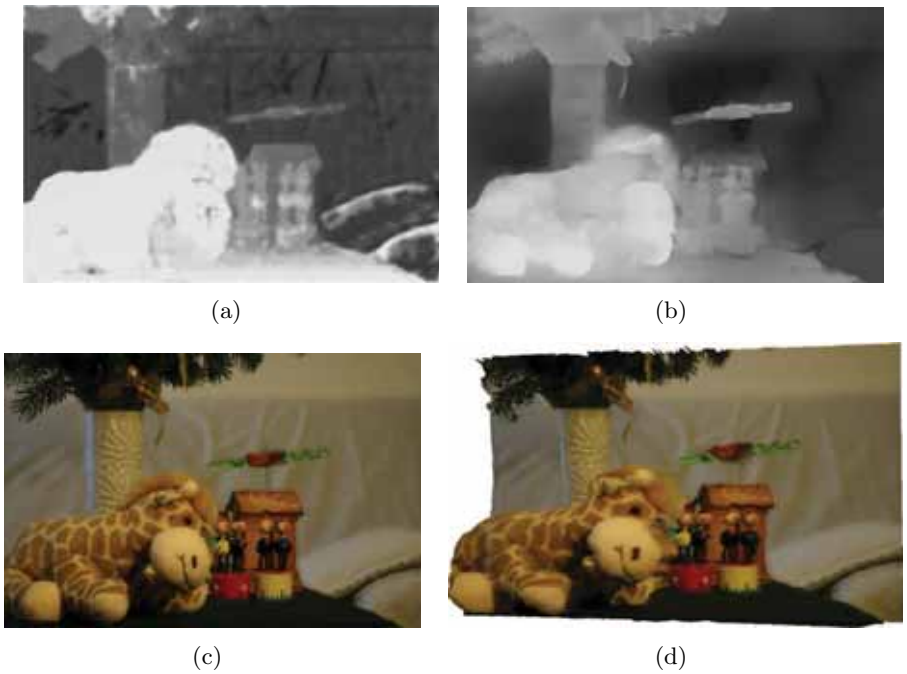
(c) (d)

Figure 7.3. Multi-focusing stereo images - (a) the near-focusing left image, (b) the near-focusing right image, (c) the far-focusing left image, (d) the far-focusing right image

6.1.

7.6 Experimental Results

A single dense depth estimation using stereo or defocus cannot produce a reliable result due to the ambiguity problem. In this chapter, a novel anisotropic disparity estimation embedding a stereo confocal constraint is proposed for real-aperture stereo camera systems. If the focal length of a real-aperture stereo camera is just changed, the depth range is localized in a focused object which can be discriminated from defocused blurring. The focal depth plane is estimated by the displacement of tensors which are derived from generalized 2-D Gaussian, since the point spread functions (PSF) in defocused blurring can be approximated by a shift-invariant Gaussian function. The isotropic propagation in blurring over invariance is localized by a sparse Laplacian kernel in the Poisson solution. The matching of real-aperture stereo images is performed by observing the focal consistency. However, the isotropic propagation cannot exactly hold a non-parallel surface to the lens plane i.e. a unequifocal surface. An anisotropic regularization term is employed to suppress the isotropic propagation near the non-parallel surface boundary. Our method achieves an accurate dense disparity map by sampling the disparities in focal points from multiple defocus stereo images. The pixels in



(a) (b)

(c) (d)

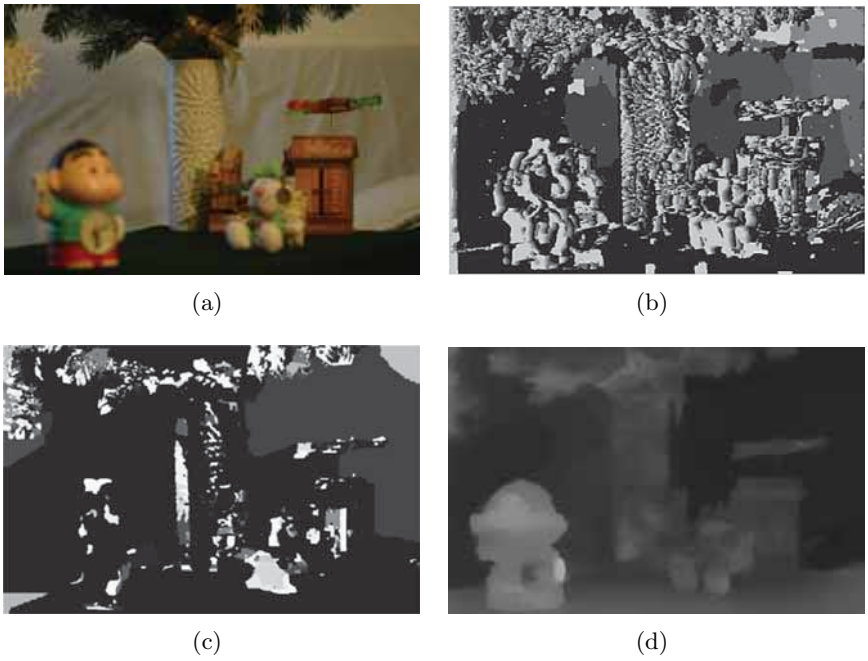
Figure 7.4. Confocal disparity estimation and pin-hole image recovery - (a) focal consistency window from 10 defocus images for the left view, (b) dense disparity map by the proposed method, (c) recovered pinhole image, (d) 3-D modeling result using the estimated disparity and the recovered pinhole image

(a) (b)

Figure 7.5. Comparison of anisotropic disparity estimations with and without confocal constraint - (a) anisotropic disparity estimation result without confocal constraint, i.e. method in Chapter 6.1, (b) confocal anisotropic disparity estimation

focal points are utilized to recover the pinhole image, i.e. an ideally focused image for all different depths.

Figure 7.3 shows a pair of stereo images which are respectively focusing on the most far and nearest object surfaces in 10 defocus stereo images. The image resolution is 720×480. First,



(a) (b)

(c) (d)

Figure 7.6. Disparity estimation for stereo images with out-focusing objects - (a) the left image, (b) disparity map by Markov Random Fields method, (c) disparity map by graph-cut method, (d) disparity map by the proposed method

a focal consistency window from defocus images is estimated to localize the range of disparity matching. Figure 7.4(a) represents the window which is estimated from 10 defocus images. The disparity map is iteratively estimated by updating the fine (i.e. focusing) structure in the window. The sampling of the fine structure derives a pinhole image with minimum aperture. The dense disparity map and its recovered pinhole image respectively shown in Figures 7.4(b) and 7.4(c) show the excellent performance. For the subjective evaluation of the results, an image-based modeling is given in Figure 7.4(d). The image-based modeling is efficient to evaluate the localization performance by mapping the texture image to the depth.

We compare the performance of confocal anisotropic disparity estimation to the anisotropic disparity estimation method described in Chapter 6. Since the previous method needs a single well-focused image pair[55], disparity of defocusing regions in the background cannot be estimated as Figure 7.5(a) represents. The novel method produces a dense disparity map preserving structure of the focusing foreground and defocusing background as Figure 7.5(b) shows.

In a far-focusing stereo camera setup, we exemplify the fact which the disparity estimation of near-defocusing object is difficult in Figure 7.6. The reliable Markov Random Fields method [176] and graph-cut optimization [177] cannot estimate the disparity in the blurred regions

55) The pinhole camera model uses focal plane assumption that means well-focused images.

due to the ambiguity as Figures 7.6(b) and 7.6(c) shows. However, the proposed method overcomes the problem as Figure 7.6(d) represents since the ambiguity in the blurred regions can be avoided by the consistency in other focus images.

Chapter 8

Applications

In computer graphics and computer vision, image- (or video-)based modeling and rendering methods rely on a set of 2-D images of a scene to generate a 3-D model or some novel views of the scene. The traditional approach of the contents generation has been to create a geometric model in 3-D and try to render a scene by reprojecting the model onto a 2-D image. Video composition arranges the 2-D segments from computer graphics and real image objects in the frame. Computer vision is focused on detecting, grouping, and extracting image features in given images, and then tries to interpret them as 3-D clues. Image-based modeling and rendering allows the use of multiple 2-D images in order to generate directly novel 2-D images or a 3-D model, skipping the manual modeling stage. However, there are a lot of fundamental key problems related to robustness and accuracy of segmentation, tracking, self-calibration, pose estimation and depth estimation, etc., and the quality is still not sufficient to apply to general purpose. In this dissertation, most related topics of the problem have been surveyed, tested and evaluated to obtain an error-robust and generalized video analysis. In this chapter, the practical applications of the all proposed methods are discussed in detail. The issues that need to be addressed are the choice of the representation, a scheme to combine algorithms required to make video-based rendering and modeling applications.

8.1 Image- and Video-based Rendering

Novel view synthesis : We have devised a method for creating novel views of a scene from any multiple camera viewpoints of the same scene. There are two methods for a novel view synthesis: camera geometric warping and scene geometry compensation. The camera geometric warping method is an approach that rays of all viewpoint images of the same scene are warped into an arbitrary viewpoint, and the warped images are interpolated. Figures 8.1(e) and 8.1(f) exemplify viewpoint warping for a virtual camera viewpoint. A ray projection and its inverse projection are known as camera parameters. Chapter 3 presented a self-calibration from camera motion to obtain the ray projection and the inverse projection. However, this

method is efficient only when the virtual camera position is near to the original camera positions, and applications of global viewpoint warping, e.g. light field rendering [12–15] and 2D-3D conversion methods, usually require many camera viewpoints from a dense camera array or a long camera path. When the scene has a lot of near objects, the linear projection model cannot handle the nonlinear divergent ray components. In this case, 3-D sampling position on the object surface should be approximated to avoid the blurring and ghosting artifacts that is a linear sampling problem for divergent rays. Figure 8.1(a) shows a blurring and ghosting artifacts. Thus, a 3-D scene geometry compensation method, i.e. depth weighted shifting, cooperated with the camera geometric warping was proposed in Chapter 4 to eliminate the artifacts as Figure 8.1(b) represents. While a camera geometric warping forms global ray projections for a focal plane as Figure 8.1(c) shows, the sampling difference between focal plane and object surface estimates can be compensated to reform the diverging ray components as Figure 8.1(d) depicts.

Light field and plenoptic function : When the scene is static, the structure from motion technique can be utilized to calculate the global ray projections related to the camera pose. This method provides an automatic post-processing scheme to generate 3-D video from 2-D video frames, although the method is limited to videos that have only global camera motion such as landscape scenes. However, dynamic motion contents consist of a lot of independent foreground object motions, and light field rendering with a large camera array is a better choice for especially 3-D contents makers. Light field is the general term that defines the amount of light traveling in every direction through every point in space, and thereby the ray sampling for a moving camera is also a kind of light field. An arbitrary viewpoint can be reconstructed by light samples at several points in a given direction by camera calibration parameters. Such a bundle of samples was described by the *plenoptic function* that has a 5-D quantity describing a continuous light flow at every 3-D spatial position for every 2-D direction. In Chapter 3, a novel ray-space definition is presented using optical flow scheme. Bundles of rays form a cylinder style segments because an image can be segmented to several regions. The optical flow light field is very efficient to define a plenoptic function using a monocular camera motion. Moreover, the new definition can describe the boundary condition problem of bundles of rays using the field-equation theory deriving the Laplace equation. For a static scene, the technical basis of dense camera array and a monocular camera motion is very similar because the methods commonly utilize the principal ray on the optical axis for rendering. However, virtual camera position should be near to the original camera position since a projection ray[56] does not correspond to all 3-D sampling positions. A lot of cameras are needed to increase the sampling rate and avoid excessive blurriness in a synthesized viewpoint.

56) It represents a line between the center of a pinhole camera and a point of the focal surface.

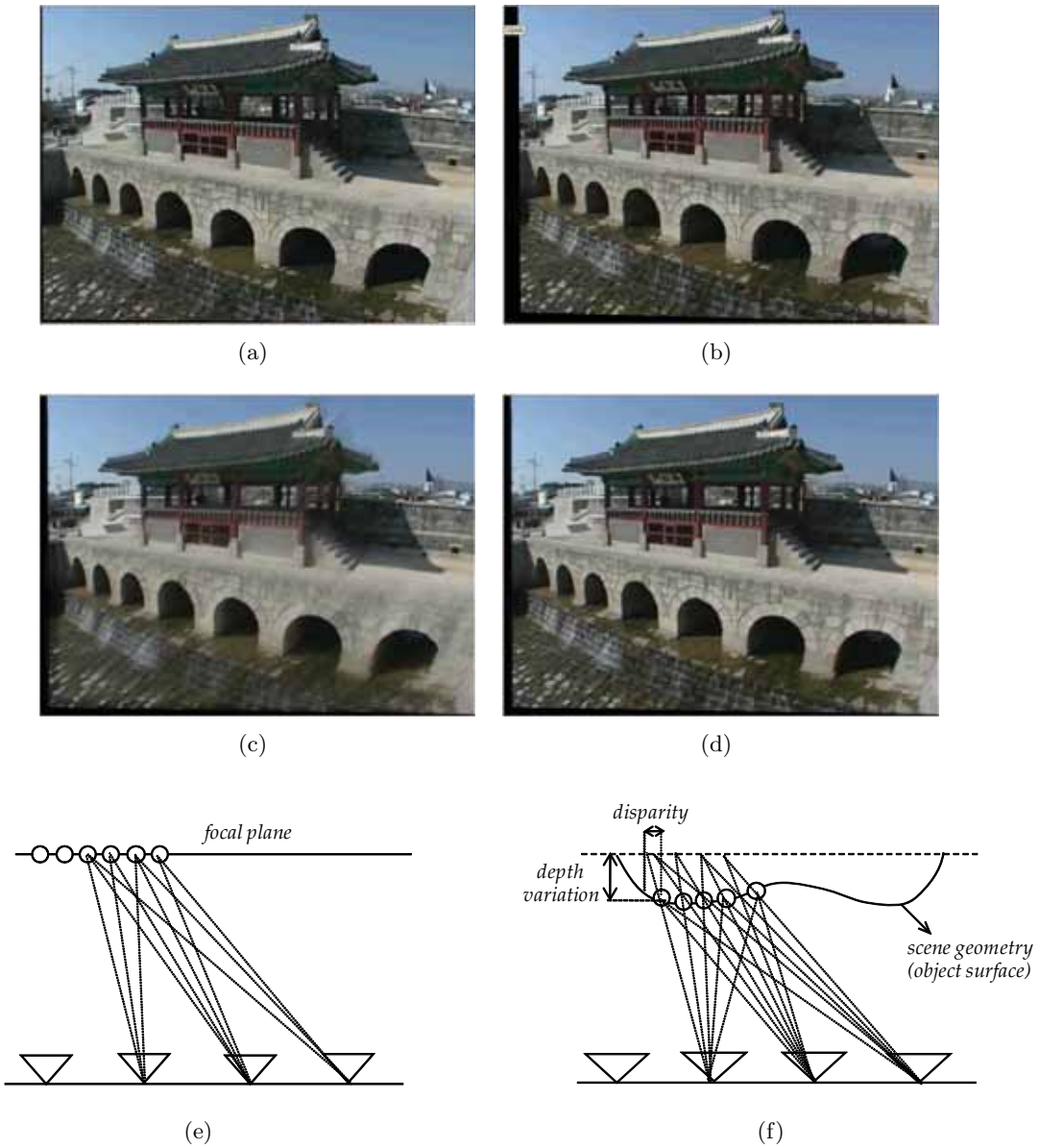


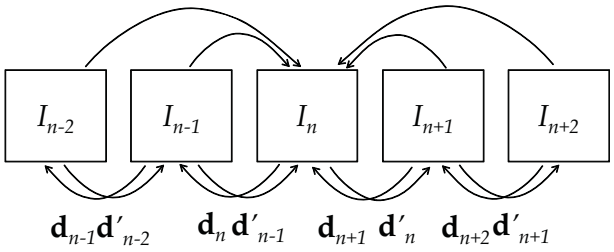
Figure 8.1. Light field rendering with 3-D scene geometry - (a) warped image for a camera, (b) another warped image, (c) blurring and ghosting artifacts, (d) warped camera image with scene geometry, (e) light field rendering with surface blurring, (f) light field rendering with scene geometry compensation

3-D sampling position : The scene geometry estimation can support to select accurate sampling positions. The distance between the two projected locations is compensated by scene geometry since the relationship among two cameras and a scene point can be obtained by triangulation. The depth estimation is difficult due to the geometrical discontinuity and hid-

den information, and a flat region without texture is troublesome for corresponding samples. Several combinations of multiple cameras, structured light, range finders, and mechanical sensing devices are used to acquire dense scene geometry since the passive stereo method is difficult to obtain 3-D geometry of natural scene. The optical flow method has been utilized to smoothly interpolate the motion of pixels moving from one camera position to another. However, large displacements between remote cameras cause local minima problems during the displacement tracking. In this dissertation, a novel sceneflow method that combines the dense stereo method and optical flow tracking is proposed to spatio-temporally estimating the 3-D geometry. For the spatio-temporal correspondences, multi-stereo allows a large depth range, while motion is being tracked to raise the sampling density. When the epipolar geometry of the stereo camera is known, an image point should be matched to an image point on the epipolar line of the other camera. The linearizations of two temporal frames form the temporal epipolar geometry. The spatio-temporal epipolar geometries can be used to approximate the 3-D scene geometry since the intersection of spatial and temporal epipolar lines approximates a reliable 3-D points. A Fourier spectral analysis of spatio-temporal ray-space efficiently shows the amount of divergent rays.

Super-sampling Light field : Exact sampling positions on the dense depth plane are useful to eliminate the spectral components of divergent rays. A depth compensated rendering using sceneflow estimates allows high density of sampling. If the sampling density is sufficiently high, the image can be enhanced by a super-sampling scheme. However, camera geometry and scene geometry estimations can be influenced by the noise and measurement errors. A unified robust estimation method that can be used for both camera parameter estimation and scene geometry regularization was proposed in Chapter 5. A super-sampling noise removal is exemplified in Figure 5.18. The camera parameters is estimated using feature points in multiple viewpoint images, the epipolar geometry is obtained by camera projections using the method described in Chapter 4.3. The scene geometry between viewpoints can be estimated by dense disparity matching in epipolar lines, and the disparity vectors between viewpoints are used for a depth compensated warping of all viewpoint images. Figures 8.2(a) and 8.2(b) respectively show multiple viewpoint images and the sampling relationship between viewpoint images. The **d**-chains are combined to approximate the 3-D sceneflow by a chain correspondence method described in Chapter 6.4. Figure 8.2(c) shows a robust dense disparity map between two views that is estimated by the anisotropic disparity estimation, and the robust error statistic efficiently removes outliers shown in Figure 8.2(d) from the disparity map. The super-sampling using all viewpoints can increase the image resolution with deblurring effects as Figure 8.2(e) depicts. The difference between original and enhanced images is shown in Figure 8.2(f).

(a)



(b)

(c)

(d)

(e)

(f)

Figure 8.2. Super-sampling view rendering - (a) multi-viewpoint images, (b) super-sampling with camera geometry and scene geometry, (c) dense scene geometry, (d) robust error statistic, i.e. white color region is outliers, (e) image enhancement with deblurring, i.e. left: original image, right: enhanced image, (f) difference between original and enhanced images

8.2 Depth Segmentation and Image Composition

Object segmentation plays a significant role for the image and video composition. The recent image and video compositing systems need an accurate segmentation without visual seams to make compositing with virtual studio, blue screen, multi-color compositing, dynamic blue screen, advertisement insertion and synthetic transfiguration etc. However, object segmentation from a natural scene, which aims to extract exactly meaningful objects from a background, is still one of the most difficult problems, even though it has been researched for more than thirty years. A lot of semi-automatic methods have been proposed using user-defined rules, e.g. feature-based, contour-based or region-based which directly point out initial object [178]. The general process is to collect only meaningful groups of pixels which can be extracted and tracked. The tracking mechanisms evolve and propagate the initial object using boundary errors. However, semi-automatic methods have a serious disadvantage, the necessity of user initialization. Thus, they are not suitable for TV applications. Fully automatic methods [179] apply the extraction rules based on specific characteristics of the scene or on the known priori information. For example, the face segmentation uses spectral characteristics of the skin-color region in the CIE chromaticity diagram. Chroma-keying, i.e. the blue screen method, accurately eliminates the uniformly colored background. However it requires a specific set-up of the background and special care for lighting conditions to avoid shadows and reflecting blue light and to maintain uniform intensity. To overcome some limitation of the chroma-keying, Z-keying [180], i.e. a depth keying method which uses distance-based measures, was proposed. A depth map can be accurately computed by cameras with depth sensors that supplement every video frame with an additional frame, the *Z-buffer*. However, the method is restricted to indoor applications since the depth range of the sensor is limited. For both, indoor and outdoor conditions, depth estimation recovered from multiple cameras substitutes the use of depth sensors. However, the depth estimation provides low spatial accuracy in the segmentation results when the objects are sparsely textured. If a segmented object is laid over a new background, old background areas which are extracted with the object can be distinctly recognized as a *seam* in the object boundary. Thus, partial replacement of the visible seam is required for pixel-wise accuracy. The precise definitions of the object boundary and coherence can be used to maintain the spatial accuracy.

Sceneflow-based segmentation : The robust anisotropic disparity estimation method proposed in Chapter 6 can efficiently segment the foreground object. A dense depth map allows separating objects at some specific depths. If cameras are calibrated and the epipolar geometry is spatio-temporally estimated, the 3-D depth can be recovered from the dense disparity estimates. A conventional depth estimation has several difficulties, e.g. the ambiguity of the local image structure due to image noise, unbalanced brightness, similar texture and occlusion, etc. If two pixels in the same image look alike, it may be impossible to find corresponding

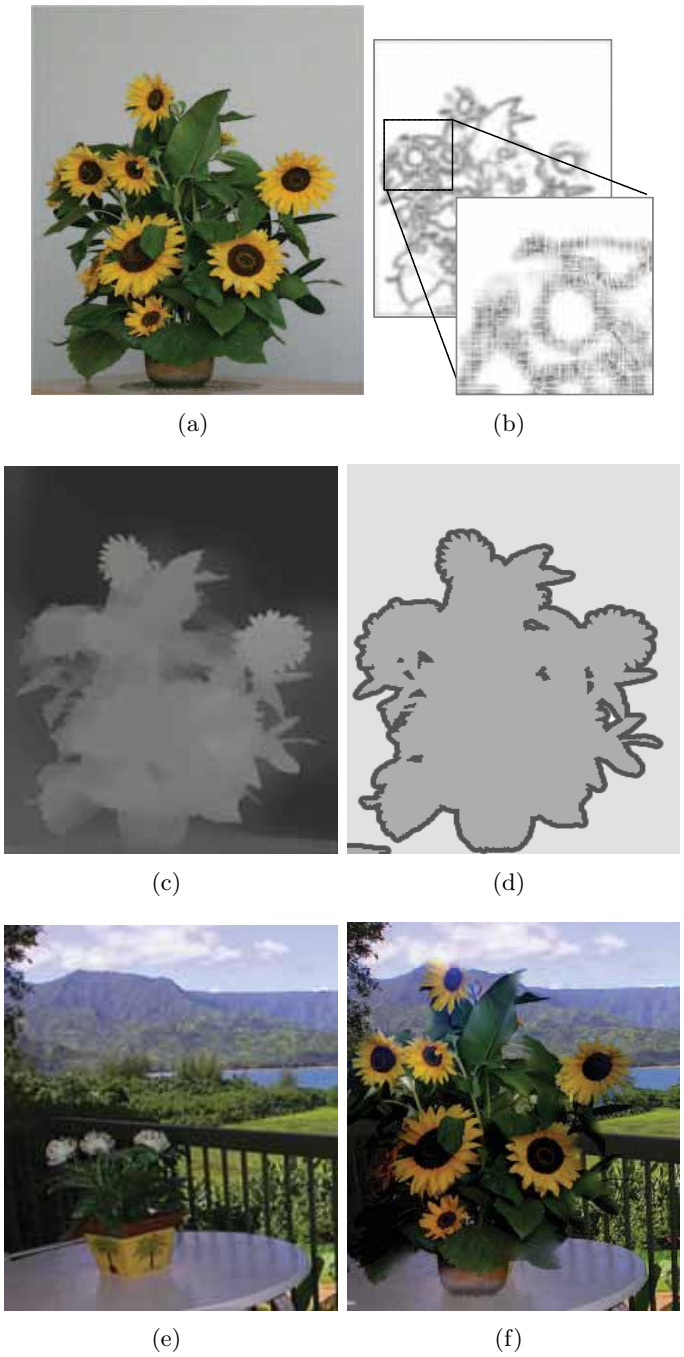


Figure 8.3. Image composition with depth segmentation - (a) a stereo images, (b) edge flow field, (c) estimated depth map, (d) trimap, (e) new background image, (f) composited image

pixels in the other image only based on the similarity. A robust region segmentation method that is performed by analyzing the local image structure was proposed in Chapters 5.4 and 5.5. We track a region by the edge flow scheme employing a smoothly varying oriented struc-

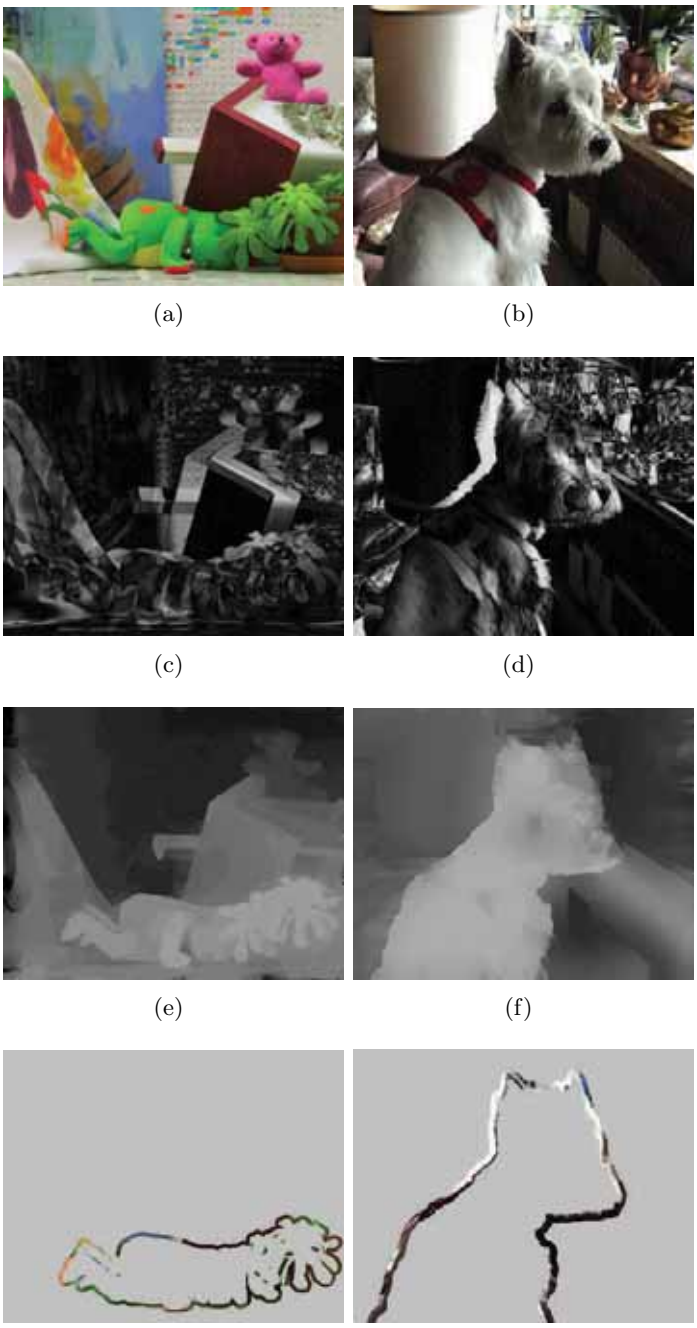
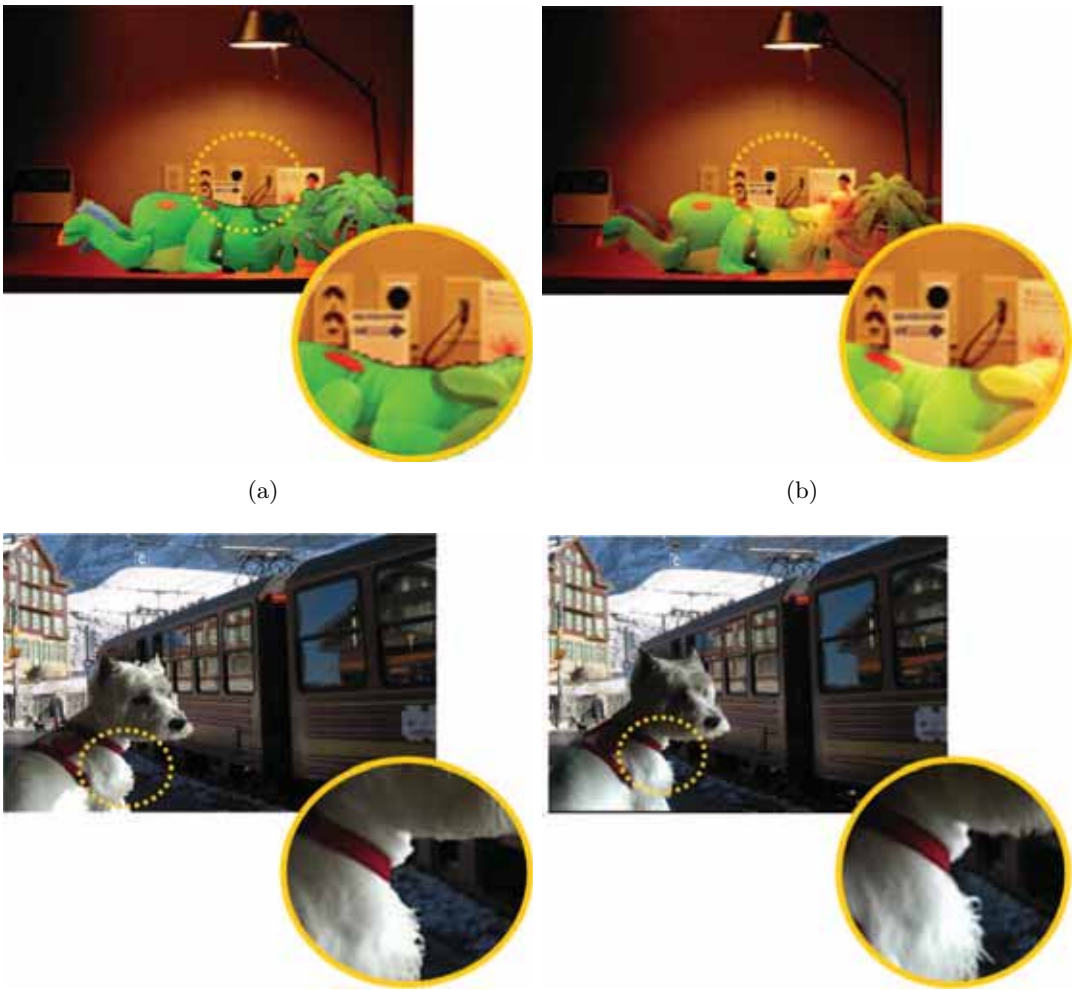


Figure 8.4. Blending with depth segmentation - (a) the left Teddy image, (b) the left Dog image, (c) difference between left and right Teddy images, (d) difference between left and right Dog images, (e) depth map of Teddy images, (f) depth map of Dog image, (g) trimap of Teddy images, (h) trimap of Dog images



(a) (b)

(c) (d)

Figure 8.5. Seamless composition - (a) the left Teddy image, (b) the left Dog image, (c) difference between left and right Teddy images, (d) difference between left and right Dog images, (e) depth map of Teddy images, (f) depth map of Dog image, (g) tri-map of Teddy images, (h) trimap of Dog images

ture. The local maxima in the structure separates invariance image regions. The proposed depth estimation is performed by using the invariance constraint of tracked regions.
Figure 8.3(b) exemplifies the edge flow estimation that is derived from the diffusion field in an image shown in Figure 8.3(a). The full resolution depth map is estimated by the anisotropic disparity estimation as Figure 8.3(c) shows. The anisotropic diffusion coefficient decreases at the edges with steep intensity gradients as Chapter 6.1 shows, an over-smoothing of depth discontinuities can be avoided. Since the anisotropic diffusion smooths homogeneous regions, the object depth is more homogeneous. The tri-map shown in Figure 8.3(d) is set up from

the discontinuities of homogeneous regions, and used as a segmentation mask. After locating the segmented object on the new background image, the edge flow method is applied to directional blending of pixels in the segmented object region and the new background corresponding to the tri-map. Figure 8.3(e) shows a new background image. By suitably incorporating the edge-preserved diffusion, blending power is near zero at the object boundary. Finally a naturally blended object can be obtained as Figure 8.3(f) shows.
Depth-based image composition for Teddy and Dog stereo images are demonstrated in Figure 8.4. Figures 8.4(a) and 8.4(b) are the left images of the Teddy and Dong stereo image pairs, and Figures 8.4(c) and 8.4(d) are the difference between the left and right images. The depth map is automatically estimated between the left and right images as Figures 8.4(e) and 8.4(f) represents. Finally, the foreground objects are segmented in the trimaps shown in Figures 8.4(g) and 8.4(h). When we apply the directional blending to the Teddy image, the composition result is a natural looking fusion with a brightness adaptation as Figure 8.5(b) depicts. The composition without a blending in Figure 8.5(a) shows some visual seams, i.e. the darker pixels on the object boundary of a green doll. In Figure 8.5(c), the membrane regions with furs of a dog are blended for the texture directions, and a natural looking composition of the particular structure is obtained as Figure 8.5(d) depicts.

8.3 Image and Video-based Modeling

The need for real-world 3-D models is increasing constantly due to the extensive use of computer graphics in the game industry, electronic commerce, the special effects and film industry. Production of photorealistic models of real scenes is still very time consuming and feasible only with expensive 3-D scanners in conjunction with manual interaction. The image-(or video-) based modeling method differs from traditional computer graphics in that both the geometry and appearance of the scene are derived from real photographs. The techniques often allow for shorter modeling times, faster rendering speeds, and unprecedented levels of photorealism.
For a rigid scene, the difference in the viewpoints is caused by a single camera undergoing motion or two cameras at different positions. Camera parameters describe the rotational and translational differences between image frames, and correspondences in the images are closely related by the collineation and the constraint. A topic in this dissertation is self-camera calibration from stereo image sequences. In Chapter 3, camera calibration parameters are estimated by a feature tracking method in an image sequences acquired by uncalibrated video cameras. First, we specify the epipolar geometry by estimating the fundamental matrix and computing the essential matrix from the fundamental matrix. The fundamental matrix estimation using parallax method was introduced in Chapter 4.4, and the computation of the essential matrix is described in Chapter 4.5. The rotation and translation of cameras

(a) (b)

Figure 8.6. Images for 3-D sceneflow Modeling - (a) Castle stereo image sequence with 15 frames, (b) Palace stereo image sequence with 10 frames

are obtained from the result by splitting the essential matrix. Next, the dense 3-D scene geometry is estimated by the sceneflow method in Chapter 6. The epipolar constraint is used to restrict the 2-D matching problem to a 1-D epipolar line. Temporal epipolar geometry was introduced in Chapter 4.8. The intersections of spatial and temporal epipolar lines form a set of stable 3-D points. The errors between the sceneflow and the camera parameters are fitted by a nonlinear bundle adjustment method. Figure 8.6 represents a test image sequence. 15 frames of the Castle stereo sequence (i.e. total 30 images), and 10 frames of the Palace stereo sequence (i.e. total 20 images) are used to reconstruct 3-D models. However, the bundle adjustment optimization needs a high computational power to minimize the back-projection errors in full sets of spatio-temporal correspondences. Although we expect that more frames can increase the quality, it was difficult to increase the number of frames in our simulation environments due to the complexity. Figure 8.7 shows video-based modeling results for the Castle and Palace image sequences.

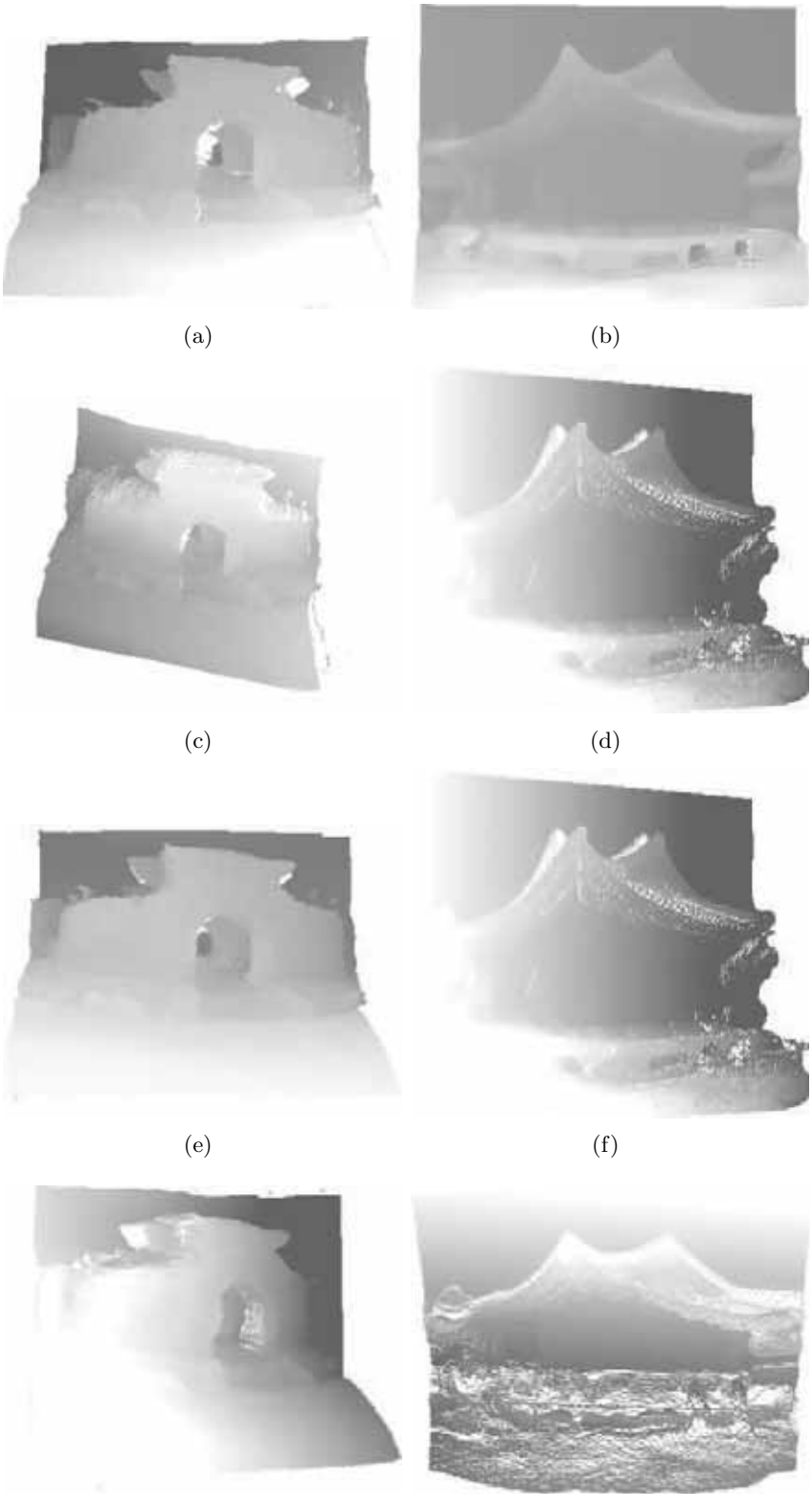


Figure 8.7. Results of 3-D sceneflow modeling - (a) recovered model from Castle stereo image sequence, (b) recovered model from Palace stereo image sequence

Chapter 9

Discussion

In this dissertation, we proposed a set of methods to recover the 3-D camera geometry and the dense scene structure from stereo camera motions. Our work is just a small progress of robust 3-D video analysis for automatic 3-D content generation. Much remains to be done, both in improving and extending the current framework. We hope that others will build on our ideas and other related works to provide more precise interpretations and applications. In this final chapter, we discuss the proposed sceneflow method for robust 3-D video analysis and further steps towards understanding stereo motions.

9.1 Sceneflow Method for Robust 3-D Video Analysis

Camera Geometry and Scene Geometry : Image- (or Video-) based modeling and rendering are active research topics in the computer vision and graphics community. In the last decade, many systems were proposed free-viewpoint generation based on collections of viewpoint images of a scene. The image- (or video-) based modeling approach is focused on the reconstruction of scene geometry, and the image- (or video-) based rendering approach is based on the projective warping of rays. In fact, the technical basis in the two approaches is closely related since the projective geometry is a non-metrical form of geometry that formalizes a ray between an infinity point and the camera center, and the scene geometry is related to the metric components in a ray. Since feature-based methods are more efficient to solve the parametric equation accurately, the camera geometry is often estimated from feature correspondence. However, feature-based methods can reconstruct only a restricted number of scene points, i.e. sparse metric reconstruction. Obtaining a dense reconstruction could be achieved by interpolation, but the result is not at a satisfactory level. These problems can be avoided by establishing dense correspondences. When the projective geometry is estimated by the feature-based calibration method between image pairs, we exploit the epipolar constraint that restricts the 2-D correspondence search to a 1-D line.

Sceneflow analysis : Stereo vision is a good modality for dense 3-D analysis of the real world surrounding us, but it suffers from some significant shortcomings. Cases posing difficulties in stereo vision are the insufficient amount of texture in the images, an unbalanced lighting condition, perspective effects, depth discontinuities, occlusions and etc. The conventional dense matching methods do not preserve the spatio-temporal discontinuities, object boundaries are often poorly localized, and occlusion holes and mismatches are located in some regions. The proposed method employs a robust image statistic in the regularization function to eliminate outliers, e.g. mismatches, mislocalization errors and occlusion holes etc. However, if there are not sufficient dense estimates as much as image pixel, the diffusion falls to local minima during solving partial differential equations, and large-baseline matching often causes this problem. In this dissertation, the 3-D sceneflow method that combines stereo matching and optical flow motion was proposed. For a rigid scene, a multi-stereo method can be temporally expended to a structure-from-motion approach. In general, the denser capturing of multiview images with a large number of cameras provides the more precise 3-D representation. A Structure-from-Motion (SfM) method is easier to obtain the spatial camera density than a multi-stereo method. The proposed method combines the two methods in the 3-D sceneflow framework. The images may come either from a set of closely spaced cameras, hand-held video camcorders, or a multi-camera rig, etc. We can obtain a lot of overlapped 3-D points from spatio-temporal correspondences. Together with the camera geometry estimation, a set of 3-D scene points is recovered based on the static scene hypothesis.

Robust estimation : Due to physical imperfections in imaging sensors and bad locations of detected features and false matches, the image data contains outliers. We consider a Gaussian distribution behavior of the location error of a point of interest. A robust estimator considers the relation between the data points and reasonable model candidates as a scale, and thereby the outliers can be removed by different contributions of scales for the data of interest. We proposed a unified scheme to eliminate outliers for fundamental matrix estimation and dense scene geometry estimation. A robust estimation criterion that is derived from the error distribution transforms the fundamental matrix by minimizing the scale-weighted error. The robust error statistic is applied to the error function of a 3-D sceneflow. The error distribution of estimates is restricted in a smoothly invariant support region, and an anisotropic regularization method efficiently eliminates outliers in the regional structure.

Spatio-temporal baseline : The spatio-temporal 3-D space can be represented by a brightness-invariance concept. The image brightness of a point in the scene not only remains constant from one camera viewpoint to another but also remains constant over the time frames. Stereo motions form the spatial epipolar geometry between two stereo viewpoints and the temporal epipolar geometry between two time frames. The sceneflow method uses

joint correspondences in the spatio-temporal image frames, and results in very strong multi-linear constraints. The intersection of spatio-temporal epipolar lines generates the sparse 3-D scene geometry with spatio-temporally stable 3-D points. The points at the line intersections is a set of reliable 3-D points. Since the camera translation and rotation parameters can be recovered from the essential matrix, the stable corresponding points yield an accurate update of the projective geometry. We use the projective geometry for optimizing the 3-D sceneflow by a nonlinear bundle adjustment method.

Application of the 3-D sceneflow : The 3-D sceneflow estimates can be used in many ways. One way would be to reconstruct a 3-D scene from all overlapping spatio-temporal views. A standard SfM method is based on feature tracking and results in a sparse 3-D reconstruction. However, the sparse 3-D method is difficult to distinguish between connected and disconnected surfaces, and thus the topology problem is a drawback. A dense matching method may directly solve the problem by triangulating[57] 3-D points from all depth maps. The wireframe model can then be textured with the real images. Another way is a virtual view synthesis directly from real views using depth-compensated warping. However, the camera geometry allows only one ray projection model for one camera, and rays diverging on a 3-D surface do not enter in the camera center and causes blurring and ghosting artifacts. Light field rendering always depends on exact and dense 3-D plenoptic sampling. Dense 3-D sampling positions on the 3-D object surface should be approximated to avoid the divergent ray problems. The 3-D sceneflow method allows high density 3-D sampling using all spatio-temporal correspondences.

9.2 Future directions

In recent years, 3-D computer vision researchers have made tremendous progress in 3-D video analysis such as multi-stereo, structure from motion, light field rendering and camera pose estimation, etc. However, they still are not sufficient for a general understanding of the 3-D scene. Most methods start with feature correspondences that are invariant against viewpoint and illumination changes and they estimate the camera geometry and the scene geometry. However, a computer that may be able to detect the image features and to find matches cannot truly understand the scene layout and its contents. For example, 3-D computer vision methods cannot utilize arbitrarily chosen images from a digital photo sharing database (e.g. Flicker, Google image search, etc.) for camera pose estimation or 3-D structure reconstruction, even when the images are parts of the same scene. Here, we propose several ideas for future works that extend our current framework.

57) The triangulation of surface points means meshing.

Affine invariant region detection : The detection of invariant regions is a crucial step in image indexing tasks. Since the perspective transform in a wide baseline setup can be locally approximated by an affine transform, commonly used detection is the affine invariant one. A feature detector can provide locations at which a local affine invariant descriptor is computed. The most challenging problem in these approaches is to find the correct scale, i.e. the spatial extension of the support region around the point. The anisotropic diffusion tensor being included in our framework can be utilized to expand a detected region or to find a stable border.

Statistical learning : Rather than trying to explicitly compute all of the required geometric parameters from images, the statistical learning, e.g. classification, recognition and training etc., can support to establish the real surface topology between all information of image features, regions and occlusion boundary, etc. In contrast to most recognition approaches that model semantic classes, we can use the 3-D geometry of sceneflow for the 3-D object description.

Usage of fast optimization : A lot of reliable correspondences can be established by the proposed 3-D sceneflow method, but there are a lot of metric differences between the 2-D visual structure and the real 3-D structure. We perform a metric bundle adjustment method to refine the 3-D sceneflow. However, the bundle adjustment optimization is mathematically simple but practically very complex. Although the method has been well applied to a lot of SfM studies, we have realized that the method is too complex to optimize a dense scene structure. A region-based adjustment method may solve the complexity problem, because image regions divided by stable features have similar depths[58]. Instead of minimizing the reprojection error for every point, the method will try to minimize the global error in each divided region.

Real object boundary recovery : In this thesis, we found the depth discontinuities by using the homogeneity of regions. However, real object boundaries are typically not defined by homogeneity, but by physical connectedness. For example, image segmentation methods rely on 2-D perceptual grouping cues, e.g. brightness, color or texture similarity, edge strength, continuity, and closure etc. However, the boundaries of such segmentations often correspond to texture, illumination, or material discontinuities, rather than true object boundaries. Although the 3-D sceneflow method provides some 3-D cues such as stereo and motion, the stereo motion cues sensitively depends on the 2-D image conditions. In future works, we will try to order all 2-D perceptual grouping cues, color cues, disparity and motion cues of prominent objects in sufficient detail to estimate the real depth boundaries.

58) This is the smoothness constraint of dense depth estimation.

Chapter 10

Conclusion

In this dissertation, we have taken a first step towards recovering the camera geometry and the scene geometry of a static scene from stereo camera motion. A lot of possible modeling and rendering experiments and implementations have been shown over this dissertation. We hope that our work will inspire others to develop more advanced algorithms and 3-D applications. We conclude with a statement of the philosophy that has driven our research:

Since camera viewpoints are relative to the fixed 3-D object, the stereo and motion can be integrally analyzed for a static scene. A multi-stereo method can cooperate with the SfM method in recovering the camera pose and the scene structure. Since rays projected from scene points can be modeled by one projective camera geometry that defines a line between the camera center and the focal point. However, the projective geometry causes image distortions according to the metric component of scene structures. The dense 3-D scene geometry estimation is a process to recover the metric geometry according to the scene structure. The quality of image-based modeling and rendering depends on the robustness and density of 3-D scene sampling. The 3-D scene geometry estimation using robust error statistics is efficient both to enhance reliable estimates and to remove the outliers without affecting the geometric discontinuities. The key to obtain higher density of reliable correspondences is a combination of stereo and motion.

Appendix A

Appendix

A.1 Relationship between Collineations and Epipolar Geometry

When observing a plane, an interesting specialization of the epipolar geometry of two views can be obtained within a collineation of $\mathcal{P}^2$ between a world plane and its perspective image. If we choose the world coordinate system so that the plane has an equation of $z = 0$, the projection equation can be expanded as

$$\alpha \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{bmatrix} \begin{bmatrix} x \\ y \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{14} \\ p_{21} & p_{22} & p_{24} \\ p_{31} & p_{32} & p_{34} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{A.1}$$

where α is an arbitrary scale. Points are mapped from the world plane to the image plane with a 3×3 non-singular matrix, which represents a collineation of $\mathcal{P}^2$. If we have a collineation from Π to the left image plane, and another collineation from Π to the right image plane, a collineation from the image plane of the left camera to the image plane of the right camera can be defined by composing the inverse of the first with the second. The plane Pi induces a homography $\mathbf{H_\Pi}$ between the views, which transfers points from one view to the other.

$$\tilde{\mathbf{m}} \simeq \mathbf{H}_\Pi \tilde{\mathbf{m}}_l \quad \text{if} \quad \tilde{\mathbf{M}} \in \Pi \tag{A.2}$$

where $\mathbf{H_\Pi}$ is a 3×3 non-singular matrix. Even though a collineation of $\mathcal{P}^2$ depends upon eight parameters, there is no contradiction with the fact that a plane depends upon three parameters. Indeed, the collineation induced by a plane must be compatible with the epipolar

geometry, i.e.:

$$(\mathbf{H}_{\Pi}\tilde{\mathbf{m}}_l)^T \mathbf{F}\tilde{\mathbf{m}}_l = 0 \tag{A.3}$$

for all points $\tilde{\mathbf{m}}_l$. $\mathbf{F}$ is the fundamental matrix. This implies that the matrix $\mathbf{H}_{\Pi}^T\mathbf{F}$ is antisymmetric:

$$\mathbf{H}_{\Pi}^T\mathbf{F} + \mathbf{F}^T\mathbf{H}_{\Pi} = 0 \tag{A.4}$$

and this imposes six homogeneous constraints on $\mathbf{H}_{\Pi}$. A collineation $\mathbf{H}$ that satisfies Equation (A.4) is said to be compatible with $\mathbf{F}$. A collineation $\mathbf{H}$ is compatible with $\mathbf{F}$ only if

$$\mathbf{F} \simeq [\mathbf{e}_r]_{\times}\mathbf{H} \tag{A.5}$$

Π does not contain c_r:

$$\mathbf{H}_{\Pi}\mathbf{e}_l \simeq \mathbf{e}_r \tag{A.6}$$

A.2 Triangulation

Given the camera matrices $\mathbf{P_l}$ and $\mathbf{P_r}$, two corresponding points $\tilde{\mathbf{m}}_l$ and $\tilde{\mathbf{m}}_r$ satisfying the epipolar constraint $\tilde{\mathbf{m}}_l\mathbf{F}\tilde{\mathbf{m}}_r = 0$, $\tilde{\mathbf{m}}_r$ lies on the epipolar line $\tilde{\mathbf{F}}\mathbf{m}_l$ and thereby the two rays back-projected from image points $\tilde{\mathbf{m}}_l$ and $\tilde{\mathbf{m}}_r$ lie in a common epipolar plane. Since they lie in the same plane, they will intersect at the some point. This point is the reconstructed 3-D scene point $\tilde{\mathbf{M}}$. Analytically, the 3-D point $\tilde{\mathbf{M}}$ can be found by solving the known parameter α_l or α_l.

$$\mathbf{e}_r = \alpha_r\tilde{\mathbf{m}}_r - \alpha_l\tilde{\mathbf{m}}_l \tag{A.7}$$

where α_l and α_r respectively corresponds to the depth of $\tilde{\mathbf{M}}$ in the left and right cameras. The three points $\tilde{\mathbf{m}}_r$, $\mathbf{e}_r$ and $\tilde{\mathbf{m}}_l$ are collinear, so we can solve α_l using the following closed form expressions:

$$\alpha_r = \frac{(\mathbf{e}_r \times \tilde{\mathbf{m}}_l) \cdot (\tilde{\mathbf{m}}_r \times \tilde{\mathbf{m}}_l)}{\|\tilde{\mathbf{m}}_r \times \tilde{\mathbf{m}}_l\|^2} \tag{A.8}$$

where $\alpha_l (\mathbf{e}_r \times \tilde{\mathbf{m}}_l) = \alpha_l\mathbf{F}^T\tilde{\mathbf{m}}_r$ and the reconstructed point $\mathbf{M}$ can be calculated by inserting the value α into Equation (3.4). Since the camera parameters and image locations are known only approximately, the back-projected rays do not actually intersect in space. Therefore Equation (A.8) solves Equation (A.7) in a least squares sense.

A.3 Homography and Canonical Basis

The method of mapping a projective representation of a plane to another is described here. Consider the 3×3 linear transformation $\mathbf{H}$, i.e. homography, and the four projective canonical points: $\varepsilon_1 = [1,0,0]^T$,$\varepsilon_2 = [0,1,0]^T$,$\varepsilon_3 = [0,0,1]^T$,$\varepsilon_4 = [1,1,1]^T$. The following equation shows that the four canonical points can be mapped to four arbitrary points $\mathbf{p_1}$, $\mathbf{p_2}$, $\mathbf{p_3}$ and $\mathbf{p_4}$, with a suitable substitution for $\mathbf{H}$. A point $\mathbf{p}$ and any scaled version of it, i.e. $\lambda\mathbf{p}$, represent the same point.

$$\begin{aligned}\mathbf{H}\begin{bmatrix}1&0&0&1\\0&1&0&1\\0&0&1&1\end{bmatrix} &= \begin{bmatrix}\lambda_1\mathbf{p}_1 & \lambda_2\mathbf{p}_2 & \lambda_3\mathbf{p}_3\end{bmatrix}\begin{bmatrix}1&0&0&1\\0&1&0&1\\0&0&1&1\end{bmatrix}\\ &= \begin{bmatrix}\lambda_1\mathbf{p}_1 & \lambda_2\mathbf{p}_2 & \lambda_3\mathbf{p}_3 & (\lambda_1\mathbf{p}_1+\lambda_2\mathbf{p}_2+\lambda_3\mathbf{p}_3)\end{bmatrix}\\ &= \begin{bmatrix}\lambda_1\mathbf{p}_1 & \lambda_2\mathbf{p}_2 & \lambda_3\mathbf{p}_3 & \mathbf{p}_4\end{bmatrix}\end{aligned} \tag{A.9}$$

where $\mathbf{H} = \begin{bmatrix}\lambda_1\mathbf{p}_1 & \lambda_2\mathbf{p}_2 & \lambda_3\mathbf{p}_3\end{bmatrix}$ and $\mathbf{p}_4 = \lambda_1\mathbf{p}_1 + \lambda_2\mathbf{p}_2 + \lambda_3\mathbf{p}_3$. Each point $\mathbf{p}$ has only two degrees of freedom. Similarly the matrix $\mathbf{H}$ has eight degrees of freedom, as it can be scaled by an arbitrary parameter λ. Thus, 4 points are needed to define the transformation $\mathbf{H}$, i.e. 2 degrees of freedom $\times$ 4 points = 8 degrees of freedom. In projective coordinates, any quadrilateral may be mapped to any other quadrilateral using the 3×3 homography $\mathbf{H}$.

A.4 Projective Transformations of a Plane

The homography $\mathbf{p}' = \mathbf{Hp}$ is a transformation which belongs to the projective group. This group has a number of subgroups which explain the connection between projective and Euclidean coordinates. The affine group consists of transformations $\mathbf{p}' = \mathbf{Ap}$, that have the form:

$$\begin{bmatrix}\lambda u'\\ \lambda v'\\ \lambda\end{bmatrix} = \begin{bmatrix}a_{11}&a_{12}&a_{13}\\ a_{21}&a_{22}&a_{23}\\ 0&0&1\end{bmatrix}\begin{bmatrix}\lambda u\\ \lambda v\\ \lambda\end{bmatrix} \tag{A.10}$$

These transformations can also be written as: $\tilde{\mathbf{m}}' = \overset{2\times2}{\mathbf{A}}\,\tilde{\mathbf{m}} + \tilde{\mathbf{a}}$, or

$$\begin{bmatrix}u'\\ v'\end{bmatrix} = \begin{bmatrix}a_{11}&a_{12}\\ a_{21}&a_{22}\end{bmatrix}\begin{bmatrix}u\\ v\end{bmatrix} + \begin{bmatrix}a_{13}\\ a_{23}\end{bmatrix} \tag{A.11}$$

The affine group corresponds to a 2×2 linear transformation and a 2×1 translation, and can be applied using standard Cartesian coordinates. The six components of this transformation

can be decomposed as two translations, two scale components, a shear and a rotation. If the transformation is restricted further, by removing the two scale components and the shear, then the Euclidean transformations of the plane is obtained as

$$\begin{bmatrix} \lambda u' \\ \lambda v' \\ \lambda \end{bmatrix} = \begin{bmatrix} \overset{2\times2}{\mathbf{R}} & \overset{2\times1}{\mathbf{a}} \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} \lambda u \\ \lambda v \\ \lambda \end{bmatrix} \tag{A.12}$$

which may also be written in Cartesian form as: $\tilde{\mathbf{m}}' = \overset{2\times2}{\mathbf{R}}\,\tilde{\mathbf{m}} + \tilde{\mathbf{a}}$. Alternatively, the projective transformation may be restricted by removing the rotations, and then the similarity group is obtained. This group contains only scale, shear and translation. This group is useful because it is used to represent the camera calibration matrix. This can be written as $\tilde{\mathbf{m}}' = \overset{2\times2}{\mathbf{C}}\,\tilde{\mathbf{m}} + \tilde{\mathbf{a}}$, or

$$\begin{bmatrix} \lambda u' \\ \lambda v' \\ \lambda \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \lambda u \\ \lambda v \\ \lambda \end{bmatrix} \tag{A.13}$$

The Euclidean and similarity groups, are subgroups of the affine group which is a subgroup of the projective group.

A.5 8-Point Algorithm

The 8-point algorithm was developed by Longuet-Higgins [24]. It is a frequently used method for computing the fundamental matrix from a set of eight or more correspondences. This algorithm has the advantage of simplicity in implementation. To make it less sensitive to noise and thus greatly improving the result, Hartley normalized (translated and scaled) the coordinates of the correspondences before calculation. This modification brings much better stability to the 8-Point Algorithm.

If the corresponding n-points have homogeneous coordinates, i.e. $\mathbf{m}_n = [u_n, v_n, 1]^T$ and $\mathbf{m}'_n = [u'_n, v'_n, 1]^T$, the epipolar constraint can be expressed as the following equation:

$$\begin{bmatrix} u'_n & v'_n & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} u_n \\ v_n \\ 1 \end{bmatrix} = 0 \tag{A.14}$$

This is equivalent to the following equation and can be solved with the rank constraint of matrix **F** with the 9 unknown coefficients:

$$\begin{bmatrix} u_1u_1 & v_1u_1 & u_1 & u_1v_1 & v_1v_1 & v_1 & u_1 & v_1 & 1 \\ u_2u_2 & v_2u_2 & u_2 & u_2v_2 & v_2v_2 & v_2 & u_2 & v_2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_nu_n & v_nu_n & u_n & u_nv_n & v_nv_n & v_n & u_n & v_n & 1 \end{bmatrix} \begin{bmatrix} f_{11} \\ f_{12} \\ f_{13} \\ f_{21} \\ f_{22} \\ f_{23} \\ f_{31} \\ f_{32} \\ f_{33} \end{bmatrix} = 0 \tag{A.15}$$

However, the high instability of the linear method is the problem since it is quite sensitive to noise, even with a large set of data points. To reduce the impact of noise on the 8-Point algorithm, the coordinates of point correspondences are normalized before running the 8-point algorithm. The normalizing transformation is applied to each of the two images independently as follows:

(1) Translate the point correspondences, so that their collective centroid is at the origin.
(2) Scale the point correspondences, so that the average distance from the origin is equal to $\sqrt{2}$

It is necessary to denormalize the results after obtaining a solution for the fundamental matrix. Denormalization is an inverse operation of the normalizing transformation.

A.6 Computation of Rotation and Translation

The rotation and translation of the camera can be obtained by splitting the essential matrix. The translation vector can only be determined up to a scale factor. The essential matrix is defined as

$$\mathbf{E} = [\mathbf{t}]_\times \mathbf{R} \tag{A.16}$$

where $\mathbf{R}$ is the rotation matrix and $[\mathbf{t}]_\times$ is the skew-symmetric matrix of the translation vector T. The essential matrix can be splitted by the singular value decomposition (SVD) as Equation 4.15 shows. The $[\mathbf{t}]_\times$ component is defined as

$$[\mathbf{t}]_\times = \lambda \mathbf{U}\mathbf{S}\mathbf{U}^T \tag{A.17}$$

where λ is scale factor, $\mathbf{U}$ is 3×3 orthogonal matrix and $\mathbf{S} = diag\,(1, 1, 0)\,\mathbf{S}^T$ is a skew-symmetric matrix.

$$\mathbf{S} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{A.18}$$

and $\mathbf{S^T} = -\mathbf{S}$. Equation A.6 can be rewritten up to a scale as

$$[\mathbf{t}]_\times = \mathbf{U}diag(1, 1, 0)\mathbf{S}^T\mathbf{U}^T \tag{A.19}$$

Then, we consider rotation matrix $\mathbf{R}$.

$$[\mathbf{t}]_\times \mathbf{R} = \mathbf{U}diag(1, 1, 0)\left(\mathbf{S}^T\mathbf{U}^T\mathbf{R}\right) \tag{A.20}$$

where

$$\mathbf{V}^T = \mathbf{S}\mathbf{U}^T\mathbf{R} = \mathbf{S}^T\mathbf{U}^T\mathbf{R} = -\mathbf{S}\mathbf{U}^T\mathbf{R} \tag{A.21}$$

$\mathbf{S}$ and $\mathbf{U}$ are orthogonal matrices, therefore $\mathbf{S}\mathbf{S}^T = \mathbf{I}$ and $\mathbf{U}\mathbf{U}^T = \mathbf{I}$. This is a SVD of $\mathbf{E}$, in which two singular values are equal and the third one is zero. Since $\mathbf{V}^T = -\mathbf{S}\mathbf{U}^T\mathbf{R}$, Equation A.6 is the same to Equation 4.15. By ignoring the scale factor λ in Equation A.6, we obtain

$$[\mathbf{t}]_\times = \mathbf{U}\mathbf{S}\mathbf{U}^T \tag{A.22}$$

Equation A.6 is equivalent to the following equation for rotation:

$$\mathbf{R} = \mathbf{U}\mathbf{S}^T\mathbf{V}^T \quad or \quad \mathbf{R}' = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{A.23}$$

For the translation vector we have

$$\mathbf{T} = \mathbf{U}\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = U_3 \tag{A.24}$$

where U_3 is the third column of $\mathbf{U}$. However, the sign of $\mathbf{E}$ cannot be determined as

$$\mathbf{T} = U_3 \quad \text{or} \quad \mathbf{T}' = -U_3 \tag{A.25}$$

A.7 Bundle Adjustment

Bundle adjustment [29] is the problem of refining a visual reconstruction to produce a jointly optimal 3-D structure and the bundles of light rays leaving each 3-D feature and converging on each camera center. The bundle of rays are parameterized by the camera calibration process. Once the camera geometry and the scene geometry have been obtained for the whole sequence, the error between them should be minimized by adjusting the camera parameters or structures. When the image error is zero-mean Gaussian, bundle adjustment is the *maximum likelihood estimator*.

Projective bundle adjustment : The 3-D point $\tilde{\mathbf{M}}_i$ and camera projection $\mathbf{P}_k$ including some errors can be refined through a global minimization of the reprojection error for the image features $\tilde{\mathbf{m}}_\mathbf{i}$:

$$\min_{\mathbf{P'}_k, \tilde{\mathbf{m}}_i} \sum_{k=1}^{m} \sum_{i=1}^{n} D(\tilde{\mathbf{m}}_{ki}, \mathbf{P}_k^{-1} \tilde{\mathbf{M}}'_i)^2 \tag{A.26}$$

This minimizes the mean squared distances between the observed image points $\tilde{\mathbf{m}}_{ki}$ and the reprojected image points $\tilde{\mathbf{m}}'_{ki}$ where $D(\tilde{\mathbf{m}}, \tilde{\mathbf{m}}')$ is the Euclidean distance. When the projective bundle adjustment refines camera parameters, then the ambiguity will be restricted to metric components [30].

Metric bundle adjustment: For high accuracy the recovered metric structure should be refined using the bundle adjustment. If one assumes that the error is only due to mislocalization of the image correspondences and that this error is uniform and normally distributed, the bundle adjustment corresponds to a maximum likelihood estimator. The camera model should be general enough so that no systematic errors remain in the data. The relative distances or angles in the scene can be used to obtain the metric information. Assume that points $\tilde{\mathbf{M}}'_i$ are the metric coordinates of the projectively reconstructed points $\tilde{\mathbf{M}}_i$, then the transformation T that updates the reconstruction from projective to metric can be obtained as

$$\lambda \tilde{\mathbf{M}}'_i \sim T \tilde{\mathbf{M}}_i \tag{A.27}$$

where λ is an arbitrary scale. By eliminating λ the equation can be rewritten as linear equations. For the image I_k including the position of the known control point $\tilde{\mathbf{M}}_i$, bundle adjustment globally minimizes the reprojection error as

$$\sum_{i=1}^{n} \sum_{I_k(u,v)} D(u_k, \frac{P_{k1}\tilde{\mathbf{M}}_i}{P_{k3}\tilde{\mathbf{M}}_i})^2 + D(v_k, \frac{P_{k2}\tilde{\mathbf{M}}_i}{P_{k3}\tilde{\mathbf{M}}_i})^2 \tag{A.28}$$

where $\mathbf{P}_k = \left[P_{k1}^T P_{k2}^T P_{k3}^T\right]^T = \mathbf{C}_k \left[\mathbf{R}_k^T | - \mathbf{R}_k^T \mathbf{t}_k\right]$.

Bibliography

[1] B. L. Anderson, and K. Nakayama, "Towards a general theory of stereopsis: binocular matching, occluding contours and fusion," *Psychological Review*, vol. 101, no. 3, pp. 414-445, 1994.

[2] S. Palmer, *Vision Science*, MIT Press, Cambridge, Massachusetts, USA, 1999.

[3] L. Li and J. H. Duncan, "3-D Translational Motion and Structure from Binocular Image Flows", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 7, pp. 657-667, 1993.

[4] R. I. Hartley, "Theory and Practice of Projective Rectification," *International Journal of Computer Vision*, vol. 35, no. 2, pp.115-127, 1999.

[5] R. I. Hartley and P. Sturm, "Triangulation," *Computer Vision and Image Understanding*, vol. 68, no. 2, pp.146-157, 1997.

[6] B. Witmer and M. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence: Teleoperators and Virtual Environments*, vol. 7, pp. 225-240, 1998.

[7] W. IJsselsteijn, H. de Ridder and J. Freeman, "Effects of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence," *Presence: Teleoperators and Virtual Environments*, vol. 10 pp. 298-311, 2001.

[8] M. Slater, A. Steed and Y. Chrysanthou, *Computer Graphics and Virtual Environments*, Addison-Wesley Publishing Company, Inc., 2002.

[9] W. IJsselsteijn, H. de Ridder and J. Vliegen, "Subjective evaluation of stereoscopic images: Effects of camera parameters and display duration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, pp. 225-233, 2000.

[10] The Wachowski Brothers, "The Matrix," Science Fiction Action Film, Warner Bros. Entertainment, Inc., Village Roadshow Pictures, 1999.

[11] Andrzej Bartkowiak, "Romeo Must Die," Warner Bros. Entertainment, Inc.,2000.

[12] M. Levoy, P. Hanrahan, "Light field rendering," *in Proceedings of ACM SIGGRAPH'96*, pp. 31-42, 1996.

[13] H. Y. Shum, L. W. He, "Rendering with concentric mosaics," *in Proceedings of ACM SIGGRAPH'99*, pp. 299-306, 1999.

[14] S. J. Gortler, R. Grzesczuk, R. Szeliski, and M. F. Cohen, "The Lumigraph," *in Proceedings of ACM SIGGRAPH'96*, pp. 43-54, 1996.

[15] D. Wood, D. Azuma, W. Aldinger, B. Curless, T. Duchamp, D. Salesin, and W. Stuetzle, "Surface Light Fields for 3D Photography," in Proceedings of ACM SIGGRAPH 2000.

[16] C. Zhang, T. Chen, "A self-reconfigurable camera array," *in Proceedings of 2004 Eurographics Symposium on Rendering*, 2004.

[17] Wilburn, B., N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," *in Proceedings of ACM SIGGRAPH, ACM Transactions on Graphics*, vol. 24, no. 3, pp. 765-776, 2005.

[18] T. H. Huang and A. N. Netravali, "Motion and structure from feature correspondences: a review," *in Proceedings of IEEE*, vol. 82, no. 2, pp.252-268, 1994.

[19] M. I. A. Lourakis, A. A. Argyros, "Vision based Camera Motion Recovery for Augmented Reality," *in Proceedings of Computer Graphics International '04*, pp. 569-576, 2004.

[20] M. Goesele, B. Curless, and S. Seitz, "Multi-view stereo revisited," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition '06*, 2006.

[21] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," *in Proceedings of European Conference on Computer Vision 2002*, vol. 3, pp. 82-96, 2002.

[22] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multi-view stereopsis," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition '06*, 2007.

[23] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry, *An Invitation to 3D Vision: From Images to Geometric Models*, Springer Verlag, 2003.

[24] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, pp. 133-135, 1981.

[25] M. Black and P. Anandan, "Robust dynamic motion estimation over time," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition '91*, pp. 299-302, 1991.

[26] A. Spoerri and S. Ullman, "The early detection of motion boundaries," *in Proceedings of IEEE International Conference on Computer Vision '87*, pp. 209-218, 1987.

[27] M. Irani and B. Rousso and S. Peleg, "Robust Recovery of Ego-Motion," *in Proceedings of the 5th International Conference on Computer Analysis of Images and Patterns '93*, 1993.

[28] P. Sturm, S. Ramalingam, "A Generic Concept for Camera Calibration," *in Proceedings of the European Conference on Computer Vision '04*, vol.2. pp. 1-13, 2004.

[29] B. Triggs, P. McLauchlan, R. I. Hartley, and A. Fitzgibbon, "Bundle Adjustment - A modern synthesis," *in Proceedings of Workshop on Vision Algorithms: Theory and Practice*, 2000.

[30] B. Micusik and T. Pajdla, "3D metric reconstruction from uncalibrated omnidirectional images," *in Proceedings of Asian Conference on Computer Vision*, 2004.

[31] A. Yao and A. Calway, "Dense 3-D Structure from Image Sequences Using Probabilistic Depth Carving," *in Proceedings of the 14th British Machine Vision Conference '03*, pp. 211-220, 2003.

[32] T. Sato, M. Kanbara, N. Yokoya, H. Takemura, "Dense 3-D Reconstruction of an Outdoor Scene by Hundreds-Baseline Stereo Using a Hand-Held Video Camera," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 119-129, 2002.

[33] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics the Approach Based on Influence Functions*, John Wiley and Sons, New York, 1986.

[34] P. J. Huber, *Robust Statistics*, John Wiley and Sons, New York, 1981.

[35] Middebury Stereo (`http://cat.middlebury.edu/stereo/`).

[36] J. R. Ohm, E. Izquierdo, "An Object-Based System for Stereoscopic Viewpoint Synthesis," *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 5, no. 5, pp. 801-811, 2000.

[37] C. Cafforio, F. Rocca and S. Tubaro, "Motion Compensated Image Interpolation," *IEEE Transactions on Communications*, vol.38, no.2, pp. 215-222, 1990.

[38] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47 no, 1/2/3, pp. 7-42, 2002.

[39] T. Kanade, P. W. Rander, P. J. Narayanan, "Virtualized Reality: Constructing virtual worlds from real scenes," *IEEE Multimedia Magazine*, vol. 1, no. 1, pp.34-47, 1997.

[40] J. Carranza, C. Theobalt, M. Magnor, H. P. Seidel, "Free-viewpoint video of human actors," in Proceedings of ACM SIGGRAPH '03, pp. 569-577, 2003.

[41] B. Wilburn, M. Smulski, H. H. K. Lee, M. Horowitz, "The light field video camera," *in Proceedings of Media Processors 2002*, *SPIE Electronic Imaging*, 2002.

[42] T. Kanade, H. Saito, S. Vedula, "The 3D room: Digitizing time-varying 3D events by synchronized multiple video streams," Technical Report, CMU-RITR-98-34, 1998.

[43] T. Fujii, K. Mori, K. Takeda, K. Mase, M. Tanimoto, Y. Suenaga, "Multipoint measuring system for video and sound: 100-camera and microphone system," IEEE 2006 International Conference on Multimedia and Expo, pp. 437-440, 2006.

[44] P. Kauff, N. Atzpadin, C. Fehn, M. Muller, O. Schreer, A. Smolic, R. Tanger, "Depth map creation and image based rendering for advanced 3DTV services providing interoperability and scalability," *Signal Processing Image Communication, Special Issue on 3DTV*, 2007.

[45] S. Wurmlin, E. Lamboray, M. Gross, "3D video fragments: Dynamic point samples for real-time free-viewpoint video," Computer Graphics, Special issue on coding, compression and streaming techniques for 3D and multimedia data, vol.28, no. 1, pp. 3-14, 2004.

[46] M. Waschbusch, St.Wurmlin, D. Cotting, F. Sadlo, M. Gross, "Scalable 3D Video of Dynamic Scenes," *the Visual Computer*, vol. 21, no. 8-10, pp. 629-638, 2005.

[47] R. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," IEEE International Journal Robotics and Automation, vol. 3, no. 4, pp. 323-344, 1987.

[48] J. Heikkila and O. Silven, "A four-step camera calibration procedure with implicit image correction," *in Proceedings on IEEE conference on Computer Vision and Pattern Recognition '97*, 1997.

[49] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330-1334, 2000.

[50] M. Pollefeys, R. Koch and L. Van Gool, "Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters," *in Proceedings of IEEE International Conference on Computer Vision*, pp. 90-95, 1997.

[51] A. Heyden, and K. Aström, "Euclidean Reconstruction from Image Sequences with Varying and Unknown Focal Length and Principal Point," *in Proceedings of IEEE conference on Computer Vision and Pattern Recognition*, pp. 438-443, 1997.

[52] A. Heyden and K. Aström, "Algebraic properties of multilinear constraints," *Mathematical Methods in Applied Sciences*, vol. 20, no. 13, pp. 1135-1162, 1997.

[53] O. Faugeras, Q. T. Luong, and S. Maybank, "Camera selfcalibration: Theory and experiments," *in Proceedings of European Conference on Computer Vision '92*, Lecture Notes in Computer Science, vol. 588, Springer-Verlag, pp. 321-334, 1992.

[54] A. Heyden and K. Aström, "Euclidean Reconstruction from Constant Intrinsic Parameters," *in Proceedings of IEEE conference on Computer Vision and Pattern Recognition*, pp. 339-343, 1996.

[55] R. Hartley, "Euclidean reconstruction from uncalibrated views," *Applications of Invariance in Computer Vision*, Lecture Notes in Computer Science, vol. 825, pp. 237-256, 1994.

[56] M. Pollefeys, L. Van Gool and A. Oosterlinck, "The Modulus Constraint: A New Constraint for Self-Calibration," in Proceedings on the 13th IEEE conference on Computer Vision and Pattern Recognition, pp. 349-353, 1996.

[57] B. Triggs, "The Absolute Quadric," *in Proceedings on IEEE Conference on Computer Vision and Pattern Recognition*, pp. 609-614, 1997.

[58] C. S. Fraser, "Digital camera self-calibration," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 52, pp.149-159, 1997.

[59] A. Gruen and H. A. Beyer, "System calibration through self-calibration," in *Calibration and Orientation of Cameras in Computer Vision*, Gruen and Huang (Eds.), Springer Series Information Sciences 34, pp. 163-194

[60] VIRTUE Home European Union's Information Societies Technology Programme, Project IST 1999.10 044. British Telecom.

[61] A. Goshtasby, S. H. Gage, J. F. Bartholic, "A two stage cross correlation approach to template matching," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 6, pp. 374-378, 1984.

[62] C. H. Chou, Y. C. Chen, "Moment-preserving pattern matching," *Pattern Recognition*, vol. 23, no. 5, pp. 461-474, 1990.

[63] J. K. Cheng and T. S. Huang, "Image registration by matching relational structures," *Pattern Recognition*, vol. 17, no. 1, pp. 149-159, 1984.

[64] R. Horaud, T. Skordas, "Stereo correspondence through feature grouping and maximal clique," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 1168-1180, 1989.

[65] S. Ullman, "The Interpretation of Visual Motion," *Cambridge, MA: MIT Press*, 1989.

[66] D. Tell, S. Carlsson, "Combining appearance and topology for wide baseline matching," *in Proceedings on European Conference on Computer Vision*, vol. 1, pp. 68-81, 2002.

[67] A. Rojas, A. Calvo, J. Munoz, "A Dense Disparity Map of Stereo Images," *Pattern Recognition Letters*, vol. 18, no. 4, pp. 385-393, 1997.

[68] S. T. Birchfield, "Depth Discontinuities by Pixel-to-Pixel Stereo", *International Journal of Computer Vision*, vol.35, no. 3 pp. 269–293, 1999.

[69] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, R. Szeliski, "High-quality video view interpolation using a layered representation," in Proceedings of ACM SIGGRAPH '04, ACM Transactions on Graphics, pp. 600-608.

[70] P. Sangteanchai and S. Madarasami, "A use of discontinuity elements in coarse and fine stereo matching," *in Proceedings of the 3rd IASTED International Conference on Advances in Computer Science and Technology*, pp. 218-221, 2007.

[71] C. Zitnick and T. Kanade, "A Cooperative Algorithm for Stereo Matching and Occlusion Detection," *Technical Report, CMU-RI-TR-99-35*, Robotics Institute, Carnegie Mellon University, October, 1999.

[72] O. Veksler, "Fast variable window for stereo correspondence using integral images," *in Proceedings of IEEE conference on Computer Vision and Pattern Recognition '03*, 2003.

[73] M. Agrawal and L. Davis, "Window-based discontinuity preserving stereo," *in Proceedings of IEEE conference on Computer Vision and Pattern Recognition '04*, 2004.

[74] T. Kanade , M. Okutomi, "A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 9, pp. 920-932, 1994.

[75] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade, "Three-Dimensional Scene Flow," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 475-480, 2005.

[76] J. Davis, D. Nehab, R. Ramamoorthi, S. Rusinkiewicz, "Spacetime Stereo: A Unifying Framework for Depth from Triangulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 296-302, 2005.

[77] K. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette across time: Part I: Theory and algorithm," *International Journal of Computer Vision*, vol. 62, no. 3, pp. 221-247, 2005.

[78] K. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette across time: Part II: Theory and algorithm," *International Journal of Computer Vision*, vol. 62, no. 3, pp. 221-247, 2005.

[79] Marc Levoy, "Light Fields and Computational Imaging," *IEEE Computer*, 2006.

[80] S. J. Gortler, R. Grzeszczuk, R. Szeliski and M. F. Cohen, "The Lumigraph," *in Proceedings of ACM SIGGRAPH'96*, pp 43-54, 1996.

[81] L. C. Evans, *Partial Differential Equations*, American Mathematical Society, Providence, 1998. ISBN 0-8218-0772-2.

[82] C. Zhang, T. Chen, "Generalized plenoptic sampling," *Technical Report*, Advanced Multimedia Processing Laboratory, Carnegie Mellon University, September 2001.

[83] R. Ng, "Fourier Slice Photography" in Proceedings of ACM SIGGRAPH 2005, ACM Press, pp. 735-744, 2005.

[84] Chai, J., Chan, S., Shum, H., and Tong, X, "Plenoptic sampling," *in Proceedings of the 27th annual conference on Computer graphics and interactive techniques, International Conference on Computer Graphics and Interactive Techniques*, pp. 307-318, 2000.

[85] J. Shi and C. Tomasi, "Good Features to Track," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition '94*, pp. 593-600, 1994.

[86] S. Birchfield, "Derivation of Kanade-Lucas-Tomasi Tracking Equation," Unpublished, Notes, 1997.

[87] T. A. Clarke and J. G. Fryer, "The Development of Camera Calibration Methods and Models" *Photogrammetric Record*, vol. 16, no. 91, pp.51-66, 1998.

[88] M. Hobbema, *The Alley at Middelharnis*, Oil on canvas, 103,5 x 141cm, National Gallery, London, 1689.

[89] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, New York, 2000.

[90] C. Zitnick and et. al., "High-quality video view interpolation using a layered representation," *in Proceedings of ACM SIGGRAPH '04, ACM Transaction on Graphics*, 2004.

[91] 3DTV Network of Excellence, funded by the European Commission 6th Framework Information Society Technologies Programme. (`http://www.3dtv-research.org`)

[92] K. Kraus, "Photogrammetric triangulation," *Photogrammetry*, vol. 1, 4th Edition, Dummler Verlag, Bonn, pp.277-279, 1993.

[93] A. Agrawal, R. Chellappa, "Robust Ego-Motion Estimation and 3D Model Refinement using Surface Parallax," *IEEE Transactions on Image Processing*, Vol. 15, No. 5, 2006.

[94] F. Fraundorfer, H. Bischof, "Detecting distinguished regions by saliency," *in Proceedings of the 13th Scandinavian Conference on Image Analysis*, pp. 208-215, 2003.

[95] J. Matas, O. Chum, M. Urban, I. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," *in Proceedings of the 13th British Machine Vision Conference*, pp. 384-393, 2002.

[96] O. Schreer and P. Kauff, "An Immersive 3D Video-Conferencing System Using Shared Virtual Team User Environments", *in Proceedings of ACM Collaborative Virtual Environments*, pp. 105-112, 2002.

[97] A. Shashua and S. Avidan, "The Rank-4 Constraint in Multiple View Geometry," *in Proceedings of European Conference on Computer Vision '96*, vol. 2, pp. 196-206, 1996.

[98] A. Shashua and N. Navab, "Relative Affine Structure: Canonical Model for 3D from 2D Geometry and Applications," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 18, no. 9, pp.873-883, 1996.

[99] C. T. Loop and Z. Zhang, "Computing rectifying homographies for stereo vision." *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition '99*, pp. 1125-1131, 1999.

[100] A. Fusiello, E. Trucco, and A. Verri. "A compact algorithm for rectification of stereo pairs," *in Proceedings of Machine Vision and Applications*, vol. 12, no. 1, pp.16-22, 2000.

[101] M. Pollefeys, R. Koch, and L. Van Gool, "A simple and efficient rectification method for general motion," *in Proceedings of International Conference on Computer Vision '99*, pp.496-501, 1999.

[102] D. Oram, "Rectification for any epipolar geometry." *in Proceedings of the British Machine Vision Conference '01*, 2001.

[103] Oxford University Multiview Image Database (`http://www.robots.ox.ac.uk/~vgg/data/`)

[104] B. Chebaro, A. Crouzil, L. Massip-Pailhes, and S. Castan, "Fusion of the stereoscopic and temporal matching results by an algorithm of coherence control and conflicts management," *in Proceedings of International Conference on Computer Analysis of Images and Patterns*, Lecture Notes In Computer Science, Vol. 719, pp.486-493, 1993.

[105] A. M. Waxman and J. H. Duncan, "Binocular image flows: Steps toward stereo-motion fusion," *IEEE Transactions on Pattern Analysis and Machine Vision*, vol. 8, no. 6, pp.715-729, 1986.

[106] Z. Zhang and G. Xu, "A General Expression of the Fundamental Matrix for Both Perspective and Affine Cameras," *in Proceedings of the 15th International Joint Conference on Artificial Intelligence*, pp. 1502-1507, 1997.

[107] J. V. Miller and C. V. Stewart, "MUSE: Robust surface fitting using unbiased scale estimates," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition '96*, pp. 300-306, 1996.

[108] K. Kanatani, *Statistical Optimization for Geometric Computation: Theory and Practice*, Elsevier, 1996.

[109] B. Tordoff and D. Murray, "Guided sampling and consensus for motion estimation," *in Proceedings of the 7th European Conference on Computer Vision*, vol. 1, pp. 82-96, 2002.

[110] P. Pritchett and A. Zisserman, "Wide baseline stereo matching," in Proceedings of the 6th International Conference on Computer Vision, pp. 754-760, 1998.

[111] P. H. S. Torr and A. Zisserman, "MLESAC: A new robust estimator with application to estimating image geometry," *Computer Vision and Image Understanding*, vol. 78, pp. 138-156, 2000.

[112] M. Irani and P. Anandan, "Robust multi-sensor image alignment," *in Proceedings of the 4th International Conference on Computer Vision*, pp. 959-966, 1998.

[113] S. Ayer, P. Schroeter, and J. Bigum, "Segmentation of moving objects by robust motion parameter estimation over multiple frames," *in Proceedings of the 3rd European Conference on Computer Vision*, vol. 2, pp. 316-327, 1994.

[114] D. Comaniciu and P. Meer, "Robust analysis of feature spaces: Color image segmentation," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 750-755, 1997.

[115] P. Meer and B. Georgescu, "Edge detection with embedded confidence," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 23, pp. 1351-1365, 2001.

[116] P. Meer, S. Wang, and H. Wechsler, "Edge detection by associative mapping," *Pattern Recognition*, vol. 22, pp.491-503, 1989.

[117] G. Li, "Robust regression," in D. C. Hoaglin, F. Mosteller, and J. W. Tukey, (Eds.), *Exploring Data Tables, Trends, and Shapes, JohnWiley and Sons*, pp. 281-343, 1985.

[118] B. Mendes and D. E. Tyler, "Constraint M-estimation for regression," *Robust Statistics, Data Analysis, and Computer Intensive Methods*, vol. 1, pp. 299-320, 1996.

[119] P. J. Rousseeuw, "Least median of squares regression," *Journal of Americal Statistic Association*, vol. 79, pp. 871-880, 1984.

[120] P. H. S. Torr and D. W. Murray, "The development and comparison of robust methods for estimating the fundamental matrix," *International Journal of Computer Vision*, vol. 24, no. 3, pp. 271.300, 1997.

[121] P. H. S. Torr, A. Zisserman, and S. J. Maybank, "Robust detection of degenerate configurations while estimating the fundamental matrix," *Computer Vision and Image Understanding*, vol. 71, pp. 312-333, 1998.

[122] F. Schaffalitzky and A. Zisserman, "Viewpoint invariant texure matching and wide baseline stereo," *in Proceedings of the 8th International Conference on Computer Vision*, vol. 2, pp. 636-643, 2001.

[123] R. D. Martin, V. J. Yohai, and R. H. Zamar, "Min-max bias robustness," *Annals of Statistics*, vol. 17, pp. 1631-1661, 1989.

[124] R. H. Zamar, "Robust estimation in the errors-in-variables model," *Biometrika*, vol. 76, pp. 149-160, 1989.

[125] P. H. S. Torr and C. Davidson, "Impsac: Synthesis of importance sampling and random sample consensus," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 25, pp. 354-364, 2003.

[126] J. Canny, "Computational approach to edge detection", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol G 8, pp. 679-698, 1986.

[127] A. P. Witkin, "Scale-space filtering", *in Proceeding of 8th International Joint Conference on Artificial Intelligence*, pp. 1019-1022, 1983.

[128] R. Deriche, "Optimal edge detection using recursive filtering", *in Proceeding of IEEE International Conference on Computer Vision*, pp. 501-505, 1987.

[129] W. Ma, B. S. Manjunath, "EdgeFlow: a technique for boundary detection and image segmentation", *IEEE Transaction on Image Processing*, pp. 1375-1388, 2000.

[130] D. Tschumperle, R. Deriche, "Vector-valued image regularization with PDEs: a common framework for different applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp.506-517, 2005.

[131] B. Sumengen and B. S. Manjunath, "Multi-scale Edge Detection and Image Segmentation," *in Proceeding of European Signal Processing Conference*, 2005.

[132] Y. Dufournau, C. Schmid, R. Horaud, "Image matching with scale adjustment", *Computer Vision and Image Understanding*, vol. 93, 2004, page 175-194.

[133] C. Menard, W. G. Kropatsch, "Adaptive Stereo Matching in Correlation Scale-Space", *in Proceedings of the 9th IAPR International Conference on Image Analysis and Processing*, vol. 1, pp. 677-684, 1997.

[134] J. C. Tilton, "Image segmentation by region growing and spectral clustering with a natural convergence criterion," *in Proceedings IEEE International Geoscience and Remote Sensing Symposium '98*, Vol. 4, pp. 1766-1768, 1998.

[135] E. Zaharescu, M. Zamfir, C. Vertan, "Color morphology-like operators based on color geometric shape characteristics," *International Symposium on Signals, Circuits and Systems*, Vol. 1, PP. 145-148, 2003.

[136] J. M. Gauch, "Image segmentation and analysis via multiscale gradient watershed hierarchies," *IEEE Transactions on Image Processing*, Vol. 8, pp. 69-79, 1999.

[137] M. J. Black, G. Sapiro, D. H. Marimont, D. Heeger, "Robust anisotropic diffusion", *IEEE Transactions on Image Processing*, vol. 7, pp. 421-432, 1998.

[138] S. K. Weeratunga, C. Kamath, "A comparison of PDES-based non-linear anisotropic diffusion techniques for image denoising", *in Proceeding of SPIE Electronic Imaging, Image Processing: Algorithms and Systems II*, 2003.

[139] J. Weickert, T. Brox, "Diffusion and regularization of vector- and matrix-valued images," in M. Z. Nashed, O. Scherzer (Eds.): *Inverse Problems, Image Analysis, and Medical Imaging. Contemporary Mathematics*, vol. 313, pp. 251-268, 2002.

[140] J. Weickert, C. Feddern, M. Welk, B. Burgeth, T. Brox, "PDEs for tensor image processing," in J. Weickert, H. Hagen (Eds.): *Visualization and Processing of Tensor Fields*, 399-414, Springer, Berlin, 2006.

[141] B. Tang, G. Sapiro, V. Caselles, "Color Image Enhancement via Chromaticity Diffusion," *IEEE Trans. on Image Processing*, pp. 701-707, 2001.

[142] O. Shahar, S. Zucker, "Hue Geometry and Horizontal Connections," *Neural Networks*, vol. 17 pp. 753-771, 2004.

[143] P. Perona, J. Malik, "Scale-space and edge detection using anisotropic diffusion", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 629-639, 1990.

[144] P. Trahanias, A. Venetsanopoulos, "Vector Directional Filters: A New Class of Multichannel Image Processing Filters," *IEEE Transactions on Image Processing*, vol. 2, pp. 528-534, 1993.

[145] K. Plataniotis, D. Androutsos, A. Venetsanopoulos, "Color image processing using adaptive vector directional filters," *IEEE Transactions on Circuits and System*, vol. 45, pp. 1414-1419, 1998.

[146] V. Caselles, G. Sapiro, D. H. Chung, "Vector median filters, inf-sup operations, and coupled PDEs: Theoretical connections," *Jour. Math. Imag. Vis.*, vol. 12, pp. 109-120, 2000.

[147] B. Smolka, M. Szczepanski, K. N. Plataniotis, N. Venetsanopoulos, "On the fast modification of the vector median filter," *in Proc. of Int. Conf. on Pattern Recognition*, vol.3, pp. 931-934, 2002.

[148] J. Kim and T. Sikora, "Color Image Noise Reduction using Perceptual Maximum Variation Modeling for Color Diffusion," *in Proceedings of International Workshop on Image Analysis for Multimedia Interactive Services '06*, 2006.

[149] International Color Consortium, *Specification ICC.1:2004-10 (Profile version 4.2.0.0) Image technology colour management - Architecture, profile format, and data structure*, 2006.

[150] M. Drew and G. Hamarneh, "Visualizing Diffusion Tensor Dissimilarity using an ICA Based Perceptual Color Metric," *in Proceedings on Color Imaging, IS&T/SID's Conference*, pp. 42-47, 2007.

[151] M. Agrawal and L. Davis, "Window based, discontinuity preserving stereo," *in Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 66-73, 2003.

[152] S. Birchfield and C. Tomasi, "Depth Discontinuities by Pixel-to-Pixel Stereo," *International Journal of Computer Vision*, vol. 35, pp. 269-293, 1999.

[153] Y. Yang, A. Yuille, and J. Lu, "Local, global, and multilevel stereo matching," *in Proceedings of IEEE International Conferernce on Computer Vision and Pattern Recognition*, pp. 274-279, 1993.

[154] L. Alvarez, R. Deriche, J. Sánchez, and J. Weickert, "Dense disparity map estimation respecting image derivatives: a PDE and scale-space based approach," *Journal of Visual Communication and Image Representation*, vol. 13, pp. 96-114, 2002.

[155] C. Strecha and L. Van Gool, "PDES-based multi-view depth estimation," *in Proceedings of International Symposium on 3D Data Processing Visualization and Transmission*, pp. 416-425, 2002.

[156] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 1222-1239, 2001.

[157] M. Bleyer and M. Gelautz, "A layered stereo algorithm using image segmentation and global visibility constraints," *in Proceedings of International Conference on Image Processing*, pp. 2997-3000, 2004.

[158] Q. Yang, L. Wang, R. Yang, S. Wang, M. Liao, D. Nistér, "Real-time Global Stereo Matching Using Hierarchical Belief Propagation," *British Machine Vision Conference '06*, pp.986, 2006.

[159] J. Kim and T. Sikora, "Gaussian Scale-Space Dense Disparity Estimation with Anisotropic Disparity-Field Diffusion," *in Proceedings of IEEE International Conference on 3-D Digital Imaging and Modeling*, 2005.

[160] J. Kim and T. Sikora, "Robust Anisotropic Disparity Estimation with Perceptual Maximum Variation Modeling," *in Proceedings of IEEE International Conference on Image Processing '06*, 2006.

[161] R. I. Hartley, "In Defense of the Eight-Point Algorithm," IEEE Transaction on Pattern Recognition and Machine Intelligence, vol. 19 no. 6, pp. 580-593, 1997.

[162] J. Kim and T. Sikora "Hybrid Recursive Energy-based Method for Robust Optical Flow on Large Motion Fields," *in Proceedings of IEEE International Conference on Image Processing '05*, 2005.

[163] N. Atzpadin, P. Kauff and O. Schreer, "Stereo Analysis by Hybrid Recursive Matching for Real-Time Immersive Video Conferencing," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 14, No. 3, pp. 321-334, 2004.

[164] M. J. Black and P. Anandan, "The robust estimation of multiple motions: parametric and piecewise smooth flow fields," *Computer Vision and Image Understanding*, Vol. 63, No. 1 pp. 75-104, 1996.

[165] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," *International Journal of Computer Vision*, Vol. 2, No. 3, pp. 283-310, 1989.

[166] L. Alvarez, R. Deriche, T. Papadopoulo and J. Sánchez, "Symmetrical Dense Optical Flow Estimation with Oclussions Detection", *in Proceeding of European Conference on Computer Vision*, pp. 721-736, 2002.

[167] A. Bruhn, J. Weickert and C. Schnörr, "Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods," *International Journal of Computer Vision*, Vol. 61, No. 3, pp. 211-231, 2005.

[168] E. Ong, and M. Spann, "Robust optical flow computation based on least-median-of squares regression," *International Journal of Computer Vision*, Vol. 31, No. 1, pp. 51-82, 1999.

[169] C. V. Stewart, "Robust parameter estimation in computer vision," *SIAM Reviews*, vol. 41, no. 3, pp. 513-537, 1999.

[170] D. Gibson and M. Spann, "Robust optical flow estimation based on a sparse motion trajectory set," *IEEE Transactions on Image Processing*, Vol. 12, pp. 431-445, 2003.

[171] T. Brox, A. Bruhn, N. Papenberg and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," *in Proceeding of European Conference on Computer Vision '04*, pp. 25-36, 2004.

[172] Image sequence server of Universität Karlsruhe. (`http://i21www.ira.uka.de/image_sequences/`)

[173] M. Subbarao and G. Surya, "Depth from defocus: a spatial domain approach," *International Journal of Computer vision*, vol. 13, no. 3, pp.271-294, 1993.

[174] P. Favaro, S. Osher, S. Soatto, and L. Vese, "3d shape from anisotropic diffusion," *in Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 179-186, 2003.

[175] V. Bove, "Probabilistic method for integrating multiple sources of range data," *Journal of the Optical Society of America*, vol. 7, no. 12, pp. 2193-2198, 1990.

[176] A. Rajagopalan, S. Chaudhuri, and U. Mudenagudi, "Depth estimation and image restoration using defocused stereo pairs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1521-1525, 2004.

[177] C. Frese and I. Ghet, "Robust Depth Estimation by Fusion of Stereo and Focus Series Acquired with a Camera Array," *in Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2006.

[178] A. Cavallaro, T. Ebrahimi, "Object-based video: extraction tools, evaluation metrics and applications," *in Proceedings of SPIE Conference on Visual Communications and Image Processing '03*, 2003.

[179] L. Lucchese, S. Mitra, "Color image segmentation: A State-of-the-Art survey," *in Proceedings of the Indian National Science Academy*, 2001.

[180] R. Gvili, A. Kaplan, E. Ofek, G. Yahav, "Depth keying," *in Proceedings of SPIE Stereoscopic Displays and Virtual Reality Systems*, 2003.