YORK UNIVERSITÉ UNIVERSITY

redefine THE POSSIBLE.

A Review of Stereo and Motion Integration: Stereomotion

Mikhail Sizintsev

Technical Report CSE-2009-06

October 26 2009

Department of Computer Science and Engineering

4700 Keele Street Toronto, Ontario M3J 1P3 Canada

A review of stereo and motion integration: Stereomotion

Mikhail Sizintsev

Department of Computer Science and Engineering
and the Centre for Vision Research
York University
Toronto, Ontario M3J 1P3
Canada

October 22, 2009

Abstract

This report provides an in-depth overview of the computer vision technique known as *stereomotion* that concerns the estimation of structure and motion from binocular video sequences. First, the modalities of separate stereo and motion estimation are discussed in light of advantages and shortcomings with respect to each other – as a result, the set of anticipated benefits of stereo and motion fusion is stated. Second, the current stereomotion literature is reviewed along three major directions: rigid structure from motion; non-rigid scene and 3D flow; temporally-coherent and spacetime stereo. Third, performance characteristics of state-of-the-art techniques are presented and compared. Finally, the report concludes by looking at open problems and speculations on future research directions.

Contents

List of Abbreviations

BCC	Brightness Constancy Constraint
BP	Belief Propagation
CTF	Coarse-to-fine
DP	Dynamic Programming
GC	Graph Cuts
KF	Kalman Filter
EKF	Extended Kalman filter
IEKF	Iterated Extended Kalman Filter
IF	Information Filter
LS	Least Squares
SSD	Sum of Squared Differences
OFCE	Optical Flow Constraint Equation
TRW	Tree-Reweighted message passing
QBPO	Quadratic Pseudo-Boolean Optimization

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Sensing the 3D environment: Fundamental challenge and motivation

Living creatures have a wonderful ability to sense the world around them. In an attempt to realize the underlying mechanisms in computer algorithms, we have to understanding how meaningful entities can be perceived from raw intensity patterns. But before even going there, its important by itself to deal with the following shortcoming:

Our spatial world is 3D, while humans or robots, being able to view only from a single location (point) at a certain instance of time, get only the projection of the world through this point. This projection, be it on a plane (conventional camera), or on a sphere (omnidirectional camera), is inherently 2D, and the third dimension, depth, is never directly observed. This idea is schematically shown in Fig. 1.1.

In this light, the ability to infer structure of the scene from its *multiple* views is of paramount importance in practical applications. One such application is robotics, where robots must operate in environments that are dangerous or unreachable by humans, such as space exploration, underground mining, military operation, and aids in driving. More generally, 3D information is required for a number of consumer applications such as Augmented Reality, 3DTV and user interfaces.

1.2 Multiview imaging and depth from triangulation

While there are numerous ways to reason about depth using only a single 2D image (e.g. shading [ZTCS99], texture [TV98] and perspective cues [CRZ00]), a fundamental approach to its true recovery is to overcome the single point location situation and obtain images from multiple viewpoints. This can be done by physically having more cameras or moving the camera in space. Once we establish the projections of the same 3D point in all cameras, we can easily infer its true 3D coordinate via triangulation as shown in Fig. 1.2.

In the light of this discussion, there are two fundamental ways to acquire multiple perspective views. Technically, there also exists a third way, which is a special case of the previous two.

1. Multiple stationary cameras with simultaneous capture; following standard terminology, we refer to this situation as stereo.

2. Single camera with motion relative to scene and sequential capture; following standard terminology, we refer to this situation as motion.

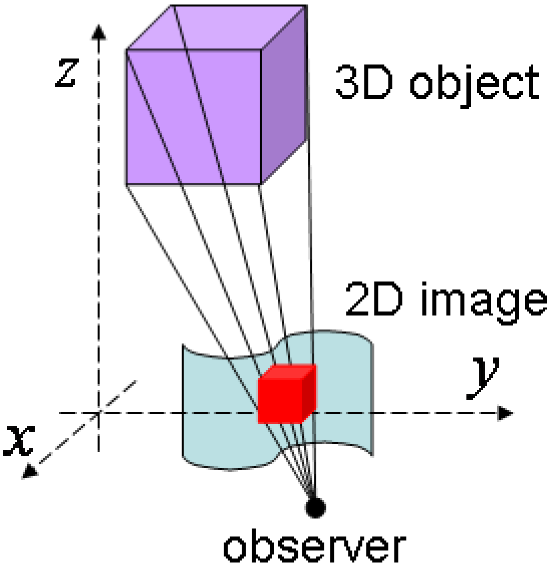


Figure 1.1: Sensing 3D world via 2D imagery. While the world has three spatial dimensions, conventional imaging yields only a 2D projection.

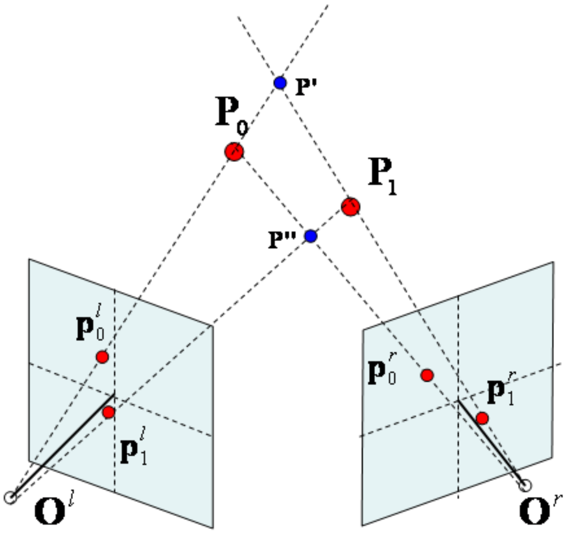


Figure 1.2: Recovering 3D structure from two 2D views. Points $\mathbf{P}_0$ and $\mathbf{P}_1$ can be easily recovered in 3D once their corresponding left and right image projections are known via the process called *triangulation* [HZ04]. Interestingly, if we confuse the projections, we end up with completely different 3D structure (e.g. $\mathbf{P}'$ and $\mathbf{P}''$ in this case) – this illustrates the paramount importance of knowing the correct correspondence

3. Single camera focused at different depths. This case is usefully analogous to camera motion along the optical axis. Interestingly, stereo systems with this particular camera configuration have been considered in the past[1] [ABG89].

Significantly, acquisition across time is necessary to recover 3D structure and dynamics of a viewed scene, while acquisition across space is necessary to exclude the dynamics effect that we do not want (like changing lighting or camera intrinsics). Considerable effort has been devoted to exploring the combination of stereo and motion, stereomotion, to yield self-sufficient techniques for reasoning about the static and dynamic behaviour of the world surrounding us.

Stereomotion using standard electro-optical cameras is a fast and reasonably efficient approach to recover information as interaction with the environment is none or minimal. In comparison, various "active methods" (e.g. lasers, projectors, sonars, or even haptics) are relatively invasive, bulky, expensive and power hungry. Furthermore, such technologies may utilize complex mechanisms which limits their ability to operate in dynamic environments and can make them mechanically fragile. Moreover, multiple active sensors of the same type can interfere with one another in conjested environments. On the other hand, the stereomotion paradigm may experience the same difficulties associated with individual stereo and motion processing, such as poor speed/accuracy trade-offs, difficulties in textureless regions and improperly reconstructing object boundaries [SW09a]. Overcoming such limitations provides key challenges in solving the stereomotion problem and inspires ongoing research.

1.3 Outline of report

This report is structured in the following way. The present Chapter has motivated this work and stated the overall objective of stereomotion. Chapter 2 describes the individual areas of stereo and motion in sufficient depth to exemplify the similarities and differences of stereo and motion to find common fruitful ground for stereomotion research. Chapter 3 outlines various matching, optimization and filtering techniques used in computer vision that are applied in the stereomotion field. Chapter 4 describes in depth the stereomotion paradigm applied in the case of rigid scenes. Chapter 5 covers the less constrained scenario of non-rigid scenes. Chapter 6 deals with a special but very important scenario when temporally-consistent estimation of structure is of interest. Chapter 7 discusses the performance of the developed stereomotion techniques and the ways this performance can be assessed. Finally, Chapter 8 ends this report with the overview of open problems and discussion of future avenues in stereomotion development.

[1] In particular, it was shown in [ABG89] that position of cameras along the optical axis resulted in reduced search range and effective half-occlusion reasoning. The obvious downside of difference in resolutions between cameras in front and back as well as the technical difficulty of simultaneous image capture made this design obsolete virtually as soon as it appeared.

Chapter 2

Stereo and Motion: Commonalities and Differences

2.1 Outline

This chapter gives a terse but in-depth overview of stereo and motion processing in computer vision. At the same time, the purpose of this chapter is not to give comprehensive reviews of these areas, for which reviews are available elsewhere [DA89, Kos93, SS02, BBH03, BFB94, Der06, BSL$^+$07]. Rather, we want to briefly set the stage for integration of these two modalities to realize their potential for complimentarily.

2.2 Stereo

2.2.1 Basics

The stereo problem is very easy to state (but, unfortunately, not easy to solve) once one considers the geometry behind it. The situation for a single 3D point and two perspective cameras [HZ04] is depicted in Figure 2.1. The basis behind the process of inferring actual depth is the search for the projection of the same 3D point across images and calculation of *disparity* – a difference in image coordinates between those projections. Once the corresponding projections are found, the absolute

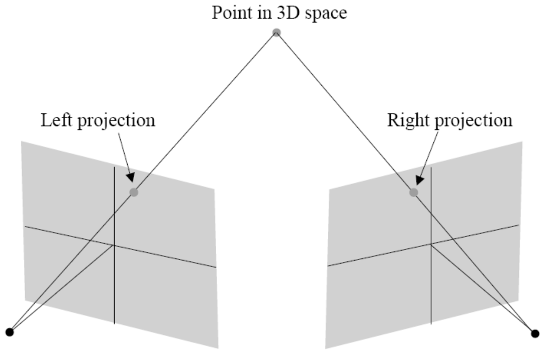


Figure 2.1: Stereo Geometry for Two Perspective Cameras. A 3D point in space is projected on two spatially displaced cameras.

3D coordinates of the world point are completely determined via triangulation, provided that the stereo rig is calibrated [HZ04]. Similarly, we can reconstruct the whole scene point by point.

Within this framework, triangulation is well understood. Calibration is usually straightforward or known in advance, and solutions up to affine or perspective transformations are possible when calibration is incomplete or unavailable [HZ04]. At the same time, correspondence remains challenging and is usually understood to be the stereo problem itself nowadays. Note that the vast majority of correspondence problem formulations explicitly or implicitly assume that matched primitives should be similar in appearances.

2.2.2 Challenges

As discussed, solving stereo essentially means solving correspondence. The correspondence problem is not easy for many reasons that are either of sensor, algorithmic, or even theoretic nature. We list most of them below.

Non-Lambertian surface

The general intuition is that points in correspondence should look alike. The definition and the concept of similarity depends on the matching entity. Talking in terms of pixels, the most widespread data in contemporary stereo, pixels of similar colour, or intensity, should belong to the same 3D point, i.e. be in correspondence. Despite its widespread use, such a simple model is only true for surfaces whose imaged brightness patterns are independent of viewpoint, e.g. so-called Lambertian surfaces[1] [TV98]. The specular component of the reflectance cannot be adequately subsumed into a simple noise model of colour-based matching, which assumes unimodal zero-mean, usually Gaussian noise. A great effort has been devoted to relaxing this “brightness constancy” assumption and various normalized measures [BBH03], rank-order statistics measures [ZW94, BN98] and entropy-based measures [Egn00, KKZ03, Hir05] have been investigated.

3D boundaries

Other fundamentally hard regions for establishing correspondences are in the vicinity of 3D boundaries. This problem is typical for computer vision processing methods that have to deal with noisy data and use low-pass filtering techniques to regularize the solution. While such methods alleviate difficulty with associated high frequency noise, they also inhibit recovery of high-frequency details, like exact discontinuity locations. This problem does not completely vanish even on the pixel level, as a pixel on the 3D boundary is the result of foreground and background colours – this matting problem has also been addressed either explicitly [SG99, XJ07, BGRR09], or implicitly by developing match measures robust in such situation [BT98, SS04].

Meanwhile, the problem of accurate and reliable recovery of 3D boundaries is very important by itself, as many applications, such as robotic manipulation, 3D reconstruction, Augmented Reality and 3DTV, critically depend on accurate depth discontinuity information. Moreover, humans are very sensitive to 3D boundaries and are able to recover them with precision greater than spacing of photoreceptors on the retina, i.e. they exhibit stereo hyperacuity [HR02], which proves that nature has a good solution for recovery of 3D discontinuities, and it is yet to be discovered.

[1] Recall that a surface is Lambertian if its luminance is the same regardless of the viewing direction, as it depends only on the cosine of the angle between the local surface normal and the illumination direction.

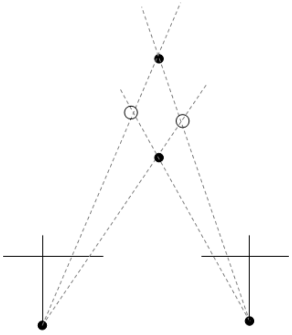


Figure 2.2: Double-Nail Illusion in Stereo. Solid dots show actual 3D configuration; open dots show typical human perception. Both configurations are equally valid from the correspondence consideration only.

Repeated texture

While points or patches in correspondence should look alike, the reverse is not necessarily true. In fact, it is very common for the majority of scene features not to be distinctive and repeat throughout the scene – a phenomenon usually referred to as *repeated texture*. They frequently arise especially in man-made structures, which tend to be produced using some templates, e.g. textured cloth, windows in buildings, etc.

One of the extreme examples of repeated texture, which is of particular interest to psychophysical studies, is the so-called "double-nail illusion" [Kv80] which is shown in Fig. 2.2. The projections onto left and right views can arise from two valid, but completely different 3D scenes due to ambiguity of matching (two nails are identical to each other). The major interest to psychologists is that humans tend to choose the wrong configuration in this particular scenario.

Textureless regions

Continuing the discussion of repeated texture, it is worthwhile to consider the special case when all surface points are the same, i.e. textureless regions. Here, the possibility of various matching assignments is virtually infinite, instead of a discrete number of valid configurations, as in the case of the double-nail illusion. Thus, visual systems try to enforce some viable surface model over this region. Regarding computer vision approaches, this model is often fronto-parallel (constant depth is implied by constant intensity), generally planar (as useful in many applications such city/building/office modeling [Fra08]) or minimum-curvature to cover a variety of smooth non-planar surfaces. Interestingly, humans seem to possess a different prior which is more closely described by minimum-Gaussian-curvature [IG06]; however, no computational realization of this suggestion has been presented to date.

Finally, textureless regions are the only ambiguity which cannot be resolved even if (infinitely) many cameras are added to the scene (the lightfield is probed as densely as possible), i.e. it is the only inherent stereo ambiguity [BSK01]. As a representative example consider Fig. 2.3, which shows a planar patch with uniform colour and a concave patch with the same colour plus possible compensation for inter-reflections. Both patches will appear identical in all corresponding camera views and even extra cues such as shape-from-shading under a Lambertian surface assumption[2] or

[2] Recall that standard shape-from-shading and uncalibrated photometric stereo solutions are characterized by the so-called bas-relief ambiguity [BKY99] – there exist multiple interpretations for lightsource positions and albedo profile for any observed shading scene. In general, only calibration, priors or various assumptions can solve for this

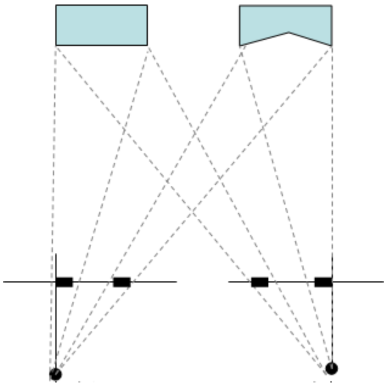


Figure 2.3: Inherent Ambiguity of Stereo. The absence of texture leaves the exact shape of the object unkhown. In this example, while the outlines of two boxes can be correctly recovered by passive multiview sensing, their interiors would stay undefined as their projections results in identical appearances in the images. In this case, common simplistic stereo priors always prefer the flat shape on the left versus concave shape on the right.

shape-from-silhouette will not distinguish between the two.

Occlusions

As established so far, computational stereo algorithms try to find points in correspondence. However, for some points in the scene the correspondence cannot be found in principle – those points are called *occlusions*, as they are seen only in one of the views of the stereo pair. With respect to the ubiquitous binocular stereo case, these points are usually referred to as *half-occlusions* [EW02]. Thus, a good stereo algorithm must not only find the points in correspondence, but also explicitly say which points have no match. Figure 2.4 shows various geometries of occlusion formation and also sketches some constraints to be discussed later in the chapter.

Historically, as early as Euclid, the basic geometric relationship that gives rise to half-occlusion was documented [Bur45]. Further, the potential perceptual significance of binocular half-occlusion has been known at least since the time of Leonardo Da Vinci [Ric77]. Much more recently, the fact that humans actually do exploit half-occlusions in making depth inferences was documented [LM67]. Subsequently, a great number of psychophysical studies of half-occlusion have supported their use by humans (see [HR02] for review); still, the enabling computations remain unclear.

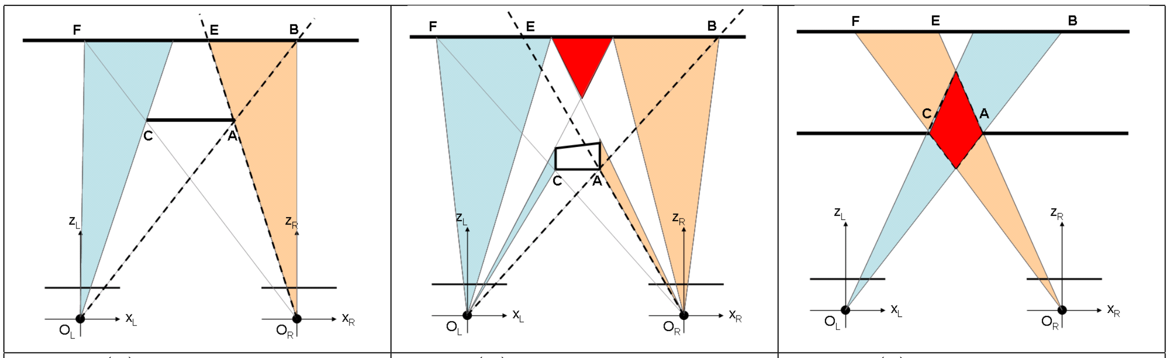
2.2.3 Constraints from image geometry

Given all of the hurdles of stereopsis discussed above and the desire to compute dense disparity fields, stereo would not be possible without several important constraints to make the problem more approachable. Important constraints arise directly from the stereo geometry setup and the scene formation mechanism itself.

Photometric constraint

This is the basic constraint to find points in correspondence across views. Most simply, it is assumed that a point in the world projects with the same brightness across views (c.f. brightness constancy

ambiguity.



(a) wide region | (b) narrow region | (c) narrow hole

Figure 2.4: Occlusion Formation in Binocular Stereo. (a) The simplest case occurs when all points on the back surface that are within the "forbidden zone" of the boundaries of the front surface are half-occluded, e.g., A is the right boundary point of the front surface. (b) More complicated situations occur when narrower front surfaces allow portions of the back surface within the forbidden zone of the front surface boundaries to be binocularly visible. Further interposed surfaces in the red (dark grey) region allow for half-occlusion relations to occur *recursively*. (c) Half-occlusions also occur in viewing back surfaces through a narrow hole in a front surface; the back surface is binocularly invisible.

in Sec. 2.3.3). Alternatively, some derived photometric image measurements are assumed to be the same (e.g. colour, edges, etc.).

Geometric projection constraint

Photometric constraints say that points, or rather small patches, in correspondence look alike. By this statement, it is usually meant that they look alike both photometrically (brightness or colour) and geometrically (shape). However, due to perspective effects, object's projections onto separate cameras can look slightly different, as shown in Fig. 2.5. Nevertheless the geometry constancy assumption is wide-spread and is viable generally due to small local support region that degrades to a single pixel in most cases. Theoretically, this constraint is only true for points (which are dimensionless) and fronto-parallel patches, while general slanted and curved surface will violate it.

Epipolar constraint and rectification

Consider a 3D point P and its projections into the left and the right cameras p^l and p^r, respectively. The cameras themselves are defined by the optical centres O^l and O^r and images planes I^l and I^r. Figure 2.6 makes matters more precise. The point P and cameras' optical centres O^l and O^r form a plane in 3D called *epipolar plane* and called Π, which intersects the image planes I^l and I^r along the lines m^l and m^r, respectively. Now, if we fix on the point p^l in the left image, we know that it is the projection of the 3D world point which could be anywhere along the O^lP ray, which means that the corresponding projection in the right image can occur *only* along the line m^r. The reverse is also true for the right point p^r. Thus, lines m^l and m^r are called *epipolar lines* – point correspondences can be found only along these lines. Further, another arbitrary point P' gives rise to a different epipolar plane defined by O^l,O^r,P' and, correspondingly, to a different pair of epipolar lines. However, every different Π will pass through the line O^lO^r by construction, which means all epipolar lines in the left image will go through the projection of O^r, called the left *epipole*, and conversely, projection of O^l onto the right image is the right *epipole*. This is the theory behind the

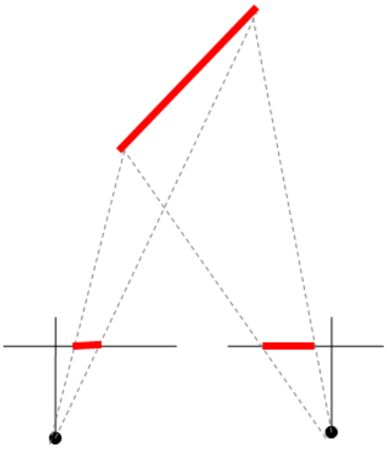


Figure 2.5: Perspective imaging in stereo processing. Due to viewpoint differences, the same 3D shape projects to different 2D shapes.

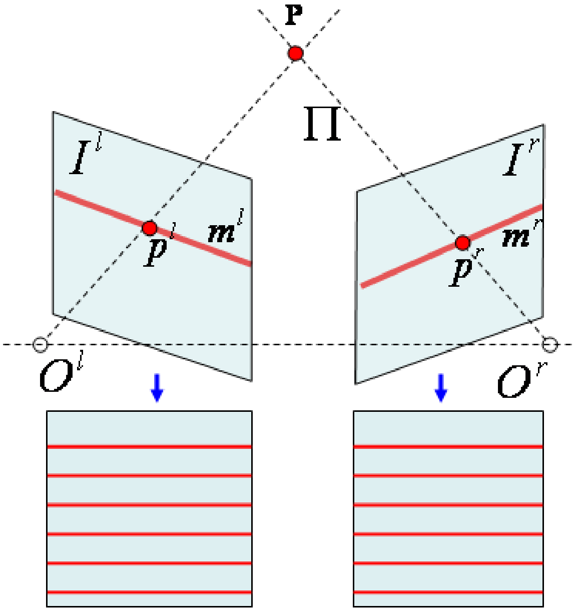


Figure 2.6: Epipolar Geometry in Binocular Stereo and Rectification. Possible points in correspondence may reside on epipolar lines only. The epipolar lines configuration is completely independent on the observed scene and can be altered via rectification – the process of warping the images that makes epipoles aligned with the horizontal scanlines for more efficient computer processing.

epipolar constraint, which reduces the general 2D correspondence search problem to 1D, i.e. search along epipolar lines only.

Once the epipolar geometry of the scene is determined, it is possible to transform left and right images such that the epipolar configuration is the simplest and most efficient from the algorithmic viewpoint – put the epipoles at infinity, which makes the epipolar lines parallel to each other and orient epipolar lines along the x-axis to coincide them with the scanline. This process is called *rectification* [BBH03, HZ04] and is very well researched to date. In fact, virtually all contemporary stereo algorithms assume rectified imagery [SS02].

Uniqueness constraint

A single point in 3D results only in one projection in each view. Thus, it is reasonable to assume that a point in one view has at most one corresponding point in the other view, i.e. introduce the constraint of the unique match. However, uniqueness is not easy to apply correctly, because it is stated for points, while stereo algorithms deal with pixels. Trivially, consecutive points on a slanted surface have slightly different disparities, which can easily be quantized to the same pixel disparity value (these difference between disparities is less that 1 pixel), i.e. they will violate uniqueness, which makes non-fronto-parallel surfaces hard to recover. Uniqueness also can be violated when dealing with (semi) transparent surfaces.

2.2.4 Constraints from object properties

The previous section discussed constraints that arise from image geometry, while even more can be gained by exploiting constraints and assumptions on the scene itself.

Depth/distance continuity

The depth/distance continuity constraint is also referred to as the smoothness constraint, coherence principle and connectivity constraint. This constraint has different instantiations, which are tightly connected to features and model used in matching.

A fundamental technique to make the correspondence solution more stable is to assume spatial smoothness, or cohesion, which means that points belonging to a single object tend to reside at a certain near-constant depth. Smoothness is usually enforced by penalizing neighbouring points that have different depths (and hence, disparities) or by assuming that neighbouring points reside at the same depth by aggregation and matching the aggregated regions. Actual mechanisms for application of smoothness can vary, but any contemporary stereo algorithm includes this constraint. In fact, algorithms that only rely on intensity-based pixel matching, epipolar geometry and smoothness are among the state-of-the-art solutions [BVZ01, SZS03].

Intensity continuity

As stated before, imprecise 3D boundary recovery is still the plague of stereo algorithms, and that is why extra constraints to define a structural, or 3D, edge are summoned. The most widespread, which works rather well in many scenarios, like office spaces, is that 3D edges can lie only along the intensity edges. This idea is also referred to as the colour segmentation cue and is abundant in contemporary stereo [SW06].

However, it is worth noting that this cue is not adequate when foreground and background objects are heavily textured in a similar fashion, as intensity edges are abundant in this case and do not reliably define any 3D structural edge. One of the most striking example of this is a random-dot stereogram; extended discussion of this phenomena can be found in [SW06].

Planarity

In many applications that aim at surface geometry recovery, it is known what kind of surfaces are present. The simplest, but also the most important and abundant, case is that of the planar surface. The surfaces can be large as in city modeling applications [Fra08]. Alternatively, the scene itself is modeled with local planar patches [HO93, CK02]. An algorithm for even more constrained scenarios of explicit modeling with a Manhattan prior (mutually-orthogonal planar regions, like in conventional buildings) has been developed [FCSS09].

Disparity limit

Whenever the search for correspondence is to be performed, there is always the question of what is the disparity range for these possible correspondences. Certainly, one can attempt to match every pixel in one image to every pixel in the other image, but this may result in 3D points that are very close to or very far away from the camera. In many applications, it is known in advance over what area the 3D model is to be obtained and that directly dictates the range of disparity values that need to be considered, i.e. define the disparity search range. Defining the search range, or range of binocular fusion, is common in spirit to Panum's fusion area in humans and animals – the area over which binocular fusion occurs [HR02].

It is worth nothing that while almost all algorithms operate over their fusion range only, there is an increasing interest to be able to do stereo processing in more constrained ranges with disparities outside of this range being correctly labeled as out-of-range [AB06], i.e. actually imitate the Panum's fusion area.

Disparity gradient limit

It was found by Burt and Julesz that the human visual system performs binocular fusion only if the disparity gradient limit does not exceed one [BJ80]. The computational realization of this constraint has been found in the form of the disparity gradient limit algorithm known as PMF [PMF85]. From the computational stereo point of view, such a constraint excludes a lot of objects with irregular geometries, (which are physically possible, though) but still lacks justification of its overall practical usefulness. Thus, it is very hard to meet it in recent algorithmic solutions, other then ones that are proposed as realizations of biologically-plausible models.

Ordering constraint

The ordering constraints states that the order of match points in both views should stay unchanged and Fig. 2.7 visualizes this statement. However, this statement is fundamentally false in the case of thin structures as shown in Fig. 2.7 and in Fig. 2.4, which restricts the use of this constraint to the certain types of objects. Despite that limitation, this constraint is still used today, as it greatly simplifies matching, as in case of dynamic programming [SS02, WIM05] and further regularizes the recovered disparity maps [Kac04].

Conceptually, the ordering constraint is very tightly related to the disparity gradient limit, as for points which violate the constraint, disparity gradient is always greater than one [Kv80].

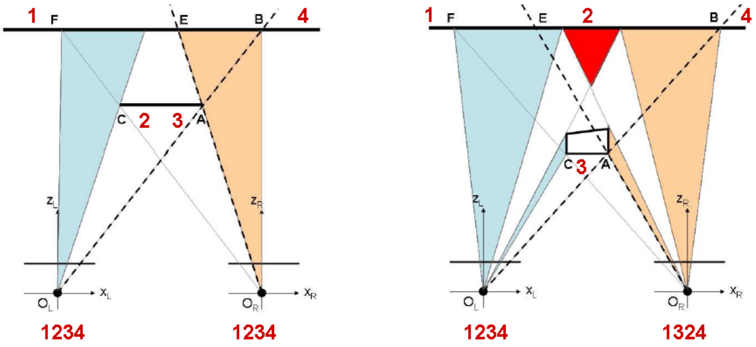


Figure 2.7: Ordering constraint and its failure. Ordering constraint state that features should appear in identical orders both is the left and right views (i.e. $1, 2, 3, 4$ in the example of the left sketch). However it is not true for thin structures in the scene that generate multiple occlusion regions, as depicted in the right sketch.

Occlusion constraint

This constraint literally says that a discontinuity in one camera corresponds to an occlusion in the other camera and vice versa [EW02]. This statement is always valid, as occlusions always arise near 3D boundaries, as explained in Fig. 2.4, and, in fact, results in one of the best occlusion detection algorithms [EW02, SLKS05]. Still, it is worth mentioning that this constraint connects disparity map in one view with occlusion map in the other view, which makes it inconvenient to use, since calculation of both disparity maps may be an unnecessary overkill.

2.2.5 Beyond 2 views

The current chapter predominantly considered the case of binocular stereo as the minimal multiview configuration, which is also inspired by biological design. In computer vision, it is possible to add more cameras to gain certain benefits, of which we can outline the following:

- The correspondence problem is more constrained simply because several points should be related. Algorithmically, this can be done by independent processing of each pair of views and then fusing results together, or doing matching in all views simultaneously (assuming calibration) [GCS06].
- Matching across multiple views is more robust and can be made insensitive to specular reflections [SJW07], transparencies [SG99, TKS06] and other challenges.
- Multiple cameras allow for more advanced occlusion processing mechanisms, e.g. [NMSO96].
- Some ambiguities, like repeated texture, are annihilated. In fact, we are left with only one inherent stereo ambiguity – homogeneous regions, when all points look alike [BSK01].

Moreover, once many more views become available new concepts can be explored, such as space carving [KS00], or 3D reconstruction from 2D slices in the Fourier domain [MAK$^+$06].

2.2.6 Algorithmic realizations

Over multiple decades of intensive research, stereo algorithms using virtually all possible combination of constraints described above have been realized.

It is beyond the scope of the present document to discuss all computational stereo realizations and readers are referred to various comprehensive reviews that appeared progressively over time. One of the earliest reviews that described predominantly feature-based methods is [DA89] that later was augmented by [Kos93]. Most recently, the review literature has culminated with [BBH03], which is almost entirely devoted to pixel based methods. A special niche was taken by [SS02] as the most comprehensive work that attempts to evaluate empirically stereo methods based on dataset with ground, which continues to be updated. New algorithms are constantly appearing and usually can be found on the associated comparative website [Mid08].

Early machine stereo algorithms made use of simple area-based correlation schemes, often motivated by photogrammetric applications. There approaches met with limit success as their lack of sophistication led to generally poor performance (see [KP81] for review). The following generation of stereo algorithms were developed more as a proof-of-concept solutions that worked predominantly on features such as Moravec interest points, Laplacian zero-crossings and edges [Gri81]. Since matching must be found explicitly for each point, stereo is considered a heavy computational process. Furthermore, matching points must be distinctive, which explained why all emphasis was put on sparse stereo algorithms and a great effort was also directed toward subsequent interpolations of sparse results [BZ87].

As time progressed, computational power drastically increased and other efficient paradigms like coarse-to-fine processing appeared, which made dense stereo processing an everyday reality. The obvious advantage of this class of method is the direct recovery of dense disparity maps, which largely removed the interpolation problem. Furthermore, formulation and assumptions of dense methods tends to be more intuitive, as, for example smoothness is naturally applied to nearby regions.

Being also concerned with biological plausibility of the computational solutions, the area of phase-based disparity estimation has been derived [FJJ91] – the architecture is based on processing responses of Gabor filters [Gab46] or their alternatives, which to some degree mimic the behaviour of stereo processing cells in the mammal vision system [HR02].

An important class of correspondence search procedures to emerge are known as "local" methods. The basic idea is to find the best correspondence for every point using some similarity measure between image patches aggregated around the matched points – the match metric itself is the area of research and numerous methods that vary from simple sum of absolute intensity differences to mutual information have appeared [BBH03]. In essence, local methods concentrate on the photometric constraint in stereo, which is not nearly enough to realize completely successful solutions in practice. Various complications such as half-occlusion [EW02], 3D boundaries, and textureless regions are not overcome.

Out of other possible drawbacks, "local" methods exhibit a limited way of utilizing smoothness constraints, which is of paramount importance, as discussed earlier in this chapter. The only way to control the level of smoothness here is to vary the aggregation window size, which must be large to behave well in noise and weakly textured regions and small to behave well near discontinuities and thin structures. Ideally, we want to have both together, which explains the need to further develop the matching framework.

Cooperative algorithms can be seen as providing a bridge between local matching and global solutions, which are described in the next paragraph. Historically, cooperative methods were inspired by computational models of human stereopsis [MP76] and are realized by diffusing reliable matches to neighbours and inhibiting values along the view-directions of the left and right eye, i.e. enforcing the uniqueness constraint [ZK00].

"Global" algorithms are one of the most recent advances in stereo vision and the primary direction in which current stereo research is headed in many respects. The name is largely meant to contrast with the local methods by stressing the difference in how the smoothness constraint is enforced. Conceptually, global methods are based on a simple energy formulation that combines the pixel-wise intensity matching term and the smoothing term to enforce continuity between nearby pixels. Usually, the latter is done by constructing an Markov Random Field (MRF) that has explicit links only for immediate neighbours. The smoothness itself is usually just a robust way to penalize the differences in disparities [BBH03], although more sophisticated priors based on differential geometry have been developed [LZ03, OWF08]. Certainly, other constraints described in this chapter have also been incorporated into this formulation. The major "art" in global methods remains solving these energy formulations and a whole variety of solvers (e.g. based on PDE, Graph Cuts, Dynamic Programming, Belief Propagation, stochastic diffusion, genetic algorithms, Tree-Reweighted message passing, Quadratic Pseudo-Boolean Optimization and others) have been developed.

2.2.7 Current level of performance

Despite confronting a number of challenges and the wide variety of real-world situations, stereo has come a long way in terms of actual performance. Importantly, the evaluation methods were developed along with the stereo algorithms themselves, so our discussion will keep these issues together. Early evaluations concentrated on synthetic data or qualitative evaluation of simple scenes. An important impact to the development of better stereo algorithms came through advancement in precision and robustness of active range-finding devices, making it possible to acquire stereo data with ground truth, even if confined to lab scenarios.

Early sparse stereo matchers were reasonable in providing sparse correspondences while trading off density. Taking the example from psychophysical stereo research, these algorithms were initially tested on random dot stereograms (RDS), as these stimuli were confined to the stereo cue only. Consequently, improvement on RDS was achieved, and solutions that completely solved them were demonstrated back in the 90's [SS98a]. At the same time, real scenes remained challenging due to low texture, aliasing, curved surfaces, non-Lambertian surface reflectance, and, finally, real image noise and radiometric calibration.

Dense local methods are capable of performing rather well for well-textured scenes (e.g. achieving subpixel precision) and various commercial realizations are available for applications, like robot navigation and semi-dense 3D reconstruction [Poi08, TYZ09]. Furthermore, recent ideas incorporating the segmentation cue gave the algorithms another boost [YK05], as their performance significantly increased and became competitive to state-of-the-art solutions [Mid08].

Current state-of-the art algorithms, mainly being global methods, are able to show very good error statistics on datasets with ground truth, which are taken in conditions close to reality [Mid08]. Not surprizingly, the best performers are the methods that combine the most generally applicable constraints and try to solve them globally while using the most descriptive priors (e.g. initial segmentation, which greatly improves region coherence). In terms of numbers, it is possible to recover the correct disparity for more than 95% of pixels in rather complex and relatively unconstraint lab scenarios. Furthermore, evaluation under sever noise conditions and radiometric miscalibration finally has an acceptable level of attention [HS08]. It also is worth saying that quality of captured test images is rather good, which may not be guaranteed in every application. Moreover, these algorithms, being complex, are rather hard to tune to the variety of scenes and stereo setups – results achieved for one dataset are quite hard to replicate for another. Interestingly, optimization algorithms have gone a long way, so that *exact* global optimum can be achieved for some scenes [MYW05], but ultimate improvement in the results has not been achieved. This fact just reinforces the necessity for more research in modeling, generalization and, finally, the understanding of the

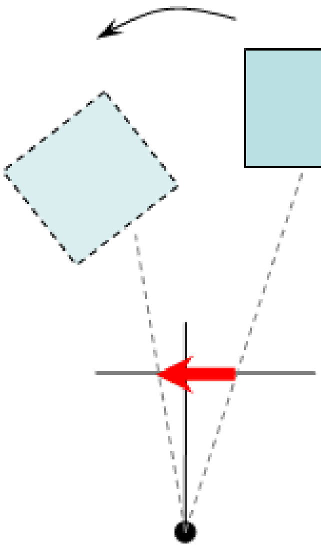


Figure 2.8: Geometry of Image Motion Formation. Moving 3D object results in 2D projections over time. Specifically, 3D motion is projected into 2D motion (depicted by arrows).

correspondence problem.

At the same time, rapid and resource-efficient performance is needed in most practical applications, which remains dominated by local and coarse-to-fine algorithms, as they are characterized by inherently lower complexity, ease of implementation, parallelization and robustness to parameter tuning [BBH03, SW09b].

Finally, since binocular stereo by definition is a purely two-frame problem, there has been no formal investigation of the stability of those algorithms over time i.e. how the quality of results behaves while the scene changes over time. This question is very important in practice, as all operations are performed over some time. Indeed, it is one of the most important questions for spatiotemporal stereo and will be discussed in detail later in this report.

2.3 Motion

2.3.1 Basics

Images taken in consecutive time frames by the same camera may reveal the temporal dynamics of the scene as well as its 3D layout. The basic geometry of image motion formation is depicted in Fig. 2.8. Image motion analysis can be performed from only two consecutive frames, the minimum possible amount of data, which makes it somewhat similar to binocular stereo processing. It is also possible to consider more frames at once and introduce more general motion analysis techniques, as discussed in Sec. 2.3.4.

Two frame motion computation algorithms are the cornerstone of image motion analysis. The first algorithms appeared in the late 60's and into the 70's and were largely motivated by video compression [Mou69, Has74, LM75, CR76, MA78, FT79]. Interestingly, solutions proposed in the early 80's are still widely used and considered very practical [LK81, HS81] (subject to careful contemporary implementation).

Usually, the end result of image motion estimation is the vector field that tells the 2D image displacement for each point. Depending on the physical scenario, the flow field vectors can arise

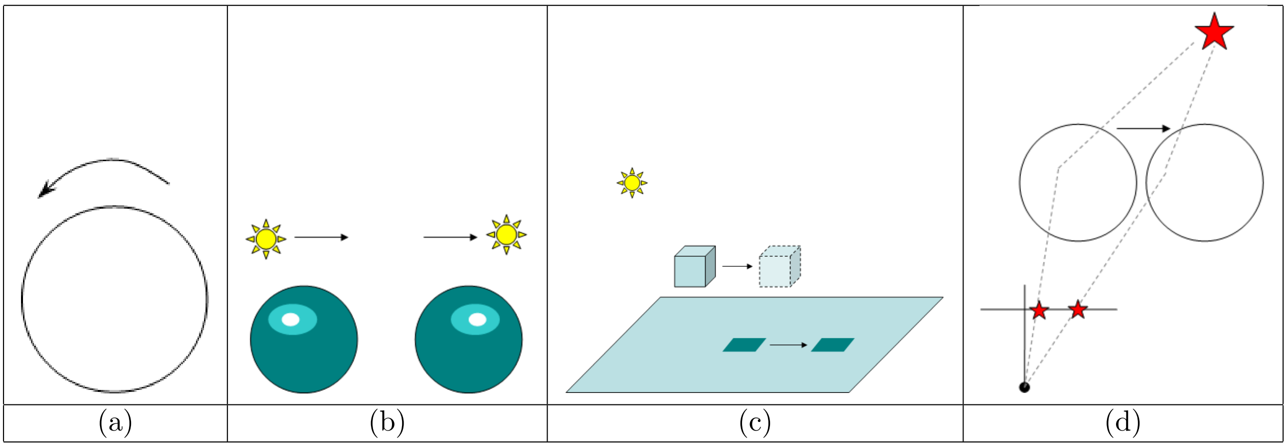


Figure 2.9: Motion Field vs. Optical Flow. (a). Motion of textureless object. (b). Motion of the light source (c). Motion of shadow (d) Specular motion.

from a single parametric model (such as translation, affine, planar [BAHH91], or even quadratic [SW01]) or be virtually independent for each point.

In any case, during the small time change, δt, between consecutive frames, each point undergoes image motion, $(\delta x, \delta y)$, and we need to employ some constraints to infer the flow vectors. The basic commonly employed constraint is photometric, a simple linear constraint that ties image brightness change to the motion vector, $(\delta x, \delta y)$. It is based on the brightness constancy assumption (an image point matching its brightness across time) which will be described below in detail.

One of the possible solutions is to consider this two-frame configuration as a general stereo problem and simply perform correspondence search subject to 2D disparity vectors. Few such algorithms exist, e.g. [Ana89], as most contemporary image motion computation techniques come from the realization of the rather useful assumption that the time interval between consecutive frames, δt, is small and flow describes differential properties of the scene.

Motion field versus optical flow

The brightness constancy assumption is one of the most fundamental in the analysis of image motion. In such formulations, image motion is often related to an "optical flow", the flow of intensity pattern in the image.

Meanwhile, optical flow is not necessarily related to projected object motion referred to as the visual motion field. Some examples are sketched in Fig. 2.9 and we will briefly discuss them below:

- In textureless regions, similar to the inherent stereo ambiguity, essentially zero image motion may in reality correspond to many object motions.
- Things other than observed objects might be moving in the scene, such as the light source, which results in dynamic change of appearance of a static object. Note that even the powerful Lambertian assumption is helpless in this situation, as incident illumination angle changes with time, unlike stereo which takes images at the *same* time.
- Moving objects usually cast shadows on the background that might be moving differently. The boundary of the shadow will create a wrongly perceived movement of the background surface (object).

- When a moving object is specular, or posses significant specular components, the object motion manifests itself as the specular flow, i.e. flow of reflected features, which is very much different from the projected 3D flow of the actual objects. Some success has been demonstrated in recovery of specular flow [ON96, RB06] and this area is receiving attention.

Optical flow vs. structure and 3D motion

In the context of this report, optical flow is taken as the image projection of some 3D object undergoing particular 3D motion, i.e. optical flow and the visual motion field are equated. If in case of stereo, we are essentially able to resolve the structure of the underlying object, it is natural to ask whether the 3D structure and motion can be inferred from the optical flow.

The answer is two-fold and has to do with the fact that optical flow describes behaviour in time. If the object is rigid over time such that optical flow can be attributed purely to camera motion, then scene structure is recoverable via "structure-from-motion". In cases when the scene deforms over time, however, an infinite number of underlying 3D structure and motion solutions are possible, thus extra information, e.g. stereo, is needed to disambiguate the possibilities [VBR+05].

2.3.2 Challenges

Non-Lambertian scene and changing illumination

Although the major constraint for motion estimation is based on brightness constancy, the latter is true only for surfaces whose imaged points retain their brightness across time, e.g. Lambertian surfaces with constant angle between illumination and surface normal direction. Furthermore, the problem is more severe than for stereo – as images are taken at different instances in time, there is usually a relative motion between a surface and a light source which results in change of incident light angle and change in the observed brightness even in the case of a Lambertian surface! This issue can be dealt with either by explicit modeling and subsequent inference of lighting, which is not currently perceived as practically possible, or simply relying on small change and describing the differential behaviours, i.e. reinforcing the underlying rule of thumb that motion estimation is only reliable when motion is small.

Aperture problem

The optical flow constraint (2.2) can be used to compute flow components, u and v, subject to brightness constancy. However, the major difficulty with such an approach is that we have two unknowns, but only one constraint, which suggests we cannot determine u and v for each point uniquely.

This limitation is fundamental, as the constraint (2.2) based on image derivatives constitute brightness gradients, and the motion component along the gradient can be determined, but not across it. In essence, it is only possible to recover the normal flow, as the flow across the gradient (brightness isocontour in this context) can be arbitrary, as depicted in Fig. 2.10. In general, normal flow is the flow vector with the smallest length that satisfies the optical flow constraint (2.2) for a given point.

This ambiguity problem is generally referred to as an *aperture problem*. Note that to the extent stereo correspondence is formulated as 1D, it is immune to the aperture problem.

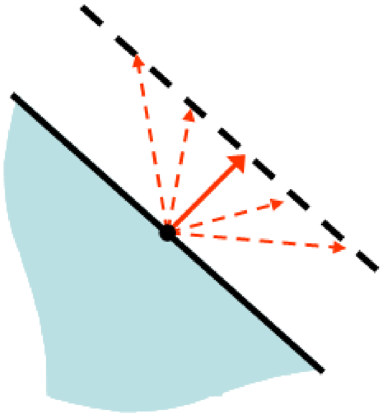


Figure 2.10: Normal Flow and Aperture Problem. Brightness constancy constraint is able to describe only the motion of the wavefront, i.e. along the gradient direction. In turn, each particular point may undergo any type of motion (depicted by dashed arrows in the sketch) as long as it is consistent with the normal flow (bold arrow).

Untextured regions

The aperture problem is about constraining two unknowns with only one constraint and can take a more severe form when zero-gradient regions are observed, i.e. purely textureless regions. In this case the constraint reduces to the tautology $0 = 0$ and no motion can locally be inferred for such regions.

Motion discontinuities

The smoothness constraint is the key for successful application of brightness constancy and recovery of optical flow. However, it is fundamentally false in cases of motion discontinuities, which arise near object boundaries – a situation identical to stereo. Both Lucas/Kanade and Horn/Schunk algorithms do not take discontinuities into account and their performance is particularly poor around those areas [BFB94, BSL+07]. Another fundamental problem of discontinuities is that differentiability assumptions used to derive (2.2) do not hold anymore, which makes the computed derivatives virtually meaningless.

The problem of discontinuities has been approached by exploiting adaptive support regions [Ana89], or robust regularization functions [BR96, BWS05]. Another common trick is to use edge constancy coherence, or a segmentation cue, as described in 2.2.4 using roughly the same machinery as for stereo processing.

Occlusions and Deocclusions

Moving object discontinuities inevitably result in some points disappearing, i.e. become occluded, or re-appearing, i.e. become deoccluded. As in the case of stereo occlusions, such points cannot have associated motion vectors as the data is fundamentally unavailable. It is worth saying, however, that due to generally small motion, occlusion regions are not as big as in the case of stereo and that is why most motion recovery algorithms either do not attempt to detect them explicitly, or simply rely on correct neighbouring values being extrapolated. Furthermore, occlusions are still not explicitly evaluated in typical quantitative experiments [BFB94, BSL+07].

2.3.3 Constraints

Photometric constraint

Photometric constraints are essential for vision algorithms, since image colour/intensity is the major and usually the only source of sensory information. In the present context, this constraint describes what should stay constant from frame to frame, which explains the name “constancy”.

Image motion is perceived as a displacement of some point or features, and typically in order for this to happen, they should be similar from one frame to another. The most straightforward way to make use of this fact is to assume that brightness, or colour, of the point does not change from one frame to another. In mathematical notation, if a point with coordinates (x, y) in the image I at time t undergoes image motion $(\delta x, \delta y)$ during the time interval δt, the following should hold:

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t) \tag{2.1}$$

This is mainly referred to as the Brightness Constancy Constraint (**BCC**), or the brightness constancy assumption. Taken as is, it justifies matching of points based on similarity of brightness.

Commonly, BCC is further manipulated to yield a relationship between image derivatives and optical flow under the assumption of smooth brightness variation and differentiability with respect to all variables. In particular, application of a Taylor series expansion can achieve the following result [LK81, HS81, TV98]:

$$I_x u + I_y v + I_t = 0. \tag{2.2}$$

Here, I_x and I_y are spatial derivatives along x and y axes, while t is a temporal derivative, which is usually just a frame difference in case of only two frames and, once again, (u, v) are flow components. Another theoretically-legitimate way to introduce additional constraints is by exploiting higher order derivatives [GVT89]. Practically, this direction is questionable as noise makes higher-order estimates all but unreliable.

The derived-above Optical Flow Constraint Equation (**OFCE**) (2.2) is the most widely used constraint in optical flow estimation. It is refereed to by different names (*optical flow constraint equation*, *brightness constancy constraint equation*, *gradient constraint equation*, *motion constraint equation*) and is derived from the brightness conservation (or constancy) assumption. Meanwhile, other conservation laws can be explored depending on the nature of the application, such as conservation of total brightness, conservation of gradient direction or magnitude [Der06], other conservation laws based on general physics models [Neg98, WB00, HF01, SAS08], or enforcing a polynomial function of temporal brightness changes that is to be estimated over time [Neg98, SAS08].

Coherence

The flow vector at each point cannot be established uniquely based on the above-formulated photometric constraint, as only one equation relates two variables. However, observed objects are usually smooth and this observation can be used to provide additional constraint. More specifically, nearby points are usually assumed to undergo the same motion, which is similar to the stereo situation, where it is assumed that nearby points have very close disparity values.

The assumption of locally constant optical flow is particularly simple to utilize, as it forms an overdetermined Least-Squares (**LS**) system of linear constraints (2.2) for the set of nearby points sharing the same unknowns u and v. Solution to this generally overdetermined system yields the flow vector. This method, though originally proposed with application to stereo vision only [LK81] is widely known as Lucas-Kanade method. Essentially, this is instance of a “local” method employed in optical flow estimation.

Another way to apply smoothness is to solve the constraints for all points simultaneously subject to regularization – an extra term that penalizes large differences between near-by flow vectors. This “global” method was first proposed by Horn and Schunk [HS81] and has been known under such name since. The optical flow recovery problem transforms into the minimization problem

$$\arg\min_{u_i,v_i} \sum_i \left(I_{xi}u_i + I_{yi}v_i + I_{ti}\right)^2 + \sum_i \alpha^2 \left(||\nabla u_i||_2^2 + ||\nabla v_i||_2^2\right), \tag{2.3}$$

where α is the regularization parameter, which weights the contribution of a smoothing prior by penalizing large first-order gradient of the flow field. This minimization problem can be solved by converting it to the the Euler-Lagrange equations, which are efficiently calculated in iterative manner, as detailed in Sec. 3.4.3.

Parametric motion

While introduction of coherence allows for estimation of general motion vector fields, it can be useful to restrict the computed vectors to follow some global parametric pattern, i.e. affine motion, quadratic motion etc. Instantly, the whole problem is transformed into the inference of a relatively small number of parameters using constraints from all points – the latter can be reliably solved via linear methods and is perceived as a generalization of Lucas-Kanade [BM04].

Large motion

Importantly, when we need to recover large displacements, the small motion assumption should still be valid in order for (2.2) to remain usable. The way to achieve this is to perform motion computation in coarse-to-fine fashion using image pyramids [BA83], whereby, estimation at any given resolution is taken to entail only small motion (see Sec. 3.4.2). Alternatively, large motions can be handled by escaping the differential assumption (2.2) and attempting to enforce BCC (2.1) more directly through explicit search over some 2D region – a generally expensive and error-prone methodology.

It is worth noting that, while most of the state-of-the-art stereo algorithms have largely abandoned coarse-to-fine processing schemes [Mid08], motion computation continues to rely on it heavily. Apparently, given that stereo is more constrained (correspondence search is reduced to 1D space in comparison to motion which usually entails 2D search), it does not strictly require the use of coarse-to-fine to further limit search.

2.3.4 Beyond 2 frames

The previous section briefly described the motion computation principles that are directly applied to two-frame motion. More generally, we deal with video sequences and the temptation to use more frames at once is great and well-motivated.

Use of multiple consecutive frames is expected to increase the quality of the computed results simply because more data is used. One way to utilize multiple frames is different, arguably more precise, computation of temporal derivatives (using temporal kernels other than $[-1, 1]$ as in the pure 2-frame case). Another way is simply to filter or smooth the recovered optical flow fields once more video frames are supplied [RAP06]. It is also a possibility to recover higher-order descriptors of motion, such as acceleration [Der06], which is fundamentally impossible having only two frames.

A quite interesting application of simultaneous multi-frame motion analysis has been discovered by Bergen et al. [BBHP92]. These authors showed that it is possible to recover multiple motions at

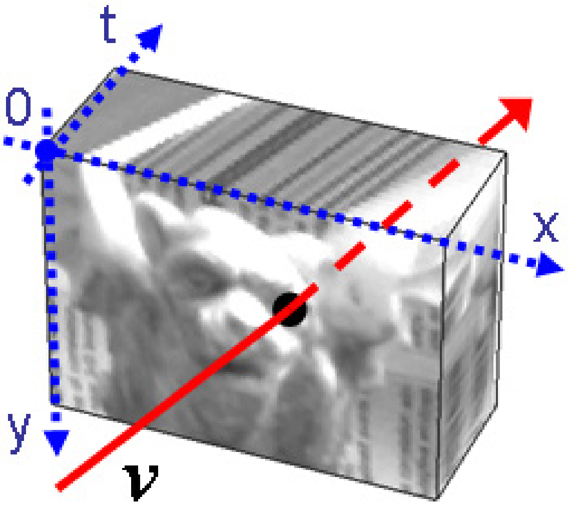


Figure 2.11: Spacetime and motion through it. Consecutive video frames stacked along the t-dimension effectively form 3D spacetime. An image motion of a point in any frame now manifests itself as an isobrightness line (more generally, contour) in this 3D spacetime. Analysis of these local orientations (and lines in particular) yields rich information about the motion in the scene.

each point, which allows one to solve for very hard situations such as transparency/translucency and better handle 3D discontinuities[3]. In particular, the authors propose to use three frames to acquire two motions and the idea is based on a simple observation that if the first component is somehow estimated and subtracted from the frames, the pairwise frame difference between 3 consecutive frames will create a pair of frames that has only the second motion, which will be readily estimated using a standard method. The idea is generalizable to n motions from $n+1$ frames [BBHP92] and has been successfully realized in the analogous case of multi-view stereo [TKS06].

Spacetime: motion as orientation

Meanwhile, consideration of the complete video sequence, rather then separate frames can give rise to alternative schemes of motion analysis.

If we consider the video as a volume by stacking frames along the time axis and refer to it as *spacetime*, then motion of the point will manifest itself as a line or curve in this volume [AB85] under the brightness constancy assumption. Figure 2.11 makes the matter more precise. Indeed, taken across two frames, a moving point defines a line.

The motion from orientation paradigm has been developed further by many researchers and various algorithms have been proposed. Orientation itself can be extracted in many ways, e.g. with steerable quadrature filters [FA91, DG05]. Orientation responses are usually organized into structural tensors, or quadrics, of nonnegative definite matrices which encode the direction of instantaneous motion as well as the first-order uncertainty associated with it. Direct regression on filter responses is also possible [Hee88].

Importantly, spacetime orientation analysis, instead of recovering flow, can be directly applied to higher-level cognitive tasks, such as qualitative characterization of motion [WB00], perceptual grouping [DW09], action recognition [SI07], or tracking [CW07].

[3] Note that the collection of points near surface discontinuities essentially exhibit motion of both surfaces.

2.3.5 Algorithmic realizations

Similar to stereo, motion computation has been researched over several decades. A variety of review papers are in the literature and we refer the interested readers to one of the most recent documents for more comprehensive coverage [Der06].

The idea of stereo disparity computation being a restricted version of optical flow recovery (albeit of generally larger displacements) has found the utilization of area-based matching methods for the purpose of optical flow computation. Dense flow recovery solutions (e.g. [Ana89, Sin90]) that entail explicit 2D search for optical displacement vectors have been proposed, and much wider utilization has been found for sparse features. The latter methods essentially boil down to finding similar primitive features, (of which there could be a great variety of definitions, from Moravec interest points [Mor81] and Harris corners [HS88] to SIFT descriptors [Low04]) between two consecutive frames – two features in correspondence will naturally define the flow vector. Furthermore, the complications associated with differential formulations (such as motion must be small) are completely eliminated.

Being also concerned with biological plausibility of the computational solutions, the area of phase-based motion estimation has been derived [FJ90] – the architecture is based on processing responses of Gabor filters [Gab46] or their alternatives, which mimics the behaviour of simple motion processing cells in the mammal vision system.

Regarding dense methods, the differential formulation (2.2) is the major modeling tool and in many respects the best current algorithms are careful combinations of the original Lucas-Kanade [LK81] and Horn-Schunk [HS81] formulations. Certainly, additional methods have been proposed, but [LK81, HS81] formulated the differential motion problem as we perceive it today. Methods to utilize first *and* second order derivatives have been developed as well [Nag83, UGVT88], but did not gain popularity due to inherent noise sensitivity of higher order derivatives.

The Lucas-Kanade [LK81] and closely related methods are essentially local algorithms for estimating optical flow. They operate by locally aggregating OFCEs, (2.2), to allow recovery by numerical methods, e.g. Least-Squares. A recent comprehensive report on such methods sheds light on may details [BM04]. This trend has further been developed by reformulating the problem in a total Least-Squares sense [WM95, NFH00, HF01] and even considering the heteroscedastic noise in data [NF03].

Further attack on the aperture problem and resolving textureless regions subject to (2.2) led to considering a single formulation that included all image points – so-called global methods, that have their roots in the Horn-Schunk [HS81] work. The performance of this class of methods varies quite a lot depending on the choice of smoothness constraint and its application. To date, various realizations that employ Partial Differential Equation solvers, Graph Cuts, Quadratic Pseudo-Boolean Optimization, level-sets (to name a few) have appeared [RG00, LX06, AK06, LRR08].

In fact, current state-of-the-art methods still position themselves as an intelligent fusion of these two original solutions, local and global [BWS05, PBB$^+$06, LRR08]. The major improvements came to the choice and learning of better priors [RB05, SRLB08], robust discontinuity handling [BA96], and better optimization strategies [RG00, LX06, AK06, LRR08]. Methods to generalize brightness constancy assumption to include other more complex specialized physical models have been proposed as well [Neg98, WB00, HF01].

Multiple frame motion estimation has been mostly realized by alternative estimation schemes predominantly from spacetime analysis and include the particular realizations of frequency-based motion analysis [YSD03] and motion-as-orientation analysis [GK95, Big98]. These methods have also developed to provide alternatives to reasoning about temporal imagery primarily in terms of optical flow. As a particular example [WB00] is able to qualitatively classify local temporal dynamics into intuitively useful categories, like leftward/rightward/upward/downward motion, scintillation and static; further work demonstrated the benefits of such categorization in the context of segmentation and grouping [DW09]. Furthermore, certain applications of optical flow, such as action recognition and motion matching, do not require explicit recovery of flow, but may be better realized using intermediate spacetime descriptors, e.g. spatiotemporal tensors [SI07]. Spacetime descriptors in the form of 3D orientation energies has also been successfully applied in spatiotemporal stereo [SW09b], as will be discussed in later chapters.

2.3.6 Current level of performance

Optical flow is an old and quite a well researched area of the computer vision field. Consequently, a huge amount of different methods with even more numerous variations have appeared over time, each of them having their own advantages and demonstrating them on quite different imagery. In this situation, the ability to systematically analyze methods to spot their commonalities and differences and know their relative performance urges special comparison studies.

The seminal paper [BFB94] was probably the first successful attempt to sketch the complete picture of optical flow estimation as it was in very early 90's. The authors largely implemented all major methods to date, including differential techniques, region-based matchers, energy and phase-based methods, and performed evaluation on a variety of datasets. A particular contribution was the introduction of realistic synthetic data, which simulated a flyover through a Yosemite valley and has complete optical flow ground truth. For many years ahead the performance on the Yosemite dataset became the major indicator of success or failure of the developed optical flow technique, which basically revealed tradeoffs between the accuracy and density of the results. As a result, the differential methods based on first derivatives demonstrated the most agile overall performance.

More recently, empirical evaluation has taken another big step when a set of real data was introduced in [BSL+07] as well as methods to semi-manually generate the ground truth for arbitrary and moderately complex scenes [LFAW08]. Furthermore, in acknowledgement that optical flow recovery may be only an intermediate step in other techniques, like frame interpolation, evaluation specific to those techniques was designed as well [BSL+07]. The necessity for complex scenes with ground truth indicates the overall maturity of the field – old datasets like Yosemite have been largely "solved" and more general, complex unconstrained and real datasets are required to drive the performance further and make meaningful general conclusions.

Even though, most state-of-the-art methods are based on early formulations, the performance has gone a long way in terms of improvements, with more specific examples following. Recovery of 3D boundaries significantly improved with the introduction of robust aggregation and non-convex priors. Textureless regions and the aperture problem are handled via smoothing, which produces visually pleasing results – interestingly enough, these results are not necessarily correct because of implicit planarity and even fronto-planarity assumptions. Also, solutions to overcome possible brightness constancy violations appeared, e.g. use of robust estimators [BR96]. Moreover, recent efforts have shown ability for real-time recovery, e.g. [BWF+05].

Overall, performance of optical flow algorithms still falls behind the performance of stereo algorithms, as optical flow solvers applied on stereo data give consistently inferior results in comparison to purely stereo methods – this largely indicates the overall less constraint and harder problem of optical flow in comparison to stereo [BSL+07].

Finally, note that most well documented successes have been achieved for the two-frame techniques, because multi-frame motion, especially with multiple layers, requires a much more significant effort in evaluation and experiment setup. The limited amount of methods existing so far mainly show qualitative results to support their claims. Moreover, as discussed above, the multiframe

	Stereo	Motion
+	Direct 3D inference via calibrated setups Epipolar constraint reduces 2D to 1D search Large baseline allows for triangulation with less uncertainty Not affected by moving lights or shadows since image capture is simultaneous	Small motion allows for differential descriptor that avoids search for match Occlusions and thin structures are less severe due to small baseline Repetitive texture and camouflage is not generally an issue Natural use of more than two frames for complex scene analysis
−	Matching is hard for large baseline due to perspective and non-Lambertian effects Thin structures are common Occlusions are significant	Epipolar constraint does not apply Aperture problem, which puts more emphasis on regularization 3D structure estimation is less effective because baseline is usually small BCC is weaker than in case of stereo since image acquisition conditions may change over time

Table 2.1: Advantages and Limitations of Stereo and Motion

spatiotemporal methods for motion analysis are finding utility in other areas of computer vision beyond optical flow estimation, which means their evaluation should be attributed to the particular application at hand.

2.4 Fusion of stereo and motion

Now that the basic theoretical understanding, as well as algorithmic and performance characteristics of stereo and motion have been discussed, it is useful to provide an explicit comparison of their relative advantages and disadvantages. Such a comparison is provided in Table 2.4. Examination of this table shows that the methods offer an intriguing amount of complimentarily, e.g. with the weaknesses of one contrasted with strengths of the other. Overall, these different characteristics arise from the different spatial camera displacement (baseline) and times that are used in frame capturing.

Ideally, joint consideration of stereo and motion (e.g. using binocular video as input) should help in removing the ambiguities and relaxing assumptions introduced by both methods separately. For example,

- Object with epipolar-aligned texture will be easier to match in stereo once at least some motion across the epipolar lines is present.
- Repeated texture, and binocular stereo camouflage can be easily broken once the scene is in slight motion.
- Movement of shadows will not be confused with apparent movement of the background surface, when stereo data is available.
- Stereo-occluded regions can get a better disparity hypothesis by considering the motion cue, rather than simple extrapolation from background.

- Multiple layers can be easily extracted once multiple frames due to motion are present.
- Absolute geometric quantities can be obtained from stereo through prior calibration.
- Dynamic scene recovery is enabled by motion.
- Temporally coherent 3D structure can be enforced via motion coherence.

Based on the outlined potentials, in the next chapters we will go over the major tools and schemes of stereo and motion fusion, discuss various applications and outline further potentials of spatiotemporal stereo.

Chapter 3

Relevant Techniques

3.1 Outline

This chapter will set the stage for more meaningful discussion of the actual algorithms for space-time stereo estimation in subsequent chapters. Specifically, we cover the notion of matching and correspondences as it has been used in stereo and motion literature. Then, multiframe processing in the case of predictive filtering will be addressed by considering possible useful realizations, such as the Kalman Filter and the Particle Filter. The last section will overview the major optimization and regularization methods used in stereo and motion estimation, which are essential for efficient solution of any ill-posed problem.

3.2 Spatial and temporal matching

A number of efforts have been interested only in 3D structure and motion recovery from stereo and/or motion data without addressing the matching problem, e.g. typically assuming correspondences are given, and dealing with synthetic data [BK83, ZHZ88, Sze88, LK90, YC90, Gov01, Der06]. However, any solution that is to work with real data has to deal with the correspondence problem in one way or another.

Stereo and motion paradigms are based on some notion of correspondence or matching. No wonder, the combination of stereo and motion resides on this major principle too. Chapter 2 discusses two basic correspondence principles in the form of explicit matching of features/pixels and in the form of a Taylor approximation of the intensity function. The first one is the most widely used for stereo processing and the second is the major constraint behind motion processing. However, these roles can be interchanged. For example, the ubiquitous motion estimation method known as Lucas-Kanade [LK81] was originally proposed in the context of stereo and the motion estimation of Anandan [Ana89] essentially mimics stereo matching for 2D disparity vectors. Thus, correspondence in the stereomotion problem can and has been approached with various combinations of matching and gradient-based principles. Another practically important characteristics of correspondence is its density. It can range from sparse to dense and even semi-dense, which usually symbolizes something in between. Therefore, it is usually convenient to separate matching methods based on this criteria and that is what we do below.

Sparse

The major rationale behind sparse matching is to identify distinctive enough features to make matching simple and robust. The Moravec interest point operator [Mor81] was one of the first proposed features detectors employed for such purposes and has been used in various efforts [JT86, TSJ92, HP94]. An improvement known as the Forstner interest operator [FEG87] also has been employed [OMSM00]. Kanade-Lucas-Tomasi tracker [TK91, ST94] employed "optimal" features and became quite popular [YO97, HC00, FRBG05, RFG07], where optimality was defined in terms of features distinctiveness and repeatability of detection. Other frameworks of combining edges, corners and vertices found its application [ZLF96]. An advancement known as the Harris corner detector [HS88] has received even more widespread use [GT95, PvP96, HC96, MB00, DH00, NNB04, ZN05]. Simpler edge-based matching also has been employed [LH91, DMC90, DM92, SZS94]. On the other hand, more complex descriptors in the form of SIFT [Low04] were employed in [SJ05]. Another set of important descriptors that were used, especially in earlier works, are lines [AW87, KA87, NDF90, ZF92]. Lines can be more robust than points, since multiple points participate in their extraction, but their matching is slightly more involved. Finally, since each feature has their own pros and cons, methods that use their combination have also been investigated. For example, [AW87, KA87] uses the combination of points and lines.

Once features are extracted, they must be matched somehow in order to establish correspondence. Some very early work treated extracted features as binary entities that were matched via model-based reasoning [JT86]. More commonly, the intensity pattern around the features is used for matching, i.e. computing correlation for patching around Harris corners, e.g. [NNB04]. Finally, some features have associated quite rich descriptors, which can be used for reliable matching[1] – SIFT [Low04] is the best example of that capability.

Dense

To date the most widely used approach to calculating dense correspondences between images involves minimization of intensity differences as aggregated over some window of support. Here, the intensity measurements can be derived directly from the images or following some preprocessing that serves to enhance the image statistics relevant for matching (e.g. bandpass filtering). Furthermore, the measure of intensity difference can take on a variety of forms, e.g. squared-difference, correlation, as well as normalization and robust measures [BBH03]. In any case, actual matching for the case of stereo is most typically formulated in terms of explicit search [SBBB95, PAT96, MSS99, DD02, NS02, AKK05, RFG07, WRV$^+$08]; whereas for motion it is more typical to invoke a gradient-based formulation [HO93, SS98b, KN03, ZN04, VBR$^+$05, WRV$^+$08].

Furthermore, there exist methods that employ the gradient-based constraint directly on range maps [Sze88, HH91, YBBR93, SJB02] (or even disparity maps [HRD$^+$99]), possibly in combination with intensity [HRD$^+$99, SAS08, SJB00]. An interesting variation is to employ the gradient-based constraint to solve for rigid motion directly [SS98b] without explicity image-to-image correspondence. Furthermore, phase-based methods (using e.g. Gabor filters [Gab46]) that were successfully used in earlier stereo [FJJ91] and motion [FJ90] paradigms also have been applied to stereomotion processing [PKRC00, SSCB03]

Recently, aggregation windows in dense matching have driven to two extremes. One extreme is to match at the level of single pixels (although still invoking some notion of spatial smoothness), which became popular fairly recently with the development of effective solvers of MRF prior models [WCR92, SBBB95, MKC00, Koc95, IM06, Gon06] and PDE equations [Sv02, MKS05, HD07,

[1] Note that, in most cases, descriptor is constructed from the surrounding intensity patterns anyway.

WRV$^+$08]. Operating on the pixel level implies working on the highest possible resolution, and is usually desired. Still, existing methods are quite computationally intensive and may not be applicable to very large models.

The other extreme to dense match aggregation is the patch-based approach, where local matches are calculated over regions defined by some segmentation (e.g. colour [ZK01]) or general notion of planar patches [TSJ92, HO93]. Methods that use various surface models and matching is calculated as a backprojection is another structure imposing scheme which became quite common [NA02, PKFH03, GM04, PKF05]. Finally, imposing meshes [Koc95, MS97] and even space carving [VBSK00, GM04] on correspondences has been applied in the field of stereomotion.

New avenues

The methods we have mentioned so far deal with the correspondence problem in the straightforward fashion of considering the correspondences between spatial and/or temporal pairs/triples/quadrupples of frames one way or another. Still, it is of interest to come up with matching paradigms that are specific to stereomotion, as they may encapsulate the advantages of simultaneous stereomotion processing and give rise to some algorithmic advantages.

The most clear example of such processing is spacetime stereo, i.e. stereo matching operating on spacetime volumes (as described in chapter 2) instead of just 2D images. In particular, stereomatching as a correlation on spacetime volumes has been been proposed as a method of stereo matching that takes care of the temporal cue automatically [DRR05, ZCS03].

The next step in the spacetime stereo paradigm can be the investigation of useful descriptors that take space and time cues into account simultaneously, with the work of [SW09b] among the first such approaches. It suggests doing matching on spacetime 3D orientated energy measurements [FA91, DG05] that are further collapsed to a quadric named the *stequel*, which naturally encodes the spatial texture and motion pattern of the imagery. Additionally, 3D motion can be readily recovered from stequels in correspondence and their matching constraint encodes epipolar geometry and calibration, which suggests that this approach may be the most general that tightly integrates stereo and motion information under the least constrained conditions.

The search for other primitives well suited for stereomotion processing is still under way.

3.3 Working over multiple frames

3.3.1 Predictive filtering

The spatiotemporal matching module of any binocular structure-from-motion solution takes care of the processing from two time-consecutive stereo pairs. The resulting correspondences allow for full description of the scene in terms of its 3D structure and instantaneous dynamics. At the same time, we deal with binocular video sequences, which naturally prompts the development of methods that take continuously available scene measurements into account and fuse the possible structure and motion parameters estimated from incremental processing.

Generally speaking, our goal now is to recover the structure parameters and motion parameters, which we denote s, having multiple (stereo) frames in time as our data measurements. In the current situation, we would get estimates for every time instance t – the results will depend on previous computations of s and current measurements. Moreover as noise and matching errors are attributes of any real systems, it is necessary to find the uncertainty of the associated estimates. To be even more general, it is interesting to recover the whole distribution of structure and motion parameters at each time instance in order to completely describe a solution.

From now on we refer to $s^t = \{s_0, s_1, s_2, ..., s_t\}$ as a state vector that encompasses all values of a state from the initial time 0 to the current time t. The state vector is simply a collection of structure and motion parameters we need to compute. Note that at each instance of time we are mostly interested in the last estimate s_t. Similarly, $m^t = \{m_0, m_1, m_2, ..., m_t\}$ is the corresponding set of measurements made by the system.

In this case, $p(s_t|m^t)$ would describe the distribution (or belief, as it is usually referred to) of the state (structure and motion parameters) at time t given all measurements from times 0 up to t. We now cover in greater detail what strategies can be taken in order to find this $p(s_t|m^t)$ distribution.

3.3.2 Bayesian filter

Standard estimation theory can be invoked to formalize further the issues in recovering the desired structure and motion parameters across multiple frames.

Since the structure model is clear (usually, it is rigid, but may change according to some model), the motion model must be agreed upon. Usually the motion model assumes a constant type of motion: zero-th order for constant translational and rotational velocities, or first order for constant acceleration and precession. The motion model is what defines the link between the motion parameters at time t and $t - i$.

We can try to solve $p(s_t|m^t)$ directly, i.e. use our motion model and constraints from all views, by employing Bayes rule:

$$p(s_t|m^t) = \frac{p(m_t|s_t, m^{t-1})p(s_t|m^{t-1})}{\int p(m_t|s_t, m^{t-1})p(s_t|m^{t-1})ds_t} \propto p(m_t|s_t, m^{t-1})p(s_t|m^{t-1}). \tag{3.1}$$

The first term can be simplified to $p(m_t|s_t, m^{t-1}) = p(m_t|s_t)$ under the reasonable assumption that measurement in the current state is conditionally independent on measurements made for all the previous frames. The second term can be written as $p(s_t|m^{t-1}) = p(s_t)$ for brevity, which is the prior for the current state derived from the previously estimated states (and measurements). So, in reality we work with the simplified form where where $p(m^t|s_t)$ is the likelihood of the observed measurements for an underlying current state and $p(s_t)$ is the prior for the current state. Thus, our posterior is simplified to

$$p(s_t|m^t) =\propto p(m_t|s_t)p(s_t). \tag{3.2}$$

While it may be possible to determine the state $p(s_t|m^t)$ directly at time t using all previous measurements, this strategy has limited practical applicability for a number of reasons:

- Computation using a large amount of data is slow and only gets worse since new data keeps coming with time.
- Computation based on frames largely separated in time is more challenging due to significant viewpoint change, scale change and increasing probability of occlusion.
- The motion and structure model may not fit perfectly and diverge (change) over time, which makes the constraints between distant frames meaningless. In comparison, deviations can be effectively adapted to with recursive filtering.

The idea of filtering is to digest the data as it comes, which is realized in a recursive fashion. Returning to Eqn. (3.2), the likelihood term is the description of how measurements arise and is defined by the underlying physics of the process. The prior term essentially "predicts" what the

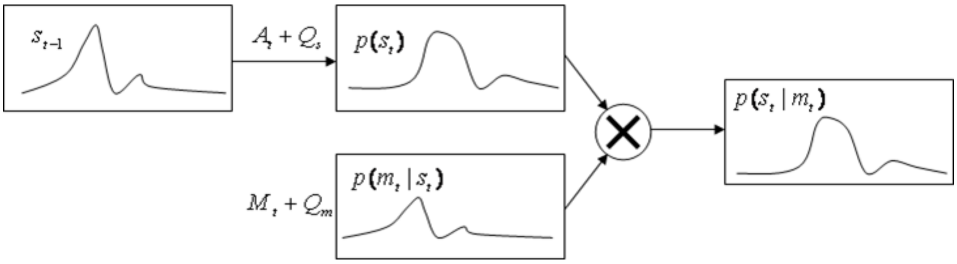


Figure 3.1: Refining estimation via recursive filtering. The probability distribution of the previous state s_{t-1} has been predicted to the next state $p(s_t)$ via adopted motion model; then it is refined using newly-arrived measurement m_t to obtain the final state distribution $p(s_t|m_t)$.

state vector may be at time s_t and we can do much better than simply making it uniform – we can use the previously computed states s^{t-1} to derive the prior for the current state.

In order to perform operations in a meaningful and tractable way, people almost exclusively use the Markov chain process assumption, which declares that state at time t directly depends only on state at time $t-1$, i.e.

$$p(s_t|s^{t-1}) = p(s_t|s_{t-1}, s_{t-2}, ..., s_0) = p(s_t|s_{t-1}) \tag{3.3}$$

The prior calculation, which is referred to as prediction, or the first step of the filtering, is simply the marginalization over the distribution of the previous stage:

$$p(s_t) = \int p(s_t|s_{t-1})p(s_{t-1})ds_{t-1} \tag{3.4}$$

Here, $p(s_t|s_{t-1})$ is the actual motion model we referred to above.

Overall, we conceptualize of filtering as a procedure consisting of two steps that are also schematically depicted in Fig. 3.1:

1. "Predict" the state at time t, by computing (3.4) using the underlying motion model
2. "Refine" the predicted state model with the latest measurement m_t by computing the likelihood $p(m_t|s_t)$

Once the general filtering framework has been outlined, more can be said about particular realizations. Note that the state vector consists of motion and structure parameters, which is usually huge; therefore, exhaustive modeling of the full distribution $p(s_t)$ may be out of the question.

3.3.3 Kalman filter

The Kalman filter is the most successful, well-developed and widespread filter used in control theory. The idea was outlined by Kalman [Kal60] in 1960 and has proved to be extremely useful over the course of history.

For our purpose, the Kalman filter can be perceived as a special case of the Bayesian filter where uncertainty can be modeled as a Gaussian distribution and both the motion model (3.4) and likelihood model are linear (3.1) in the state vector variables.

In this case, both prediction and measurement likelihood update stages can simply be expressed as linear equations of the form

$$\begin{aligned} s_t &= \mathsf{A}_\mathsf{t} s_{t-1} + \mathbf{u} + q_s, \\ m_t &= \mathsf{M} s_t + q_m, \end{aligned} \tag{3.5}$$

where the motion model and measurements are matrices $\mathsf{A_t}$ and M, respectively, and q_s and q_m are zero-mean Gaussian noise processes with corresponding covariance matrices Q_s and Q_m. Note the presence of the so-called "control" component $\mathbf{u}$, which influences the state vector in a known way and is needed for the case of robotics applications, but may not be needed in passive stereo sensing setups. The iterative update equations can be derived from (3.5) and can be found in many works [Kal60, YC90, FH01, Thr02, Hog].

Such an elegant formulation is able to take care of the entire conditional and marginal probability distributions associated with the state variables in an analytic form. However, these distributions are limited only to a Gaussian model. The particular disadvantage of this limitation is that it has only a single peak and is not heavy-tailed, which means that computations will not be robust to outliers. Another problem is that the state update and measurement equations are linear, which applies to a rather limited set of situations.

3.3.4 Extensions of Kalman filter

The assumptions that observations are linear functions of the state and that the next state is a linear function of the previous state are crucial for the correctness of the Kalman Filter and has to do with a the fact that any linear transformation of a Gaussian random variable is also a Gaussian [ER04]. Meanwhile, if the next state or measurement relationship are not a linear function of the previous state, the result is not necessarily Gaussian.

Since the state and measurement equations are rarely linear in practice, Kalman filter approximations have been considered that linearize state and measurement equations via Taylor series (the Jacobian Matrix serves as a major description of the relationship). This realization has been given the term Extended Kalman Filter (EKF). In essence, the EKF calculates a Gaussian approximation to the true belief and its goal is to efficiently estimate the mean and covariance of the state representation instead of the exact posterior (which may take an arbitrary form).

Since the EKF is just a linearized approximation to the true state posterior, it is practically advantageous to refine the estimations in a linear fashion. The idea is extremely similar to the Lucas-Kanade image registration technique, which ubiquitously uses the iterative scheme to refine estimated parameters [BM04]. This version of the filter has been named the Iterative Extended Kalman Filter (IEKF). It takes a newly estimated state s_t and iterates the estimation procedure again using the previous covariance matrix.

Another way to achieve the linearization of the state and measurement equations is to explicitly fit the linear regression model by stochastically sampling the mean state vector around the current mean estimate. This method has been coined the Unscented Kalman Filter (UKF). The method requires only a linear number of samples in terms of vector dimensionality, and, since it does not explicitly compute the Jacobian, the UKF and EKF are roughly computationally equivalent. However, UKF still models the state posterior only as a unimodal distribution; thus, it does not overcome one of the bottleneck simplifications of the original Kalman filter.

Information Filter

It is also worth acknowledging the parallel development of the so-called information filters (IF) and their corresponding extensions into Extended Information Filters, etc. The algorithm is very much like a Kalman filter but the inverse of the state covariance matrix is calculated and updated directly. Such a different parametrization allows for computationally simpler calculations, but does not change the scenarios or results of the KF and IF.

3.3.5 Particle filter – a non-parametric filter

Ultimately, we want to estimate the full posterior for the state vector in the least constrained fashion to yield the best possible robust and unbiased structure and motion parameters. A possible solution would be to represent the multidimensional state-vector distribution as a histogram and apply straight-forward update and measurement rules. In this case we would get a typical Histogram Filter [TBF05]. The obvious drawback of a histogram filter is the exponential increase in memory and computation with the increase of dimensionality as well as dependency on the histogram binning. A better idea is to stochastically represent the distribution by a set of samples and that is exactly what has been realized in the so-called Particle filter [GSS93, Kit96]. In computer vision it is also known under different names, e.g. Condensation [IB98].

Particle-filter algorithms represent a probability density through a weighted set of N samples (where weights are called importance factors). The samples are initially drawn randomly and evolved via (possibly non-linear) state and measurement equations. The specific implementation is subject to multiple details that are out of scope of this paper. Independent of these details, particle filters can be characterized by the following limitations:

- The quality of result is proportional to the number of particles, which is directly related to the computational and memory requirements. This is different from the analytic closed-form and efficient solution available for the Kalman Filter.
- The number of needed particles grows exponentially with the dimensionality of the state vector. In reality, only up to 3-4 dimensions can be reasonably represented with particle filters with current computational technology (i.e. having around 10-100 particles to sufficiently represent each dimension).

Interesting elaborations of particle filters have been proposed to deal with the shortcomings described above. One of them is the so-called Rao-Blackwellized Particle filter [KBD04], where the main idea is to exploit the fact that some of the dimensions of the state variables can be reasonably represented as a Gaussians – this allows use of the Kalman filters for the corresponding subspaces in the particle filter framework, but with greatly reduced dimensionality.

3.4 Optimization techniques

Both stereo and motion problems, as overviewed in Chapter 2, are ill-posed, because disparity and motion vectors must be recovered for each point; whereas, only one measurement (typically noise corrupted) per point is generally available. Fusion of stereo and motion faces essentially the same problem; therefore, efficient and effective enforcing of smoothness constraints is essential in any solution. Consequently, a large variety of techniques have been proposed to address this problem.

Here, we do not consider very restrictive model-based techniques (like affine motion, or 3D planar scene), which can be described by a finite number of parameters. Also, rigid motion estimation problems, as shown in Chapter 4, do not suffer from this particular problem, as it involves the solution of a finite number of unknowns, which is usually far less than the number of observation points. In both these cases, the problem is sufficiently over-constraint and standard Least-Squares (possibly robustified) formulations proved to be sufficient.

3.4.1 Local methods – intrinsic smoothing

These methods are almost entirely based on overconstraining the local matching formulation by summoning more data via local aggregation. Typical Sum of Squared Differences (**SSD**) type local

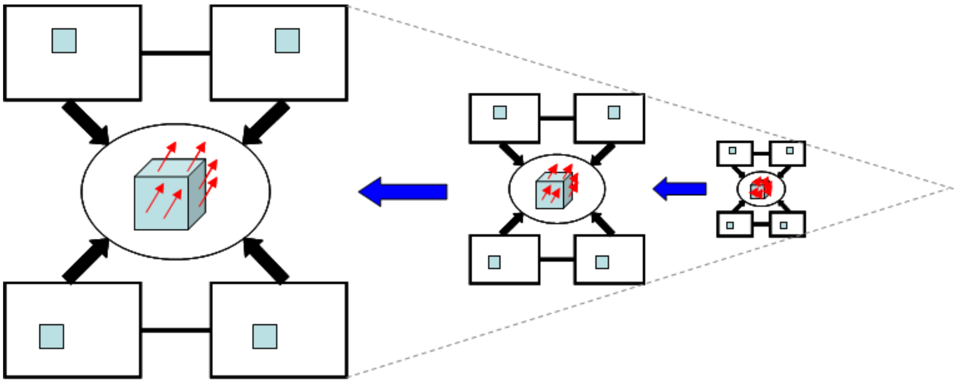


Figure 3.2: Coarse-to-Fine stereo motion processing. Image pyramids for the left and right time-consecutive frames are constructed. Estimation is performed progressively from coarse to fine levels, using the coarse level results as an initial estimates for the finer level refinement.

stereo matchers [SS02] and Lucas-Kanade-type optical flow estimators [LK81] are both excellent examples of local techniques. No additional explicit optimization needs to be performed by these methods; however, the solutions can be regularized further by virtually any method that we cover below.

3.4.2 Coarse-to-fine estimation techniques

A central idea in stereo and motion processing is coarse-to-fine computation, which will be abbreviated as **CTF** for brevity. Figure 3.2 depicts the paradigm when consecutive binocular stereo pairs are the input (4 frames) and 3D structure with 3D flow are the output.

The basic idea is to compute the necessary quantities using coarse scale data and refine them later using the same data but at higher resolution. This recursive refinement procedure has several very important advantages. Specifically, it helps remove local minima in correspondence search, as well as significantly reduces the storage and processing time, since large disparities at fine resolution can be recovered at coarse resolution with smaller search. Out of the disadvantages of CTF we can note the greedy approach that may be trapped in local errors, more uncertainty near 3D boundaries and difficulty in recovery of thin structures [SW06]. The first two flaws may be greatly diminished by careful computation in the CTF framework, like adaptive scalespace matching [SW09a]. Further, solutions for better thin structure resolution exist as well, e.g. [Siz08].

Coarse-to-fine has been used extensively in stereo [SW09a] and is of paramount importance in motion [Der06]. Additionally, CTF is needed for motion processing, because the differential photometric constraint is true for small motions only. Being largely responsible for the early success in stereo and motion, CTF is essential in joint stereo motion processing too – virtually any method, especially ones considering 3D range flow relies on CTF processing. Conveniently, in order to diminish the disadvantages of CTF it can be used only for interframe correspondence calculation (motion pairs), but not for intraframe correspondences (stereo pairs) for reliable 3D thin structures recovery, e.g. [HD07].

3.4.3 Differential priors – PDE

The pointwise brightness constancy constraint is simply not enough to perform estimation of disparity or motion vectors. A viable alternative to the local methods mentioned in Sec. 3.4.1 is to introduce priors for nearby image locations to regularize the solutions, such as force structure and motion vectors to be locally smooth. The priors themselves come from a variety of formulations that can be as simple as minimization of gradient (2.3) as pioneered in Horn-Schunk [HS81] or more involved robustified intensity-edge preserving penalizers [Der06].

The optimization problem formulation (2.3) is based on the fact that continuous, rather than discrete, entities are to be recovered, i.e. optical flow vectors (u, v) in this context. In turn, this problem is transformed into the task of finding the extremum of the 2D functional $f(u, v)$ that minimizes the integral:

$$E = \arg\min_{f(u,v)} \int_{\Omega} (I_x u + I_y v + I_t)^2 + \alpha^2 \left(||\nabla u||_2^2 + ||\nabla v||_2^2\right) dxdy, \tag{3.6}$$

where Ω is the working domain over which $f(u, v)$ is defined, usually being the whole image.

Minimizing this convex functional is a typical calculus of variations problem [Str86], which involves the solution of the corresponding Euler-Lagrange equations of the form

$$\begin{aligned} 0 &= \Delta u - \frac{1}{\alpha}\left(I_x^2 u + I_x I_y v + I_x I_t\right) \\ 0 &= \Delta v - \frac{1}{\alpha}\left(I_x I_y u + I_y^2 v + I_y I_t\right) \end{aligned} \tag{3.7}$$

where Δ is the Laplace operator ($\Delta := \partial_{xx} + \partial_{yy}$). Needless to say that careful numerical approximation to differentiation of I, u and v is crucial for the satisfactory practical performance.

The objective now is to simultaneously solve the system of these linear Partial Differential Equations, from which the name of this class of method largely comes – *PDE* solvers, or *variational methods*. Since equations (3.7) form a very large and sparse matrix, direct estimation such as Gauss-Jordan elimination is not applicable and various iterative methods, e.g. Gauss-Seidel or Successive Over-Relaxation (SOR), are widely used [HS81, BFB94, BWS05, PBB$^+$06]. Various multi-grid extensions to further increase the speed and robustness are also possible. Furthermore, the system of equations (3.7), being homogeneous, does not define the functional $f(u, v)$ uniquely and boundary conditions are necessary to disambiguate the solution.

Finally, the functional of the form (3.6) uses a rather simple regularization prior that has many shortcoming including the corruption of depth discontinuities. Various extensions that considers zero-crossings, intensity edge cue [Der06] or even local-aggregation based cues [BWS05] have been proposed. Obviously, a particular prior will result in different Euler-Lagrange equations; however, they will be solved in a similar way.

Since variational methods are extensively used both in stereo and motion literature, they have been successfully applied to stereomotion recovery as well [Sv02, PKFH03, PKF05, HD07, WRV$^+$08, SAS08].

3.4.4 MRF model

Regularization of the underconstrained problem is usually done by trying to enforce smoothness between neighbouring points, because the estimate in every point depends on each other, with the dependency typically diminishing with distance. Graphical models tremendously help to visualize and make use of this dependency. Here, the general idea is to declare each point (or pixel) as a graph node and dependency as a link. Considering all possible pairwise dependencies, also known as binary cliques, (not even considering higher order cliques) results in intractable formulations, as the number of connections grows at least quadratically with the number of pixels.

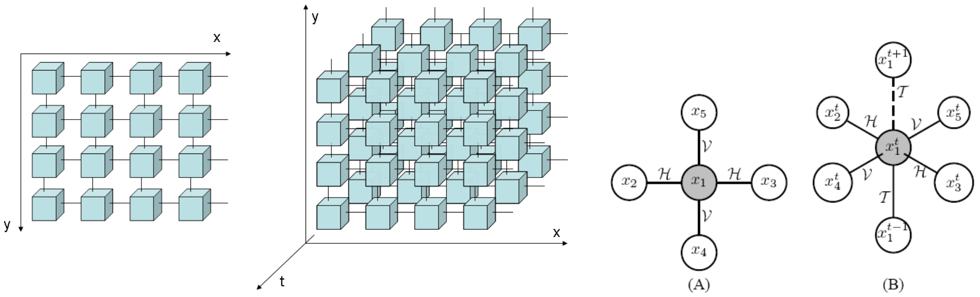


Figure 3.3: 2D spatial and 3D spatiotemporal MRF grids. Whole grids (on the left) as well as their basic generating elements (on the right, adapted from [WIM05]) are shown.

The idea which comes to the rescue is the Markov assumption, already used in predictive filtering of Sec. 3.3 for 1D chains. In the case of 2D, instead of considering all possible links, we consider only the immediate 4-connected neighbour of the graph, which results in a simple planar graph, as shown in Fig. 3.3. In the case of 3D [LALS04, WIM05] and even 4D MRFs, [IM06], we will consider 6-connected and 8-connected neighbors, respectively. Finally, it is important to note that graphs of higher connectivity are possible (e.g. 8-connected neighbourhood instead of 4-connected for 2D grids), but gain in performance is hard to achieve due to increased difficulty in actual optimization.

Regularization within the graphical model is achieved by forming a global energy function which consists of two terms

$$E = \sum_i E_i + \lambda \sum_{j \in \mathcal{N}(i)} E_{ij}, \tag{3.8}$$

where E_i is the dataterm (derived from the measurements) for each pixel and E_{ij} is the binary term that defines the energy dependency between points i and j from the neighbourhood, $\mathcal{N}(i)$. The relative weighting, λ, controls the amount of smoothing in the final solution. Each component E takes different values depending on the (disparity or flow) label the point takes. Note that the set of labels is discrete, which is quite favourable for computer realization and makes it different from the PDE-type solution derived for continuous data.

The last major question is how to optimize the described energy function. The rest of this section overviews major methods developed and used to date. Looking ahead, we mention that the choice and success of every method largely depends on the dimensionality of the problem (1D disparity, 2D image flow, 3D range flow), the graph (1, 2, 3 or 4D) and the form of the regularization prior (e.g. convex vs. robust).

Early endeavors

Forming and solving MRF models for various computer vision problems, like image restoration, segmentation and stereo per se received attention for sometime, with one of the seminal works being [GG84]. The original solution of the 2D MRF smoothness formulation was found via stochastic relaxation [GG84], which is a variation of the Markov chain Monte Carlo (MCMC) method – random sampling from the highly multidimensional distribution. Notably, this generic randomization method suffers from several flaws: sensitivity to initial condition as well as very slow and unproven convergence and, as a result, is far from practical.

Simulated annealing is an alternative method for solving very generic global optimization problems [KGV83]. The idea comes from the name itself, which performs multistage optimization at different “temperatures” gradually “cooling” the system and refining the estimation for lower temperatures. The notion of "temperature" is related to scale – thus, this idea has much in common with the coarse-to-fine procedure discussed previously. While the simulated annealing method in theory is capable of finding a global minimum and certain convergence properties can be proven [GKR94], the execution time is extremely slow and can even be worse than brute-force enumeration in the case of a discrete finite search space. That makes simulated annealing more of an archaic method that is not currently in wide use by the machine vision community.

"Diffusion" is one of the more recent methods that was applied in the context of stereo vision algorithms [SS98a]. The idea originates from the name of the method, where the local posterior distribution of the determined value, e.g. disparity, is the mix of the analogous posteriors of its neighbours (which are determined by links in the MRF graph). This intermixing goes on for multiple cycles virtually diffusing the pointwise likelihood distributions into the global posterior. While this method was able to show reasonable improvement over its contemporary analogues, it did not evolve further as it was overshadowed by Loopy Belief Propagation, discussed below. The latter is similar in spirit but has a more sound theoretical foundation and produces significantly better results.

Iterative conditional modes (ICM) [Bes86] uses a deterministic "greedy" strategy to find a local minimum – in essence, for each node, it chooses the label giving the largest decrease in energy, and repeats this process until convergence. The method is particularly sensitive to its initial conditions and prone to get trapped in local minima; at the same time it is very simple, fast, lowers the original energy and may be useful when the original condition is already very good. Significantly, ICM has been rigorously compared with many contemporary optimization methods on various computer vision tasks [SZS$^+$06, SZS$^+$08] and the comparison shows complete incompetence of ICM, which allows us to safely exclude it from the tools for joint stereo and motion processing.

Genetic algorithms (GA) [RN03] are another class of general-purpose AI algorithms designed to simulate evolution and make use of this principle to solve a variety of tasks. There is great freedom in choosing the particular evolution and mutation rules; however, the claimed versatility of the algorithm is not usually supported by sound theory and convergence results. We can find some utilization in the spatial stereo literature [GY01], but GA has not been used in stereomotion literature to our knowledge.

Dynamic Programming (DP)

Dynamic Programming (DP) is a classical technique in algorithm design that, like the divide-and-conquer method, solves problems by combining the solutions to subproblems. In the context of graph optimization, DP is ideally applicable to the tree arrangement, i.e. connected graphs without loops. Thus, 1D MRF chain graphs can be solved with DP optimization. Regarding 2D and higher-order graphs, while DP has been used directly on MRF grids pruned to tree structures [Vek05, LSY06], a more typical application of DP is found via organizing multiple linear paths over 2D and even 3D grids. One of the instructive examples that turned out to perform well in practice is the semi-global stereo matcher [Hir05] that constructs disparity correlation space by traversing the MRF 2D grid along 16 orientations over 2D space. Finally, DP has already been successfully utilized in the stereo motion research [LALS04, WIM05] and further application of DP is expected due to its speed and relative simplicity.

Belief Propagation (BP)

Belief propagation (BP) is an iterative algorithm that solves the graphical model at each node by passing at each iteration its posterior distribution, called the message, to its neighbours and in turn updating its own posterior using analogous messages from the neighbouring nodes. Originally, it was perceived as a technique somewhat equivalent to DP, since BP is guaranteed to converge and find a global minimum only on graphs without loops. However, the completely local nature of the algorithm allows it to execute on graphs with loops, which may perform extremely well in practice even in absence of theoretical proofs, as demonstrated in the case of stereo matching and image restoration [FH04, SZS$^+$08]. Thanks to its versatility, BP is already used quite heavily in contemporary stereomotion research [WIM05, IM06, LMPF07] to optimize over 3D and even 4D graphical models.

Graph Cuts (GC)

Graph Cuts (GC) was first introduced to the computer vision as a method that finds the global minimum for 2D MRF grids with binary labels. The method was quickly extended to handle the multiple linear label problem (1D disparity scalars), and was still able to find the global minimum, if the prior term was convex [Ish03], or guaranteed to find a local minimum not greater than twice the global minimum in the case of the robust Potts smooth prior model [BVZ01]. Currently, GC is the best optimization for 2D MRF formulations, virtually independent of its application [SZS$^+$08] with the most success achieved in segmentation, image stitching and stereo processing. Unfortunately, the aforementioned nice theoretical and experimental results do not generalize to either MRF of high order (e.g. 3D or 4D) or higher dimensional problems (like 2D optical flow or 3D range flow). Furthermore, GC is quite slow and is difficult parallelize. Nevertheless, GC can be used in stereomotion processing and has already been successfully applied in a spatiotemporal stereo method of [SW09b].

Future trends

Despite vast utilization and constant improvement of the aforementioned MRF optimization strategies, new interesting trends are under constant development. Below we briefly outline two examples that have been brought into the vision community and have already shown their usefulness.

The tree-reweighed message passing (TRW) technique originally developed by [WJW05] was inspired by the problem of maximizing a lower bound on the energy. The basic idea is to perform optimization on various tree decompositions of the 2D MRF grid and combine results in a convex fashion. The method is reported to achieve close to state-of-the-art results [Kol06, SZS$^+$08]. An important advantage of the method is the ability to estimate the lower bound of the energy – thus, we can get a feel for how good the solution is; moreover, in many situations it can find the proven global minimum [MYW05].

The Quadratic Pseudo-Boolean Optimization QPBO algorithm [RKLS07], addresses the basic binary label optimization problem. This method takes the interesting approach of changing the labeling to ternary – the objective now is to label nodes as 0, 1, or "unknown". The method is of great practical interest, as such labeling algorithms already exist, and in many situations there will be many nodes that can will be unambiguously labeled 0 or 1. The second stage of the QPBO procedure is to resolve the "unknown" nodes by trying various heuristics and even applying labeling results of other algorithms. The method is adaptable to virtually any type of energy functions and has already been successfully used in stereo processing to enforce a second-order smoothing prior model [OWF08] and optical flow to combine discrete and continuous optimizations [LRR08]. This method may be well suited to general non-rigid stereomotion processing.

Chapter 4

Rigid Scene and Egomotion

4.1 Outline

This chapter will extensively cover the problem of structure and motion recovery in the case of rigid scenes. This constraint is quite powerful and restrictive, but finds many practical applications. First, the modeling of rigid structure and motion is discussed and followed by appropriate estimation procedures for the monocular, and especially binocular case. Specific useful realizations, e.g. working in disparity space and "direct" methods for structure and motion, will be discussed. Description of various ways to deal with multiple frames acquired over time will follow. The chapter ends with a discussion of the role of calibration in the rigid structure-from-motion situation.

4.2 Rigidity as the constraint on the scene

The problem of simultaneous stereo and motion recovery has quite a few degrees of freedom, which means it should be effectively constrained. A special constraint which is of paramount practical importance is the rigidity of the scene, i.e. objects are not deformable and do not move relative to each other. This situation is physically realized by moving the rigid object in space (with camera being stationary) or moving the camera in a static environment.

The first case is more related to 3D object modeling when the camera observes the moving 3D object. Sometimes the motion is known or severely restricted, which gives rise to a set of special algorithms with interesting properties, as will be discussed in Sec. 4.4.3. The second case is more related to navigation in the field of robotics and is well known as simultaneous localization and mapping (SLAM), which will be briefly discussed in Sec. 4.7.3.

Availability of stereo data has been extensively used by previous researchers to get the 3D structure from stereo for frames consecutive in time, and then estimate the ego-motion between these frames from the recovered 3D points. Out of many other advantages, the egomotion estimation in this case becomes much simpler and results in linear methods. The following sections will describe in depth various solutions and how they are realized.

4.3 Modeling of rigid structure and motion

It is usually convenient to consider the scene static and the camera undergoing rigid motion. Rigid motion itself is completely characterized by the 3D rotation and 3D translation, provided that intrinsic camera model parameters stay unchanged during the motion capture sequence. However,

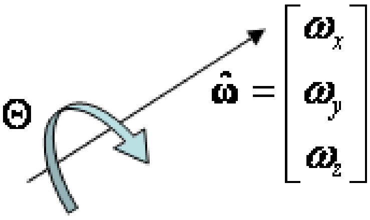

Θ
$\hat{\omega} = \begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix}$

Figure 4.1: Rotation matrix parameterized by axis of rotation $\hat{\omega}$ and the angle of rotation Θ around this axis according to the right hand screw rule.

the latter requirement can be relaxed to allow more flexible motion capture, which might be initially uncalibrated – discussion of this matter is delayed to Sec. 4.8.

The rigid motion of a 3D world point (or rather a collection of points) $\mathbf{P}$ is described in terms of rotation R and translation $\mathbf{V}$:

$$\mathbf{P_1} = \mathsf{R}\mathbf{P_0} + \mathbf{V} \tag{4.1}$$

The same statement can be expressed in homogeneous coordinate space, which is useful in a variety of situations we consider later in the paper:

$$\begin{bmatrix} \mathbf{P_1} \\ 1 \end{bmatrix} \simeq \begin{bmatrix} \mathsf{R} & \mathbf{V} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{P_0} \\ 1 \end{bmatrix} \tag{4.2}$$

Here, $\mathbf{V}$ is the translation 3D vector, which could be arbitrary and is arguably the best way to model translation. On the contrary, rotation matrix R must be orthonormal with unit eigenvalues, i.e. this 3×3 rotation matrix has only 3 degrees of freedom. The reliable computation and temporal update of the rotation matrix directly is challenging and numerically unstable. That is why there exist several alternative representations that have found utility in computer vision. The most important ones are Infinitesimal motion, Quaternion, Rodriguez formula, Screw decomposition, which are abundantly described in the literature, e.g. [HZ04].

The choice of representation for R is important because it usually dictates the estimation procedure. A particularly good choice from the pedagogical point of view is the Rodrigues rotation formula, as it explicitly involves the axis of rotation

$$\hat{\omega} = \begin{bmatrix} \omega_x & \omega_y & \omega_z \end{bmatrix}^\top$$

and the angle of rotation θ around this axis, as visualized in Fig. 4.1. In these terms, the rotation matrix becomes

$$\mathsf{R}_{\hat{\omega}}(\Theta) = \mathsf{I} + \Omega \sin\Theta + \Omega\,(1 - \cos\Theta)\,, \tag{4.3}$$

where

$$\Omega = [\hat{\omega}]_\times = \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix} \tag{4.4}$$

Another particularly important case of the Rodrigues formula is the special case of small motion. In this case the rotation angle Θ along the axis is small such that $\cos\Theta = 1$ and $\sin\Theta = \tan\Theta = \Theta$, i.e.

$$\mathsf{R} = \mathsf{I} + \Omega\Theta = \begin{bmatrix} 1 & -\omega_z\Theta & \omega_y\Theta \\ \omega_z\Theta & 1 & -\omega_x\Theta \\ -\omega_y\Theta & \omega_x\Theta & 1 \end{bmatrix}. \tag{4.5}$$

Further convenient manipulation may involve subsuming Θ in to the modulus of $\hat{\omega}$, i.e. make its modulus Θ rather than 1.

Throughout the rest of this chapter we will mainly stick to this infinitesimal motion constraint, as many situations comply to the small motion assumption and it allows for tractable linear solutions, as described in Sec. 4.4.

4.4 Solving for rigid structure and motion

4.4.1 Monocular case

In order to better visualize the advantages of stereo setups in rigid motion recovery, we first derive the case for monocular structure from motion.

Having points matched across the views, the rigid motion, Eqn. (4.1), needs to be rewritten in terms of image coordinates. Let Z_0 and Z_1 be the depths of the point $\mathbf{P}$ at time instances zero and one, respectively. Then we can express the world point coordinates via their image projection coordinates (under the unit focal length assumption) as the following:

$$\mathbf{P} = \begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} = P_z \begin{bmatrix} P_x/P_z \\ P_y/P_z \\ 1 \end{bmatrix} = Z\mathbf{p} \tag{4.6}$$

The rigid motion Eqn. (4.1) becomes

$$Z_1\mathbf{p_1} = Z_0\mathsf{R}\mathbf{p_0} + \mathbf{V} \tag{4.7}$$

which is really

$$\mathbf{p_1} = \frac{Z_0}{Z_1}\mathsf{R}\mathbf{p_0} + \frac{\mathbf{V}}{Z_1}. \tag{4.8}$$

Note that translation $\mathbf{V}$ is recoverable only up to a scale factor, as it is coupled with the unknown Z_1 (a well known fact). But the major problem comes from the fact that the unknows $\frac{Z_0}{Z_1}$ and R are coupled, which means there is no simple elegant solution to solve for those simultaneously. Luckily many iterative schemes, e.g. alternating structure (e.g. Z) and motion (e.g. R and $\mathbf{V}$) recovery have been developed in the past and can be found in many places [HW88, Hee90, Han91]. Subspace methods also are applicable to “splitting” the somewhat complicated equations that relate visual motion with 3D structure and motion into a simpler set of equations for solution [HJ92].

4.4.2 Multi-ocular case

Consideration of binocular (and generally multi-ocular) stereo setups allow us to significantly advance on the problem of structure and motion computation outlined above.

Immediately, the major improvement comes from the fact that (calibrated) stereo allows for the direct recovery of depth. Now we are able to operate on (4.1) directly, because we can simply treat the derived points $\mathbf{P_0}$ and $\mathbf{P_1}$ as observations, i.e.

$$\begin{bmatrix} P_{x1} \\ P_{y1} \\ P_{z1} \end{bmatrix} = \begin{bmatrix} 1 & -\omega_z & \omega_y \\ \omega_z & 1 & -\omega_x \\ -\omega_y & \omega_x & 1 \end{bmatrix} \begin{bmatrix} P_{x0} \\ P_{y0} \\ P_{z0} \end{bmatrix} + \begin{bmatrix} V_x \\ V_y \\ V_z \end{bmatrix} \tag{4.9}$$

The constraint becomes a simple system of linear equations for motion parameters[1].

$$\begin{bmatrix} 0 & P_{z0} & -P_{y0} & 1 & 0 & 0 \\ -P_{z0} & 0 & P_{y0} & 0 & 1 & 0 \\ P_{x0} & -P_{y0} & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \\ V_x \\ V_y \\ V_z \end{bmatrix} = \begin{bmatrix} P_{x1} - P_{x0} \\ P_{y1} - P_{y0} \\ P_{z1} - P_{z0} \end{bmatrix} \tag{4.10}$$

Having multiple measurements (i.e. triangulated points), it is possible to solve the corresponding multiple constraints in a linear system of equations (that always have a finite number of unknowns) in a Least Squares sense to get the absolute quantities for the rigid motion parameters.

To simplify the estimation procedure even further, it is possible to separate estimation of rotation from translation. This can be done by subtracting out the mean of the 3D coordinates from the point clouds at two time instances and estimating the rotation matrix in the LS-sense. The 3D translation is simply the difference vector between the mean of the 3D coordinates at times 0 and 1, which happens to be the centre of mass of the system. Interestingly, this approach follows instantly from a closer look at the matrix structure of (4.10), where the first 3 × 3 subpart is dedicated to rotation and the last 3 × 3 subpart happens to be the identity matrix.

The schematic outline of combining stereo for 3D structure estimation and motion for the subsequent rigid motion parameter computation is quite straightforward and has been extensively used throughout the literature [BK83, HB85, KA87, ZHZ88, YC90, LK90, LH91, WCR92, Koc95, HRD$^+$99, OMSM00, MB00, DD02, NNB04, ZN04, AKK05, Der06, RFG07]. While the aforementioned methods are quite different in their final goals, matching features, choice of predictive filtering framework, and so on (these differences will be elaborated later in this chapter), all of them rely on the linear rigid motion parameter extraction stage from spatially and temporally matched primitives.

Nevertheless, alternatives have been developed. Some work simply performs two separate structure-from-motion computations on the left and right views treating them as monocular sequences; the results are later fused using a Kalman Filter similar to the one described in Sec. 4.7.3 [BE95]. The stereo motion problem also has been treated as a fusion problem to compute 3D structure – estimation is done using stereo, then structure-from-motion using one of the views and results are combined using the same Kalman Filter [GST89]. Other work considered the rather contrived case of single planar surface estimation using stereo motion [AH90]. Nevertheless, it results in a particularly simple and stable method that does not require any explicit correspondences – a linear relationship of points belonging to a plane is enough to formulate and compute the 3D description of the plane and the direction of its translation.

A more interesting and quite different approach to rigid motion estimation from left and right image flows has been proposed in [WD86] and improved upon in [WD96]. There, the binocular flow constraint has been derived and stereo matching is performed on the flow vectors directly by aggregation; the constraint itself states that the ratio of disparity change to the disparity itself must stay constant in the local region. The combination of binocular flow matching with traditional stereo matching using appearance-based features is also possible [WD86]. This approach is the most interesting from a theoretical perspective, as the requirement of supplying left and right flows is too restrictive to be practical – afterall, optical flow is not directly observable, and its computation might be more challenging than stereo itself, as described in Chapter 2.

[1] Note that similar reasoning is applied to the other representations of rotation, such as the full matrix R that must be re-normalized to be orthonormal once the corresponding 9 parameters have been recovered in the linear manner (as in [Der06]).

Finally, it is also possible to make use of the original rigid motion Eqn. (4.7) without first solving for the depth via a stereo setup. This class of methods tries to utilize the constraints earlier in the estimation and directly estimate the needed parameters via by-passing the computation of intermediate quantities (which might be error prone and noisy) – they were given the name "direct methods" and are discussed in detail in Sec. 4.6.

4.4.3 Rigid scenes with known or restricted camera motion

When motion is known, we effectively have the equivalent of the multiview stereo setup and we would be better off by exploring more specialized solutions from that particular area (for examples, consult [GCS06, SCD+06]). This analogy can indeed be explored more directly in special cases of (known or constrained) motion that may allow for certain neat algorithmic solutions.

A typical motion for the acquisition of the 3D models of small and medium size objects is the turn table. Specifically, constant rotational motion of the turn table makes "cyclograph" stereo [SK02] possible – spacetime slices from videos are extracted and stereo is performed on these multi-perspective images. The result is a full 3D model without explicit need of combining 3D pieces acquired from different image pairs.

Another useful type of controlled motion is a simple translational motion, as on conveyor belt in the $x - y$ plane of the camera. Such motion excludes in-depth motion and makes stereo matching on simultaneously many frames possible without taking care of the discrepancy of image flows. Thus, special algorithms can be derived, as the DP method [MKC00]. This particular technology has also been termed "motion stereo" to exemplify that stereo pairs for matching arise from the controlled motion [Nev76, MKC00, PKRC00].

4.5 Working in disparity space

While most papers have been pre-occupied with rigid motion characterization in 3D space, certain advantages exist by performing computation in disparity space directly, which is native to stereo processing. An important difference of the latter methods is the absence of the triangulation step to "upgrade" disparity information to 3D coordinates, which suggests the following benefits:

- Meaningful triangulation assumes some kind of calibration. It may be Euclidean, affine, or missing, which would correspond to metric, affine, and projective reconstructions, respectively. This dependency of triangulation on calibration is a burden for any structure from stereo and motion algorithm and reasoning in disparity space becomes very handy to remove this problem altogether.

- Noise modeling is of critical importance to all computer vision algorithms. Since the conversion from the disparity-camera to the world coordinate system is a non-linear process, it results in rather complex error distributions with respect to the 3D points, even if the noise on image coordinates was reasonably approximated by a Gaussian (or other more realistic, e.g. heavytailed, but tractable distribution).

The general idea is simple. Though the relationship between 3D coordinates X, Y, Z and image coordinates x, y, d are non-linear and expressed as

$$\begin{aligned} X &= x * b/d, \\ Y &= y * b/d, \\ Z &= f * b/d, \end{aligned} \qquad (4.11)$$

with f being focal length and b being baseline between the left and right cameras, once we upgrade image and world coordinates into homogeneous representations, it is possible to get the equivalent *linear* relationship:

$$\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \simeq \frac{1}{f} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & f \\ 0 & 0 & 1/b & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ d \\ 1 \end{bmatrix} = \Gamma \begin{bmatrix} x \\ y \\ d \\ 1 \end{bmatrix}, \text{ where } \Gamma = \frac{1}{f} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & f \\ 0 & 0 & 1/b & 0 \end{bmatrix} \tag{4.12}$$

Considering that a 3D point undergoing rigid transformation can be written as a matrix operation in the homogeneous space as in (4.2), we can directly express the rigid motion in terms of image/disparity coordinates as a transformation according to a homography

$$\begin{bmatrix} \mathbf{p}_1 \\ 1 \end{bmatrix} \simeq \Gamma^{-1} \begin{bmatrix} \mathsf{R} & \mathbf{V} \\ 0 & 1 \end{bmatrix} \Gamma \begin{bmatrix} \mathbf{p}_0 \\ 1 \end{bmatrix} = \mathsf{H} \begin{bmatrix} \mathbf{p}_0 \\ 1 \end{bmatrix} \tag{4.13}$$

with the consequence of being able to use all convenient and well-developed algorithms for homography estimation [HZ04]. This convenient encapsulation was originally used for 3D rigid motion processing by [DH00] and later improved upon and realized in real-time in [AKK05]. The more general rigid structure and egomotion estimation problem has been developed in [DD02] and enhanced by later contributions like [Der06] to include more advanced statistical noise models.

4.6 "Direct methods" for stereo and motion

The majority of algorithmic solutions for rigid 3D motion estimation are preoccupied with inference from image motion (either dense, or feature-based). This 2D motion is usually assumed as observed, but in reality it has to be computed first.

The idea, pioneered by Horn and Weldon [HW88], is to combine these two steps into one and avoid accumulation of errors that inevitably arise in practice. In its original form, the method tries to combine the gradient-based optical flow constraint derived in a previous chapter and rigid motion estimation. This is essentially achieved by differentiating the (infinitesimal) rigid body motion representation (4.1) as in e.g. [BBHP92]:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \frac{1}{Z} \begin{bmatrix} -f & 0 & x \\ 0 & -f & y \end{bmatrix} \mathbf{V} + \begin{bmatrix} (xy)/f & -(f^2+x^2)/f & y \\ (f^2+y^2)/f & -(xy)/f & -x \end{bmatrix} \Omega, \tag{4.14}$$

where Z, Ω and $\mathbf{V}$ are structure and motion parameters, as before. The idea became instantly popular and developed further to include multiple moving frames [Hee90], other forms of brightness constancy constraints and much more [Han91, IA99].

More importantly, direct methods can be extended by simultaneous consideration of more views as in [HO93, MSS99]. It is worth noting that this strategy is applicable to having more than two cameras, and does not substantially add fundamentally new algorithmic advantages [HO93], since stereo image pairs and motion image pairs are considered in the same way. The main gain comes in terms of results, because stereo and motion offer different information, such as structure estimation mainly coming from stereo near motion's focus of expansion, or estimates coming from non-epipolar aligned motion when image texture is epipolar-aligned for stereo.

Nevertheless, the major issue of having structure, Z, and translation motion component, $\mathbf{V}$, coupled remains, which means that uncoupled solutions are still not available and all methods that use the direct rigid formulation (4.14) are forced to perform recursive alternating estimation of structure and motion in one way or another.

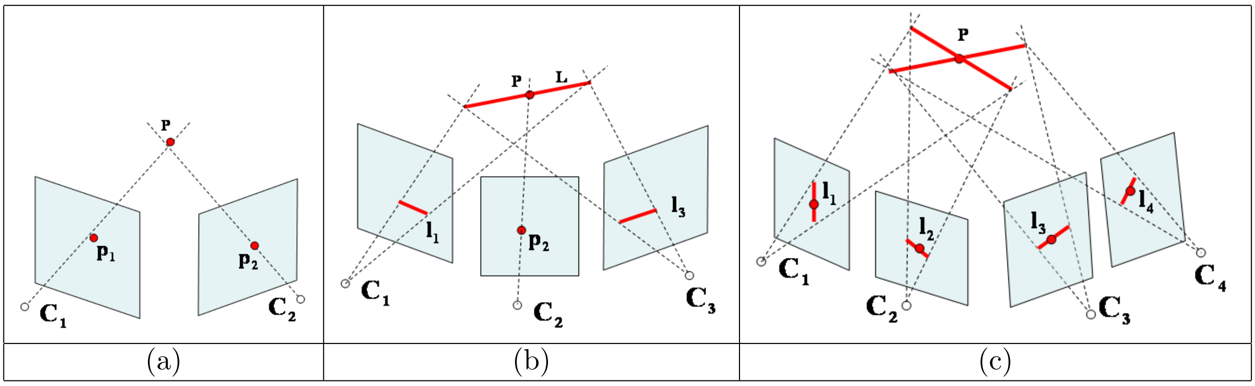


Figure 4.2: Fundamental constraints for multiple views. Intuitively, multiview constraints characterize the exclusive arrangements of lines and planes in 3D space. (a). Epipolar constraint between two points. Image points $\mathbf{p}_1$ and $\mathbf{p}_2$ form 3D lines $\mathbf{C}_1\mathbf{p}_1$ and $\mathbf{C}_2\mathbf{p}_2$, respectively. These lines must intersect at a 3D point P, which is an exclusive configuration, because two arbitrary 3D lines do not intersect. (b). Trifocal constraint between a point and two lines. Image lines $\mathbf{l}_1$ and $\mathbf{l}_3$ define 3D planes $\mathbf{C}_1\mathbf{l}_1$ and $\mathbf{C}_3\mathbf{l}_3$, respectively, that intersect along the 3D line $\mathbf{L}$. Point $\mathbf{p}_2$ in the second view defines another line $\mathbf{C}_2\mathbf{p}_2$ that must intersect with $\mathbf{L}$. (c). Quadrifocal constraint between four lines. Four lines in four views define four planes in 3D that must intersect at a single point. This is an exclusive configuration, because even if three general planes always intersect at some mutual point, this point generally does not belong to the fourth plane.

A step toward linear motion estimation was taken in [SZS94], where the authors estimate depth maps from stereo and consider Z as known when linearly solving for (4.14) via Kalman filter updates. In this case we can perceive this method as an improvement over the standard method outlined in Sec. 4.4 that does not require temporal matching, but still requires spatial. In other words, stereo significantly helps motion estimation, but motion does not help stereo at all.

An interesting avenue has been explored by [SS98b] to directly combat the coupledness problem of stereo and motion parameters by considering the stereo setup. There, the authors manage to find a linear solution, which simultaneously exploits stereo and motion constraints in linear fashion. A simple idea was to express the infinitesimal brightness change constraint for 3 pairs of views (out of 4 possible images in the time-consecutive stereo pairs). Essentially, it results in equations of the form (4.14) for different image pairs that have different motion parameters (because stereo cameras have some known non-zero rigid transformation between then) but same depth, Z, (as they describe the same point) – marginalizing Z gives rise to linear constraints that describe only the motion parameters in the uncoupled form. The development of such algorithms is especially beneficial from a practical point of view and could be perceived as one of the main intrinsic goals of the stereo-and-motion area.

4.7 Working with multiple frames

4.7.1 Multiple-frame constraint

As extensively discussed above, the structure and motion parameters can be estimated from a time-consecutive stereo pair. Obviously, results from the next consecutive stereo pair are in strong correlation and must be combined. However, prior to pursuing this path, let us pay attention to the fact that there exist fundamental multi-view constraints beyond the two view epipolar constraint that can be utilized.

It is well known by now that there exist fundamental relationship between 2, 3 and 4 views. There are no further fundamental constraints for 5 and more views, aside from the aforementioned 2, 3 and 4-view constraints [HZ04]. All these fundamental constraints are depicted in Fig. 4.2.

The binocular constraint states which points may be in correspondence. The constraint has a paramount utilization in stereo matching – it basically reduces the stereo correspondence search to a 1D search problem. Moreover, the constraint is stated for points, which is useful for pixel-based algorithms that are currently popular.

The trifocal constraint links a point and two lines in 3 images. Technically, this constraint can be used in methods that combine point and line correspondences, e.g. [KA87]. In practice, it found a neat utilization in a direct method of [SS00] – the authors solve for rigid structure and motion using a point in a reference view and brightness isocontour lines (usually derived from the image gradients) in the other two time-consecutive views. Obviously, pairwise epipolar constraints are still valid for a triple of frames and are utilized by many algorithms. Another early statement of the constraint relating three images in somewhat different terms can be found in [SA91].

The quadrifocal tensor links 4 lines in 4 images. While one may consider applying it in time-consecutive stereo frame pairs for brightness isocontours (similar to [SS00]), the quadrifocal constraint seems never to be applied in such a way, especially in the case of the stereomotion estimation problem. Interestingly, it has been shown in [SW00] that the quadrifocal tensor can be constructed from two trifocal tensors and one fundamental matrix – this decomposition has been taken advantage of by [CMR07] for the purpose of robust egomotion estimation from two consecutive binocular stereo frames.

Thus, despite the interest and seeming usefulness in application to stereo motion, which naturally involves at least 4 frames considering consecutive time instances, the idea of explicitly using trifocal and quadrifocal tensors has not seen sufficient utilization. While there may be many reasons for this fact, the following may be the most profound:

- The epipolar constraint for two images is the minimal constraint which is stated for points; trifocal and quadrifocal constraints involve lines. Since most of the methods, especially the most recent ones, are interested in dense pixel-based reconstruction, constraint over lines has less obvious utility in those frameworks.

- Trifocal and quadrifocal tensors are entities of high order – third and fourth, respectively. Thus, any attempts to simply estimate them will struggle with severe noise amplification in the estimation process; likewise, any attempts to express them in terms of meaningful parameters (such as rotation and translation values) will result in high-order combinations that are challenging to decompose.

- Owing to relatively recent discovery and limited use of higher order tensors, these structures are not yet well understood and further theoretical investigation is very desirable.

In conclusion, we will not return to the discussion of inherent multi-frame constraints and leave the ultimate question of their practical benefits open.

4.7.2 Structure-from-motion as factorization

The rigid motion transformation (4.1) is a linear transformation by itself. The non-linearity arises only in the camera projection stage, when we relate the world point $\mathbf{P}_1$ to the corresponding image point $\mathbf{p}_1$. However, if the camera were affine (with 2×4 intrinsic matrix K_a), then the relationship of 3D world coordinates, motion and the directly observed image points would have been completely linear:

$$\mathbf{p} = \mathsf{K}_a \begin{bmatrix} & \mathsf{R} & & \mathbf{V} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{P} \\ 1 \end{bmatrix} \tag{4.15}$$

Thus, all measurements $\mathbf{p}$ could be collected into one big matrix and factored into 2 rank-4 matrices that would correspond to the uncoupled motion and structure parameters.

The original formulation along these lines was proposed for orthographic cameras [TK92], which was later generalized to affine cameras [PK94] and handling of missing data (occlusions and deocclusions of features); furthermore, a few (iterative) algorithms for projective factorization exist as well [ST96]. The obvious peculiarity of such methods is their truly simultaneous recovery of structure 3D points and rigid motion parameters, i.e. treating the problem in a calibration-estimation bilinear fashion [Kv97].

Along these lines, an interesting approach to consider is [HC00] as it extends factorization to the stereo motion case resulting in truly fused structure and ego-motion estimation. The formulation is done for affine cameras only, which limits its practical use. Interestingly, stereo correspondences are not needed to be found explicitly – measurement matrices on which factorization is to be performed for left and right cameras have been collected independently and aligning their columns via rank analysis would be equivalent to finding stereo correspondences.

4.7.3 Two-frame constraints – batch process and predictive filtering

A natural extension to the instantaneous rigid two-motion estimation is the incorporation of multiple frames over time. One of the ways would be to organize the Least-Squares system for multiple frames all at once, as has been suggested in [Gov01], where all possible two-frame constraints from monocular sequences are used. The described LS minimization of the appropriate global error functional is also referred to as batch processing [BC91].

A much more typical situation is to process the data as it is coming in – it is more natural and has the needed flexibility of not over-extending the assumption of constancy of the motion parameters. Nevertheless, batch processing may still be used to initialize the estimation for the first few frames, as in e.g. [WCR92].

Generally speaking, our goal now is to recover the structure parameters Z and motion parameters Ω, $\mathbf{V}$ having multiple (stereo) frames in time as our data measurements. In the current situation, we would get estimates for every time instance t – the results will depend on previous computations of Z, Ω, $\mathbf{V}$ and current measurements. Moreover as noise and matching errors are attributes of any real systems, it is necessary to find the uncertainty of the associated estimates. To be even more general, it is interesting to recover the whole distribution of Z, Ω, $\mathbf{V}$ at each time instance in order to completely describe a solution. In the case of a stereo setup, the depth can be obtained directly from stereo; thus, it may be considered enough only to robustly estimate motion parameters, which significantly constrains the state set to a small number of dimensions, e.g. 6 in our case for Ω and $\mathbf{V}$ (it could be more in case when quaternion representation is used for rotation, or if precession is to be estimated as in [YC90], or even rotational acceleration [BE95]).

Out of various filtering architectures, the Kalman filter, even in its basic formulation, has received the most widespread attention and has been applied to the recovery of rigid structure and motion in a variety of work over decades [GST89, LK90, Koc95, BE95, YO97, FRBG05, AKK05, RFG07]. To overcome the linearity restriction, Extended [CMR07] and Iterated Extended Kalman Filters have been used in the stereomotion literature as well [YC90, WCR92]. A version of a robust Kalman Filter that operates under a heavy-tailed noise model (constructed as a mixture of broad and narrow Gaussians) also appeared [TSJ92].

Importantly, the technology of rigid structure and motion computation from visual information has been widely used in robot navigation in solving the Simultaneous Localization and Mapping (SLAM) problem. The basic idea of SLAM is to sense the environment, built its map and localize the robot in this environment with subsequent purpose of navigation. We cannot possibly outline even major works and will largely refer to the numerous surveys and discussion papers, e.g. [FH01, Thr02, Hog]. It is important in the context of this paper that SLAM is not directly interested in the problem of efficient and joint estimation of structure and ego-motion perse (i.e. combining stereo and motion) but rather just uses these results as measurements and subsequently works on them in some filtering framework to achieve robust and fast performance. The bottom line is that sensing is typically disjoint from the SLAM estimation, but that is what computer vision is the most interested in (although, some recent work has started to couple these matters more closely [Hog]). Thus, SLAM solutions are not directly applicable to solving the fundamental problem of spatial and temporal correspondences, as pursued in this paper.

4.8 The role of calibration

Having moving stereo camera(s), we get fully metrically calibrated estimation of 3D structure, which is generally more reliable than what is recoverable from a monocular moving camera. However, in many practical situations we may not have the luxury of relying on calibrated stereo setups: Calibration may be unknown or in need of continuous update, e.g. in various complicated frameworks with moving parts (an active stereo head) or harsh/varying environments (temperature drops, mechanical vibrations, or even change in sensing medium in the case of an amphibious autonomous vehicle). Finally, the human visual system demonstrates the peculiarity of self-calibration – when normal vision is altered by the use of external optics such as prisms, glasses, video see-through displays, hand-eye coordination will be very quickly restored after minimal interaction with the environment.

Interestingly, the problem of calibration has been indirectly addressed by Richards [Ric85]. The author showed via theoretical analysis that neither stereopsis nor motion parallax alone can recover the 3D shape when it is observed by two [orthographic] cameras with unknown vergence[2] – surprisingly, the linear solution to the whole problem is available when motion and stereo are integrated. In other words, 3D shape is recoverable from integration of stereo and motion without any prior calibration whatsoever.

So, it is of interest to the community to be able to operate the sensor without assuming its calibration, which can be principally explored in two ways: (i) what can be computed and inferred without calibration and (ii) how can calibration be performed automatically. Certainly, solving the second problem will allow us to forget about the first, but resources, time and effort will be conserved if calibration will be done only where it is really of essence.

4.8.1 Working with uncalibrated setups

Having only two perspective cameras which we do not know anything about (i.e. cameras are uncalibrated, or the corresponding *intrinsic* camera parameters are unknown), it is possible to find points in correspondence and make certain conclusions about the underlying observed 3D structure. In particular, it is well-known that it is possible to recover the scene up to a projective transformation from two uncalibrated views [HZ04], i.e. from the uncalibrated setup.

[2] Note, that only angular disparities are directly observable in this case.

These results essentially mean that stereomotion analysis can be performed without calibration and its results will be true up to some projective transformation. Specifically, it is possible to recover structure and motion and even segment objects that undergo different motions [DH00, DD02]. The actual computations are usually done directly in disparity space (rather than in 3D space, transformation to which actually requires the knowledge of calibration) and is discussed in further detail in Sec. 4.5.

More specific use of uncalibrated setups was demonstrated with respect to recovery of non-metric quantities. For example, [GT95] accepts the unit measure of the 3D world in terms stereo baseline and uncovers the time to impact to an obstacle. In their case robustness is much more important than accuracy – use of stereo setups is justified from this position, plus calibration will not improve on the robustness of the sensor.

4.8.2 Autocalibration

Nowadays it is a well-known fact in the computer vision literature that it is possible to calibrate a moving camera just from several views (at least 3) if it observes a *general* rigid scene [HZ04]. The topic of autocalibration of a single moving camera has received much attention; here we are mainly interested in the calibration of a moving stereo setup. As a representative example, [AP95] estimates structure, motion and focal length (intrinsic camera parameter), leaving the absolute 3D world distance measure as the only ambiguity in the solution. It is important to note that certain legitimate motions (like pure translation, and, of course pure rotation (no-parallax case)) will not allow us to do this simultaneous calibration. In this light, are there any advantages of using multiple (in our case two) uncalibrated cameras?

One of the necessary assumptions behind the monocular camera calibration is that it stays constant with respect to its intrinsic parameters. This can be reasonably assumed for multiple cameras as well. In this case two consecutive time frames has 4 views, which is already enough to perform calibration.

Certainly, unconstrained calibration should recover both relative (e.g. relationship with respect to the left and the right cameras) and intrinsic camera parameters. Moreover, the only constraints we can operate on are pairwise epipolar constraints between various combinations of views. This strategy has been implemented first by [Zha95] to recover only the extrinsic parameters, and later improved by [ZLF96] to recover intrinsic parameters as well[3] by solving Kruppa equations, which are non-linear. A slightly different approach of stratification [Fau95] has been taken in [HC96]. Recall that stratification in this context refers to gradual model fitting from simpler to more complex but precise models – projective reconstruction is upgraded to affine reconstruction which is advanced to Euclidean and finalized with metric reconstruction. Thus, the authors of [HC96] first estimate affine calibration (i.e. determine the plane at infinity) and then upgrade it to metric (by explicitly estimating the absolute conic). The advantages of this method are linear estimation procedures, completely general form of the intrinsic camera matrix (which may be constrained, but no apparent benefits of doing this have been found) and more detailed noise analysis (which emphasizes the importance of the affine calibration) – all of which results in a more practical autocalibration algorithm.

A more interesting advantage of using a stereo setup is to estimate intrinsic camera parameters when they are not constant. Specifically, [PvP96] performs Euclidean 3D reconstruction from stereo sequences with variable focal lengths. The authors estimate the change in focal lengths for each camera by analyzing the motion of epipoles, negate this effect by weak pre-calibration and further

[3] These authors used some very reasonable additional assumptions about intrinsic camera matrices (like zero skew and central principle point location) to derive a feasible algorithm.

calibrate the whole setup in the stratified fashion (i.e. affine followed by Euclidean calibration). This method would even work for pure translation (i.e. general motion is not needed), but will not work for a parallel stereo setup because the epipoles are at infinity in this case and their movement due to focal length change cannot be analyzed in principle.

Chapter 5

Non-Rigid Scene and 3D Flow

5.1 Outline

In this chapter the rigidity constraint will be relaxed and methods that recover structure and motion in this less constrained case will be discussed. The discussion of methods progresses according to how tightly the stereo and motion cues are combined. First, 3D range flow recovery from range data (mostly obtained from stereo) will be overviewed. Then paradigms for combination of separate disparity and image motion processing are overviewed. Finally, various endeavors toward joint estimation of stereo and motion will be given a comprehensive overview. Possible treatment of stereo and motion over multiple frames will be touched upon at the end of the chapter.

5.2 Relaxing the rigidity constraint

In the general situation, operating in the unconstrained environment should be done without assumption of scene rigidity. In this case, the scene itself changes from one time instance to another and this change in 3D structure cannot be described by one global egomotion transition. Thus, each scene point, while having 3D spatial coordinates is also characterized by its motion in 3D to capture full temporal dynamics. Such general and hard problems have been mainly considered in the context of scene flow estimation.

In practice, we may get a very good idea of the 3D motion by observing only its 2D projection in a single camera. At one extreme, if the scene is rigid, its structure and motion can be completely inferred (up to a global scale factor). In general, single image motion encodes much information about the structure of the scene and its motion [Pra83], such as direction of egomotion (via focus of expansion/constraction) and relative depth. Meanwhile, the existence of the information does not necessarily mean the possibility of its extraction, and, more importantly, the existence of tractable and otherwise practical extraction algorithms.

Conceptually, the important missing piece in the image flow is the depth change, and that is what can be estimated from multiple image flows. By analogy with spatial stereo, flows from multiple images brought into correspondence make general 3D motion recovery possible, e.g. [VBR^{+}05]. Importantly, the necessity of the aforementioned correspondences essentially is linked to 3D structure – thus, the problem of 3D structure and motion are, once again, tightly connected.

5.3 Stereo as a range sensor – 3D range flow

The idea of stereo being a direct range sensor has been extensively used for rigid motion estimation, as described in the previous chapter and, quite expectedly, also has been employed in general unconstrained 3D motion estimation. As a result, the notion of *Range flow* has emerged as a subarea of its own.

In its original formulation, range flow is merely an extension of the optical flow paradigm to the 3D data case, i.e. the constancy constraint is formulated for 3D spatiotemporal derivatives. The derivation of the range flow constraint is identical to the BCC outlined in an earlier chapter, but the z-component now participates. Specifically, starting with

$$I(x + \delta x, y + \delta y, z + \delta z, t + \delta t) = I(x, y, z, t), \tag{5.1}$$

following with a Taylor series expansion, retaining zeroth and first order terms only and allowing for small time change results in the constraint on the 3D velocity vector $(u, v, w) = \left(\frac{\delta x}{\delta t}, \frac{\delta y}{\delta t}, \frac{\delta z}{\delta t}\right)$

$$I_x u + I_y v + I_z w + I_t = 0, \tag{5.2}$$

where $I_\star$ is the partial derivative with respect to $\star$.

The term "range flow" has been coined by Spies et al. [SJB02], but similar ideas have been developed by many researchers over time. One of the earliest works addressed the problem of estimating the range motion from sparse range data using Bayesian inference and applying a thin plate surface model [Sze88]. Subsequent research estimated range flow in the context of deformable 3D surfaces by enforcing a mesh-like structure [YBBR93]. Another contribution put major effort into enriching the range flow formulation from depth with image flow from intensity information [SJB00]. Finally, [SAS08] were concerned with extending the differential formulation of range flow from the camera grid [SS06] to the case of varying-over-time illumination conditions.

One of the apparent advantages of using stereo is the simultaneous registration with intensity images, where the optical flow (the projection of 3D flow modulo differences with the visual motion field) can be written out and used in the estimation. Note that the range flow constraint describes the constraint between 3 variables using only one equations, which introduces more ambiguities than in the classical 2D optical flow version [SJB00, YBBR93] – normal and full optical flows are counterparted by line, plane and full range flows.

5.4 Combining separate disparity and image motion results

A moving 3D scene results in 2D image projections on cameras separated in space and time. In essence, any subsequent structure and motion reasoning are based on corresponding points across stereo and motion image pairs. Thus, obtaining those correspondences explicitly in the first place gives a good start for the 3D analysis – this straightforward line of thought has already been followed my numerous stereomotion papers.

In one of the the earlier works, [Mut86] computes the collision point[1] from the disparity and image instantaneous velocities in left and right views. That paper states the clear problem of when the camera will pass the point/object based on the binocular velocity input. The solution was derived from the observed velocities of an object in the left and right image. Stereo correspondence was reduced to a trivial notion of detection of an object, as the scene essentially contains a single objects in the experiments. Around the same time [JT86] considered a much more general problem

[1] In the authors notion this is simply the point of intersection of the object trajectory with the $Z = 0$ plane.

of estimation of 3D structure and tracking its 3D motion over time. The authors perform feature matching over spatial and temporal image pairs and realized model-based tracking to work over the whole stereo video sequence. The method was quite elaborate, especially for its time, as it carefully considered constraints on stereo (Panum's fusion area) and motion (visual momentum), was able to deal with occlusions/deocclusion via merge-and-split operations in the feature tracking mechanism, and enforces overall smoothness to the solution via use of numerous local heuristics.

More recently, [ZF92] infer depth from stereo and infer the motion of features (line segments in this case) directly in the 3D coordinate frame. Essentially, the method tracks features and groups them into coherent rigid objects. Significantly, many practical modeling and implementation details were elaborated such as: representation of 3D line segments, a token tracking architecture to allow multiple matching with uncertainty estimation, efficient segment matching, an EKF framework for enforcing a rigid-motion model in segment grouping into multiple objects, and extensive experiments with real data. A later approach also solves the stereo motion problem via token tracking [YO97]. In this work, the emphasis is on generation of virtually all possible stereo matching tokens and refining them with motion matching information, while introducing an efficient token discarding mechanism to avoid the exponential growth of tokens. General non-rigid motion is considered[2] and token tracking itself is performed with a Kalman Filter.

Even more recent efforts have emphasized dense modeling over sparse token-based approaches. One such deals with this problem by first performing dense stereo to have a correspondence between the left and the right points [ZK01]. Then, the objective becomes recovery of consistent optical flow maps by formulating a joint Least-Squares error criterion. Once the left and the right flow is jointly computed, 3D flow inference becomes trivial as discussed later in Sec. 5.5. Furthermore, the general idea of simultaneous estimation of optical flows from multiple cameras has been extended by [VBR$^+$05] – in that case, the authors use the spatial correspondence information (obtained from space carving [KS00] rather than standard multi-view stereo) to compute 3D flow from the differential BCC constraints over multiple video sequences.

Range flow to model skinning[3] in 3D has been utilized in a quite straightforward way using the stereo camera [NS02] – stereo is performed at each time instance to yield a 3D range map and the 3D motion is determined from the consecutive range maps; temporal correspondences are simply found by matching time-consecutive left images using the same stereo search procedure[4]. A more elaborate way to compute binocular 3D flow and its multi-view extension has been demonstrated by [PKFH03] and [PKF05], respectively. [LS08] take their own cut at the problem by fusing the stereo and left and right flow estimations performed in the CTF fashion with uncertainty modeling. There, the authors rely on a variational formulation and utilize the photometric matching constraint as a projection on the 3D surface under recovery: initially, stereo estimation is performed; subsequently, the 3D flow vector field is found using the same variational principle to register input images captured at different times. As another interesting alternative, [KN03] proposes a scheme to estimate 3D structure and 3D motion by combining stereo disparity, 2D image motion and even shading flow in a large pixel-based Kalman Filter. Again, all cues are computed separately to make the framework easier and more tractable.

Finally, special attention can be devoted to typical works in car aid navigation that estimates structure and 3D scene motion from stereo cameras. Specifically, we are interested not in mere navigation, which is almost exclusively done via SLAM, but in *moving* obstacle detection, e.g. detection

[2] Actually, it is assumed that the objects move independently of each other, and have a small enough size to be represented by a single token. In turn, motion is assumed to be small and smooth enough such that the first derivatives of the 3D position can be considered as constant over a sampling period.

[3] This particular application tries to simulate skin deformation during limb motion.

[4] Disparities are 2D vectors in this case.

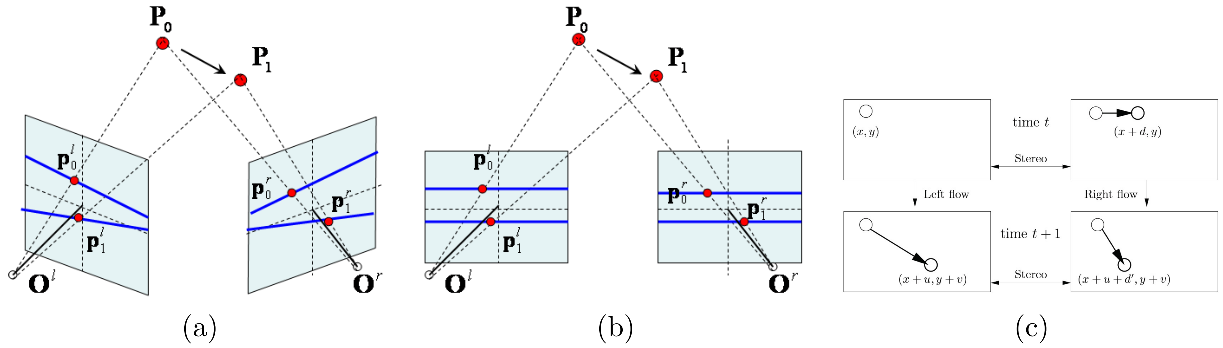


Figure 5.1: Stereomotion in the Stereo Rig – 4-frame imaging constraint. Point **P** undergoes a 3D motion from time instance 0 to 1 and results in corresponding projections on the the left and the right images. Blue lines are corresponding epipolar lines of the stereo system. From left to right: (a) General configuration, (b) Fronto-parallel configuration (c) Four frame fronto-parallel configuration (Adapted from [HD07]).

of pedestrians and cars. One of the early successful attempts was demonstrated in [DMC90, DM92]. More recent advances are fairly represented by [FRBG05, WRV+08]. The latter methods are based on separate disparity and flow computations. In particular, the stereo disparities are separately estimated and 3D motion is derived from left and right image flow constraints – as will be derived later in Sec. 5.5. Furthermore, image points with disparities can first be converted to 3D coordinates as in [FRBG05].

5.5 Toward joint estimation of stereo and motion

Since the investigation of the stereo problem starts with having 2 images (taken from different space locations) and the investigation of optical flow needs 2 images as well (taken at different time instances), it is reasonable to view the stereomotion problem considering 2 stereo frames consecutive in time, i.e. consider 4-frame constraints.

The investigation of the 4-frame stereomotion problem has been described in the previous section and here we continue considering the general motion and structure computations while increasing the "tightness" of the coupling between the stereo and motion cues.

During the following discussion refer to Fig. 5.1a. A point **P** undergoes a 3D displacement **V** and is projected into $\mathbf{p}_0^l$, $\mathbf{p}_0^r$ at time instance 0 and $\mathbf{p}_1^l$, $\mathbf{p}_1^r$ at the consecutive time instance 1 on to the left and the right frames, respectively. Stereo epipolar relationships exist between the pair $\mathbf{p}_0^l$, $\mathbf{p}_0^r$ and $\mathbf{p}_1^l$, $\mathbf{p}_1^r$. Temporal matching between pairs $\mathbf{p}_0^l$, $\mathbf{p}_1^l$ and $\mathbf{p}_0^r$, $\mathbf{p}_1^r$ as well as cross-matchings $\mathbf{p}_0^l$, $\mathbf{p}_1^r$ and $\mathbf{p}_0^r$, $\mathbf{p}_1^l$ are generally unconstrained and are subject to 2D search. Certainly, if the motion is rigid, the epipolar constraint can be applied there too.

Parallel stereo setup

The parallel or otherwise rectified stereo setup plays an extensive role in computational stereo because epipolar lines are aligned with the horizontal scanlines. This has the great advantage of simplifying the underlying algorithms for two main reasons: (i) disparity search is merely search along the x-axis, (ii) computation and memory management are significantly improved upon, because images are conventionally stored in a row-major order in memory.

Thus, we implicitly consider the situation of Fig. 5.1b for the parallel setup. Moreover, as we will see, almost every paper that relies on binocular stereo assumes it is of parallel configuration. A particular convenience of the fronto-parallel setup is the fact that vertical components of left and right image motions must be the same, because the epipolar lines which $\mathbf{p}_i^l$ and $\mathbf{p}_i^r$ reside on are identical scanlines in the left and the right images.

A more clear picture, especially for the purpose of matching, can be obtained by explicitly considering 4 frames and their relationship as in Fig. 5.1c. There, a point in the left image (x, y) has disparity d and undergoes an image+disparity motion (u, v, d') (here, d is disparity, u, v and d' are the x-, y- and disparity flow components, respectively). Arrows in the diagram connect the image pairs for which the photometric matching constraints are formulated. Cross combinations also possible, but are hardly ever used (with the exception of [SS98b]).

5.5.1 Pairwise matching constraints

The situation depicted in Fig. 5.1 can be handled in a number of ways and the most computationally simple thing is to use pairwise matching cost functions.

One of the earliest endeavors is [LS93], where the authors develop a system that performs Dynamic Programming-based stereo matching along epipolar lines and coarse-to-fine matching along left and right temporal streams. The relationship between stereo and motion vectors is described, but it is unclear how it is actually enforced. Moreover, the authors were mostly interested in the 3DTV video coding application, which is more specific and in many respects less challenging than 3D structure (and motion) recovery. Another naive combination appeared in [HP94] and consisted of a simple iterative pairwise matching performed in a circular fashion[5] until convergence. Only uniqueness and smoothness constraints were enforced.

Sudhir et al. took a more principled approach by integrating the pairwise matching error constraints into a global minimization framework together with a regularization term for smoother results [SBBB95]. Estimation is performed in interleaved stereo and motion fashion and a special treatment for discontinuities is taken by explicitly estimating the latter as a binary map. Subsequent work estimates the disparity and flow in both images taking the previous disparity field as known [PAT96]. Taking into account the latter fact, the authors formulate a global error function that consists of simple SSD-based error terms for stereo and motion pairs plus a smoothing prior. The minimization problem is solved iteratively via diffusion. Other authors also considered using explicit matching terms for stereo and motion in a single minimization function [MS97]. In their case, only the right-based image motion is utilized (3 frames out of 4) and structure estimation is done with the help of a mesh model. Even thought this particular endeavor may not be the best for actual structure estimation, it performed quite well for the authors' application of stereo video sequence coding.

Strecha and van Gool [Sv02] propose to use more sophisticated 3-frame combinations out of 4 possible for stereomotion correspondence encapsulated into a joint minimization criteria to be optimized using a PDE solver. However, they explicitly make use of the epipolar constraint during consideration of temporal matching, which, strictly speaking, is only true when motion is rigid. However, rigid motion parameters were never computed. A more recent contribution also combines the stereo and motion pairwise matching constraints plus regularization into a single PDE formulation [HD07]. Late linearization to the optical flow constraint is performed and a more sophisticated solver (in comparison to [Sv02]) is used.

Gong [Gon06] explicitly performed motion computation in disparity space, (x, y, d), and also

[5] Stereo pair at time 0, right image motion pair, stereo pair at time 1, left image pair, etc.

considered the pairwise constraint, but his formulation is solved using the dynamic programming algorithm (performed on horizontal and vertical lines). Here, it is worth noting that trinocular stereo has been used in the experiments for more robust matching and occlusion handling. Yet another work addressed the same problem by introducing an edge-preserving matching and smoothness term for better results [MKS05, MS06]. A further addition was to incorporate a joint estimation constraint to bring all 4 images in a single match together – however, it was used only for verification purposes to get rid of inconsistent matches.

5.5.2 Simultaneous four-image matching

In a general stereomotion scenario each point is to be characterized by its depth (or disparity) and 3D motion (be it in disparity space or world coordinates). Thus, during the general stereo matching processing, each point should be characterized by a 4D vector (e.g. $\begin{bmatrix} d & u & v & d' \end{bmatrix}$, where d is the disparity and the remaining terms are the components of the 3D disparity flow). In turn, direct solution for this vector will involve search in 4D space, and must involve all four images simultaneously. In essence, the matching procedure would reduce to the construction of this 4D disparity space, i.e. finding an error at every point for all possible (discrete) values of $\begin{bmatrix} d & u & v & d' \end{bmatrix}$ and then choosing the vector that results in the minimum error. In turn, the overall error formulation could be purely local, or correspond to some global energy function.

Obviously, the search in 4D space is a more challenging problem due to its computational demands as well as other complications, like a choice of good smoothness prior. The idea of using all 4 images simultaneous to explicitly search in 4D space has been discussed in [ZK01]. Meanwhile, it can be argued that only [IM06] suggests a fully cooperative stereomotion image matching formulation that simultaneously consider four images. However, this particular method has not been widely used or improved upon due to its inherently higher computation and memory demands. Furthermore, results only on very small images have been demonstrated to date.

5.5.3 Stereomotion as a 6D concept

Continuing the discussion started in the previous section, we can say that a scene is an entity in 3D space and at every point motion can be characterized by another 3D vector. Thus, a general unconstrained scene in space and time can be fully represented as a 6D space construct. For example, [FRBG05] uses the term "6D-vision" by merely fusing the results of stereo matching and optical flow computation in a single estimation framework. However, the 6D concept has been fully exploited only in [VBSK00]. The authors essentially extended the space carving idea [KS00] to spacetime – carving was performed in 6D parameter space directly using the colour constancy constraint. Recall that the space carving algorithm essentially tries to recover the photo hull – a 3D shape that subsumes all possible reconstructions that are photo-consistent with the multiple input images – by an algorithm operating in the 3D space of voxels: Operation proceeds by sequentially "cutting" out voxels that result in colour-inconsistent projections onto the input images. An interesting twist to the problem has been proposed by [GM04], where the authors perform a time-consistent space carving. A 4D spacetime (x, y, z, t) is considered where a slice along the time dimension results in a 3D isosurface – a 3D scene at a particular instance of time. The 4D smooth isosurface is inferred from frame-wise space carving results using a level set method.

Finally, it is important to note that this class of method is practically useful only when we have many views of a scene – preferably, we must be able to view the whole object at once to reason about its visual hull. Thus, this class of approaches is not of primary interest in the current context, as we deal predominantly with binocular stereo setups, possibly observing both space-bounded objects and arbitrarily large unbounded scenes.

5.6 Multiple frames – filtering in the case of non-rigid motion

Section 4.7.3 goes into depth on the filtering framework for rigid motion scenes. The same framework, especially variations of the Kalman Filter, have been extensively used in the general non-rigid motion formulation. The major difference now is that each feature/scene point undergoes its own 3D motion and it is explicitly modeled by the filter. Thus, the corresponding state vector now becomes a $3N$-dimensional vector just to capture the motion parameters, where N is the number of points, as before, plus an additional N dimensions to compute the structure. Since the dimensionality is extremely high, non-parametric filters stay out of question, and only the Kalman Filter [YO97, FRBG05] and Extended Kalman Filter [ZF92, KN03] have been used successfully. Here, EKF has been more popular due to its ability to deal with non-linear state and measurement models.

Chapter 6

Spatiotemporal Stereo, Temporally-Coherent Stereo, Spacetime Stereo

6.1 Outline

This chapter will discuss in depth the paradigm of spatiotemporal stereo (also known as spacetime stereo and temporally-coherent stereo) – a 3D structure estimation technique over time. We start with discussion of the obvious idea of utilizing optical flow to assist the disparity search, continue with the idea of exploiting continuity in space and time during matching and the notion of spacetime matching as the end-result. We also discuss ways for binocular spacetime matching and analysis to operate on higher level features than raw pixel intensities.

6.2 3D structure estimation over time

Calibrated stereo camera are becoming a more common type of sensor. Importantly, in many cases we are predominantly interested in structure estimation with respect to the camera only, and, since we do it over time, it is naturally to ask our estimations to be temporally consistent. For such cases, we largely ignore explicit (3D) motion recovery (although it might be possible) and concentrate on structure only. Given this simpler, but still a very practically important objective, the main goal becomes the estimation of structure at each instance of time, subject to reduced noise and

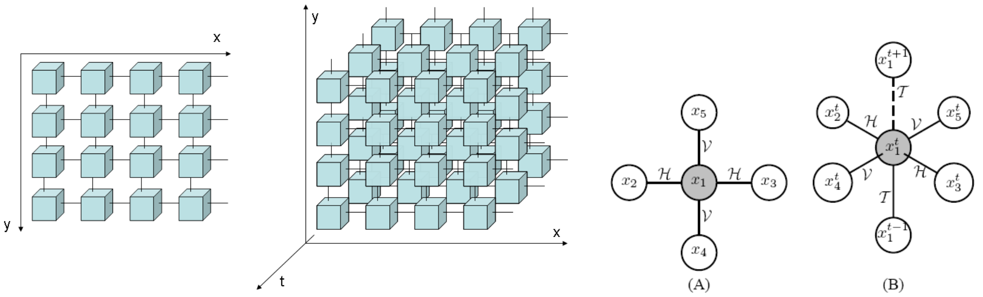


Figure 6.1: Spatial and spatiotemporal MRF grids. Whole grids (on the left) as well as their basic generating elements (on the right, adapted from [WIM05]) are shown.

temporal consistency.

6.3 Optical flow to "seed" disparity search

The natural availability of stereo frames over time has been noticed long ago and various efforts have tried to exploit it to help otherwise static stereo matching. The earliest endeavors, e.g. [Nev76], perceived the presence of time consecutive stereo data mainly as a way of simplifying the match by reducing the feature search range; this will automatically result in speed improvement. Here, the authors tried to avoid the problem of motion estimation (which, being less constrained, is even harder) in stereo matching reinforcement. Instead they considered the motion as predetermined, or small enough so that it becomes trivial for sparse feature methods. Essentially, the same approach was revisited multiple times over the years and advances in stereo and motion processing yielded more compelling results. Along these lines, two different auxiliary objectives were tried, which introduced an interesting tradeoff, as follows.

Crossley et al. introduced temporal coherence to stereo matching to significantly reduce the search range for an edge-based correlation system [CLTS97]. At the same time, errors in flow estimation yielded incorrect predictions for the stereo matcher. Thus, the accuracy of the algorithm generally was reduced (despite making it somewhat faster), as correlation was restricted to a relatively small portion of the entire disparity space. The introduction of various stochastic searches to combat the errors in flow resulted in some accuracy improvement, but did not advance on this matter considerably. On the other hand, a much later work [Gon06] introduced temporal consistency specifically to improve matching results. In particular, the authors estimate left and right image flows and use them only to improve the correlation scores of the corresponding stereo matching hypotheses – the search is still performed across the whole disparity space. Such a simple scenario still allows finding the correct disparity match even if the flow estimation is wrong. Certainly, it results in much more significant computational load, but the authors still manage to demonstrate real-time performance, though on small images[1].

Finally, it worth noting that propagation of disparity estimates from one frame to another can be done not only densely via optical flow, but also via tracking more complex entities, like edges and lines [Sha02].

6.4 Space and time continuity in stereo matching

The idea of reinforcing spatial matching using temporal correspondences may be viewed slightly differently. Recall from Chapter 2 that stereo matching is an inherently ambiguous problem and various other cues and constraints must be used to make it well-behaved. The cue of paramount importance, especially in dense matching, is the continuity constraint, with the MRF 2D grid being one of the most popular realizations. Thus, a straightforward extension to spatial continuity is an upgrade to temporal continuity, which can be organized as a MRF 3D grid. For example, [LALS04] simply extend the MRF stereo algorithm with coarse-to-fine dynamic programming to consider the temporal pixel links between frames at time t, $t-1$ and $t+1$: The algorithm remained essentially unchanged, but temporal passes were added to the horizontal and vertical spatial passes. The typical MRF 2D grid is now replaced with 3D grid as shown in Fig. 6.1. Others considered the problem in a more careful fashion by appropriately modeling occlusions, learning all model parameters and using Belief Propagation instead of Dynamic Programming [WIM05]. Furthermore,

[1] The fact that stereo matching and optical flow estimation are performed in a disjoint fashion is beneficial from the speed point of view – estimations may be done in parallel using, for example a graphics card [Gon06].

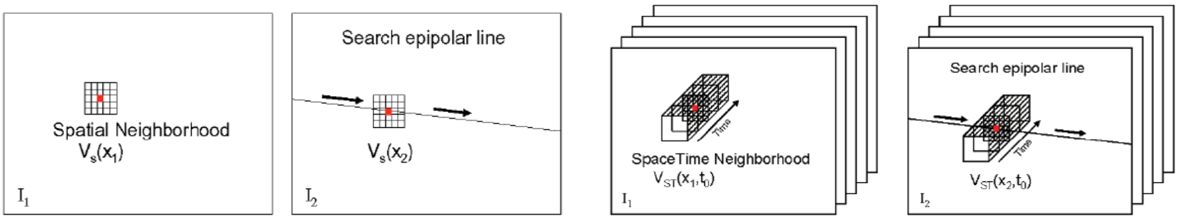


Figure 6.2: Spatial vs. spatiotemporal aggregation. Adapted from [DRR05].

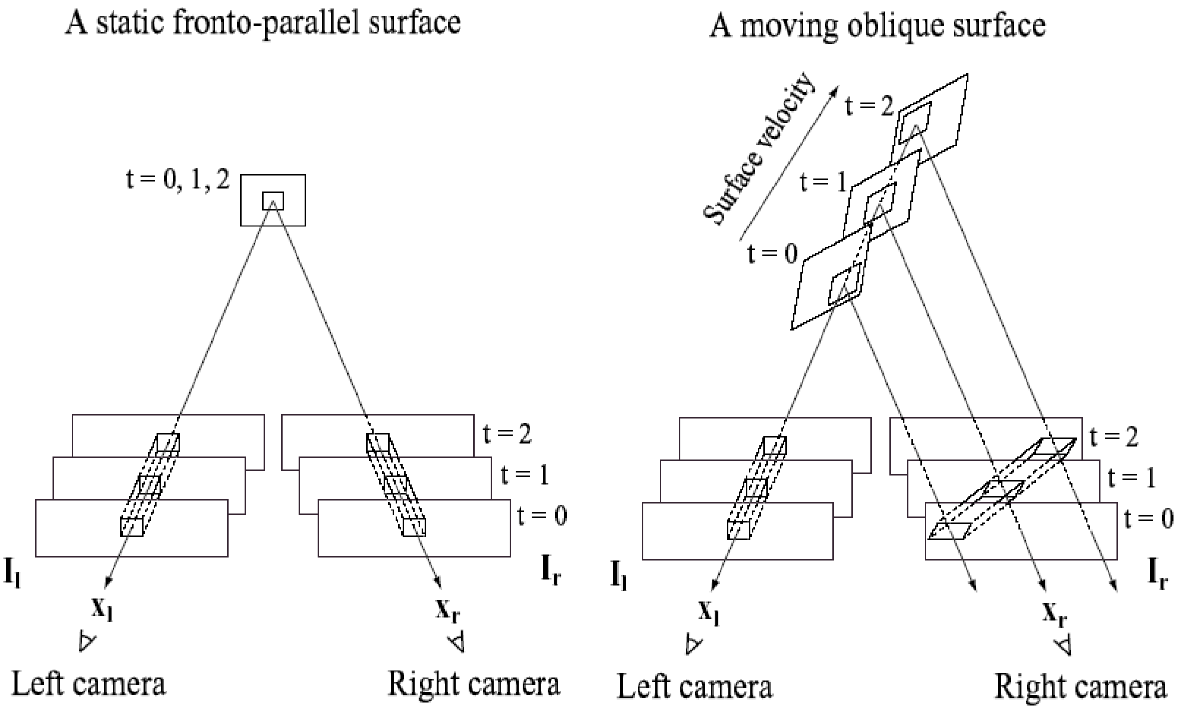


Figure 6.3: Compensation for depth motion and slants in the spatiotemporal aggregation for rectified binocular stereo. Adapted from [ZCS03].

the authors acknowledged the fact that simple temporal links are not correct when motion is present and introduce a simple motion detector to break invalid links. Finally, [LMPF07] treats the modeling of temporal MRF links appropriately by forming them using pre-computed optical flow and designing an elaborate belief propagation algorithm specifically for temporal MRFs[2].

6.5 Spacetime stereo

The next step in generalizing continuity in space to continuity in space and time is to change the perception of video as a collection of time-consecutive 2D images into that of a single continuous spacetime 3D volume – and that is what the “spacetime” term predominantly denotes.

[2] The authors design specific propagation rules and introduce the ability to break temporal links in case of inconsistent messages from spatial and temporal streams. This refinement tends to be the most beneficial in places such as 3D boundaries where optical flow experiences the most difficulties.

Despite the long history of using the spacetime volume, especially in the area of motion processing [GK95], this paradigm has been introduced to spatiotemporal stereo fairly recently. For example, [NA02] extended the local aggregation in correlation-based matching into the temporal domain and used this idea to create temporally varying meshes from multiple cameras. The fact that motion results in different spatiotemporal profiles in the left and the right views was not addressed.

Meanwhile, a clear and definitive statement with respect to spatiotemporal matching has been given independently by two research groups [ZCS03, DRR05]. The authors proposed to treat stereo over time as a matching problem on 3D spacetime volumes, where matching itself was performed using 3D aggregation, as shown in Fig. 6.2. The basic idea behind using aggregation in time is that we can reduce spatial aggregation and thus greatly ameliorate the problems near 3D discontinuities that are paramount for aggregation-based stereo [SS02, BBH03].

In reality, since point intensity per se does not drastically change over time, a temporal texture cue must be created in order to allow for substantial advantages, if the goal of reducing spatial aggregation to single pixel support is to be realized. In essence, the methods [ZCS03, DRR05] and subsequent improvement [ZSCS04] were finalized as active range techniques that use stereo cameras and an active projector that creates random or pseudo-random texture on the scene, which also explains the exceptionally good results they offer. Importantly, those methods are only interested in stereo processing – actual motion is very undesirable, as temporal aggregation cannot be done along straight lines anymore. Thus, motion is either completely ignored ([DRR05]), or attempts at compensation are made ([ZCS03, ZSCS04]) by resolving the difference of aggregation window orientation between the left and the right views, as shown in Fig. 6.3, using a linear differential model. The downside of this approach is that spatial aggregation cannot be arbitrarily small (i.e. 1×1 spatial support is not possible anymore), which does no longer offers an automatic treatment of discontinuities. Furthermore, presence of large, non-trivial motions is hard to be successfully compensated for [ZCS03]. Moreover, the aggregation window, being straight in time for the reference view, is not aligned with the motion trajectory of the central pixel and thus will result in less accurate matching.

6.6 Spacetime stereo – beyond intensities

So far, the spacetime paradigm has been used mainly to summon extra data from the temporal dimension to make matching more discriminative. Here, motion has been either ignored, or minimal attempts for compensation have been taken.

Meanwhile, spacetime volumes can tell a great deal about motion. Toward that end, [SW09b] recently proposed to use a novel descriptor called “stequel” to perform spatiotemporal stereo. The stequel descriptor is based on 3D oriented energy analysis within the spacetime volume. In essence, as visualized in Fig. 6.4, stequels simultaneously capture the spatial appearance and temporal dynamics of the scene. Furthermore, spatiotemporal epipolar constraint and corresponding stereo matching principles have been derived that operate directly on stequels. A particular benefit of this approach is that motion does not need to be extracted and the primitive works quite well for non-trivial motions. Further advantage lies in fact that 3D motion is readily recoverable from stequels in correspondence, which no other spacetime stereo methods currently support [ZCS03, DRR05].

More intriguing advantages of spacetime matching may arise from 3D orientation energy distributions themselves (bypassing the stequel creation stage), as they are able to describe the temporal dynamics not constrained to simple single motion at a point scenarios. Specifically, scintillations and transparent moving surfaces manifest themselves as structures with multiple orientations, which

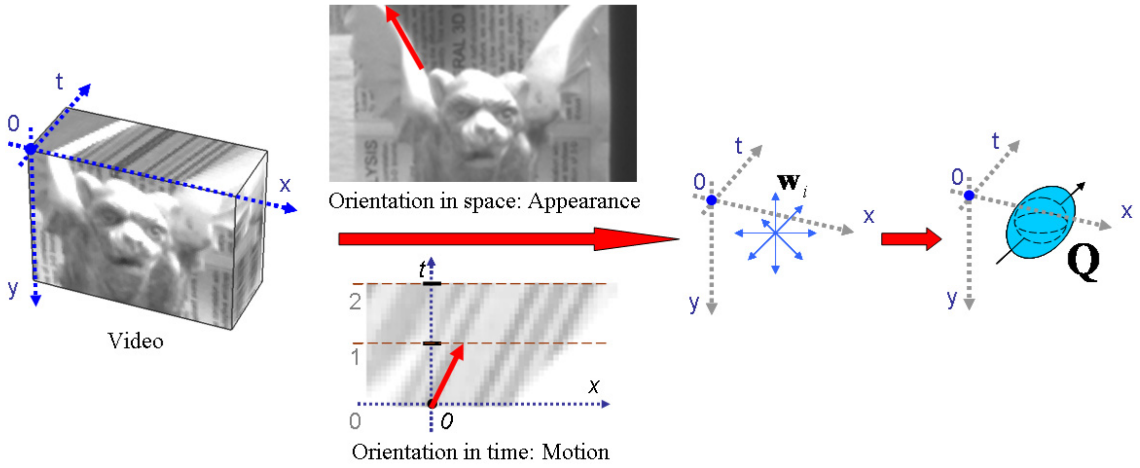


Figure 6.4: Extracting stequels from the spatiotemporal volume. Orientations in 3D represent both spatial appearance and temporal dynamics. At each point, energies from the set of orientations that equally span x-y-t space are collected. This collection is latter collapsed into a quadric, Q, termed the stequel for subsequent matching. Adapted from [SW09b].

means that stereomotion supplies principal cues for processing scenes with multiple layers and specular reflections, because the latter can be described by an equivalent multi-layer arrangement.

Chapter 7

Performance of Stereomotion Techniques

7.1 Outline

Once the variety of methods for stereomotion processing has been discussed, it is necessary to highlight their performance in order to understand the state-of-the-art. Coverage is done in the order of the three previous chapters – rigid structure and motion, nonrigid structure and motion spatiotemporal stereo. Furthermore, within each section, there is an extended discussion of how the performance may be assessed and what datasets must be used.

7.2 Rigid scenes

Chapter 4 described the problem of rigid structure and motion recovery and discussed various solutions. This section will give a sense of performance achieved in the rigid structure and motion recovery problem. To do so in an informative fashion, we need to clarify how performance can be assessed, including consideration of datasets.

7.2.1 Performance assessment

The methods of concern recover both rigid structure and motion. Nevertheless, one or the other modality may be more important in a particular application, which may influence the development of the algorithm and, correspondingly, the evaluation methodology.

Rigid motion

Rigid motion describes the global motion of the camera through the scene (or vice versa), which means that it can be described by a relatively small number of parameters. In the overwhelming majority of situations it is 3 for rotation and 3 for translation, though some other quantities like acceleration [BE95] or precession [YC90] may be required as well. Furthermore, it is easy to create a scene with controlled rotation and translation using commercially-available motorized stages, or a robotic arm.

In this light, the evaluation of rigid motion parameters becomes the trivial task of simple comparison of scalar values – possibly calculating the 1D distributions with mean and variance of error for each parameter. Specifically, rigid motion parameters are conveniently described by $x-$, $y-$, $z-$ translations and roll, pitch, yaw rotations.

Structure

Since most relevant methods are capable of solving structure and motion directly in 3D, it is possible to compute 3D structure (in terms of a point cloud, let's say) and directly compare it to ground truth 3D structure. Thus approach is used in many methods, especially when the scene data is not particularly hard (i.e. obstacle avoidance scenarios).

On the other hand, a potentially more meaningful way to evaluate structure is in terms of a range map associated with the reference image view. Range maps have intrinsic data continuity enforced by a pixel grid. Thus, comparison to the ground truth easily boils down to pointwise scalar value comparisons. This technique is extensively used in stereo evaluation, where disparity is used instead of range [SS02, BBH03, Mid08]. The aggregate error statistic is usually RMS error over all pixels, or percentage of pixels where absolute error is greater than a predefined threshold. The latter comparison turns out to be more informative, as it is not usually important how big the error is if pixel is wrong and the choice of the threshold allows one to see various aspects of performance, e.g. primary pixel-wise matching (for threshold around 1 disparity value) vs. subpixel interpolation (for threshold around 0.5 disparity value).

7.2.2 Datasets

Synthetic

One way to assess the performance is to have full control over testing scenarios and generate the data synthetically. This straightforward approach has a great advantage of constructing scenes of any complexity in terms of geometry and motion as well as possession of the ground truth, which allows for the quantitative assessment of the algorithms and meaningful comparison between competitors. Noise is usually added to simulate more realistic conditions and test the performance of the algorithms in the presence of such corruption. Meanwhile, there is no guarantee that real noise will match synthetic noise, which is very often simply modeled as Gaussian. Furthermore, there are other complicated effects that usually are not modeled in synthetic datasets, e.g. non-Lambertian scenes, sub-surface scattering, specularities, etc. Meanwhile, these issues are directly connected to the power of the modeling tools – for example, sophisticated graphics models and algorithms can be used to render very realistic non-Lambertian effects and transparencies. In any case, though some conclusions made regarding performance on synthetic data can be extrapolated to real scenarios, there is no fundamental guarantee of the correctness on such implications.

Real data

Currently, given a pair of cameras and computer it is a trivial matter to capture meaningful stereo sequences of reasonable quality. Thus, results on real data are easy to obtain and are a must in demonstrating the usefulness of stereo motion algorithm.

The obvious drawback of real data is difficulty in knowing the ground truth, but there are various ways to deal with this challenge

- Disregard the ground truth. It may be enough just to demonstrate the performance of the algorithm in situations when the field is just growing and any solutions are welcome, or a result is so compelling and visually pleasing that it documents value by itself, even though it cannot be directly assessed numerically. However, this minimalist point of view becomes too limited as the field matures and requires rigorous assessment of algorithm performance either for the purpose of comparison with alternatives, or for speed-accuracy tradeoff analysis.

- In absence of the ground truth it is possible to reason about stability of the solution, e.g. variance of recovered disparity values, entropy, etc. Furthermore, this auxiliary analysis is useful even when ground truth is available.

- A more meaningful indicator is to consider a view rendering task – for example, having 3D structure and motion information it is possible to render other views from the reference view, and compare the results with the true views. Measures such as RMS intensity and colour errors, or SNR may be reasonable. Moreover, this strategy is optimal when the major application of an algorithm is coding, or new view synthesis, while geometry per se is not of immediate interest.

- Since people have a choice on what dataset to capture, they have some control over complexity of the 3D structure environment and, more easily, the rigid motion the camera undergoes. In case of 3D scenes, the shape of simple objects (like cubes, or planar walls) and their distances from the camera may be known exactly.

- Finally, a good alternative is to use some superior technology to acquire the 3D measurements, as is currently done in the stereo field [SS02, Mid08] . The situation is simpler for rigid motion, which can be programmed precisely by positioning cameras on a robotic arm.

7.2.3 Level of performance

Early works in rigid stereomotion (including some recent ones as well) were more like a proof of concept and experimental evaluation was limited to simulations on synthetic data; [AW87, BK83, HB85, ZHZ88, BE95, LK90, YC90, Der06]. Gaussian noise was added in most situations and algorithms were tested with various levels of noise to investigate conditioning of the Least-Squares ego-motion estimation, or convergence of the predictive filtering frameworks. Interestingly, a few works did not have even synthetic simulations [HH91, TSJ92].

With the availability of cheaper electronic cameras and growth of their popularity, real scenes started to be introduced to the experimental evaluation replacing synthetic ones [GST89, NDF90, WCR92, Koc95, GT95, PvP96, HC96, HRD$^+$99, MSS99, HC00, DH00, MB00, DD02, ZN04, AKK05]. Still, many contributions, including the most recent ones, use a mix of synthetic data to quantitatively assess the performance and real data to show the algorithm is applicable to real situations. Additionally, many methods demonstrate the result of warping of the reference view to the other matching views as a consistency check to signal that meaningful structure and motion parameter values were recovered.

Interestingly, in some cases real sequences with ground truth for ego-motion parameters were introduced rather early, and more recent methods use GPS as a gold standard [WD86, KA87, AH90, HO93, SZS94, GT95, WD96, SS98b, OMSM00, NNB04, CMR07].

The availability of ground truth and quantitative evaluation for some methods allows us to get a general sense of performance, especially for the egomotion parameters, rotation and translation. A few important trends may be outlined:

- In general, improvement over time can be noticed, even though each method has been tested on its own dataset. Processing over multiple frames via predictive filtering helps significantly, sometimes showing even one order of magnitude error decrease.

- Direct methods, due to their close integration of cues and avoidance of complex intermediate computations show better results. For example, one of the first methods [SZS94] with error reported at 15%, improved to less than 5% in the later work [SS98b, MSS99].

- Doing computation directly in disparity space and avoiding 3D reconstruction is able to give huge improvement [DD02]. For example, the relative translational error drops from 0.35% to 0.025% once essentially the same estimation is performed in disparity space instead of metric 3D coordinates.

- Very good results have been demonstrated in robot navigation across various terrains, from desert and forrest to cluttered urban scenes. Work almost a decade-old [OMSM00] was able to estimate rover trajectory across 210 stereopairs with error of only 1.2%, while more recent work [CMR07] shows only 0.004% and 0.03% drifts in translation and rotation, respectively, over a 360 meter long sequence. With respect to absolute localization of moving robots, significant improvement has been achieved – a 1% to 5% error was dropped to 0.1% by [ZOS+07].

In case of structure, quantitative description is extremely poor and is usually contrived to very simple scenes where object location is known – for example [WD86] reports the recovered distance of 38 in vs. 45 in, which is quite a large error.

In conclusion, to date, virtually every paper used its own datasets with attempts simply to demonstrate the viability of the approach. No explicit comparison between rival methods has been done. At the same time, current state-of-the-art solutions have been brought from research labs into the industrial world and demonstrate success (e.g. [NNB04]), especially in the field of robot navigation.

7.3 Non-rigid scenes

The outcome of a non-rigid stereomotion algorithm is the 3D structure (usually in the form of range maps) and a pointwise 3D flow. Since evaluation of structure estimates has been discussed in the previous section, we concentrate on evaluation of 3D flow below.

7.3.1 Performance assessment of 3D flow

The end result of a 3D range flow estimation procedure would be the 3D vector at every point (where an attempt is made to estimate flow). Indeed, it can be perceived as an extension of the optical flow recovery result in 2D, which means that ideas of performance assessment in that area can be reused here. The question of proper assessment boils down to how to compare the recovered to the ground truth 3D vector and get an aggregate numerical indicator over the whole scene.

A few measures have been appealed to in the past.

- Norm of the difference, E_n, between vectors $\mathbf{v}$ and $\mathbf{v}_{gt}$ of recovered and ground truth flow, respectively:

  $$E_n = \|\mathbf{v} - \mathbf{v_{gt}}\|_2 \tag{7.1}$$

  The measure and its aggregate statistics are trivial to compute and have been used in optical flow comparison before [ON94]. In this formulation, large motions will always have greater contribution and the direction of the flow cannot be accessed. Thus, various weightings can be used to calculate aggregate statistics, such as inverse proportion to the ground truth vector length to equate the contribution of small and large flow vectors.

- Angular error, E_a, between vectors calculated as a normalized dot product:

  $$E_a = arccos\left|\frac{\mathbf{v}^\top \mathbf{v_{gt}}}{\|\mathbf{v}\|_2\|\mathbf{v_{gt}}\|_2}\right|. \tag{7.2}$$

This measure is the most widespread as it tells the error in the estimated motion orientation – arguably a more important and descriptive quantity than vector difference, considering that the motion is usually small, which means that the magnitude of the vectors is small too. The aggregate statistic is simply the average of angular errors. This measure has been popularized by the seminal paper [BFB94].

- Magnitude error, E_m, between vectors, which may be use as a complement to the angular error formulation:

$$E_m = |\|\mathbf{v}\|_2 - \|\mathbf{v_{GT}}\|_2|. \tag{7.3}$$

7.3.2 Dataset

Synthetic datasets are definitely possible to construct for range flow estimation problems. Along these lines, a completely synthetic datset Yosemite [BFB94] has been used as a major indicator of performance for more than a decade.

Real scenes are valuable and abundant, but obtaining the ground truth for them is more challenging than in the case of stereo or rigid motion. Leaving aside the application-specific evaluation methods like view interpolation, there are very few ways to obtain the ground truth known so far:

- Move through a known scene in a rigid fashion, but run algorithms without the rigidity assumption. It is much easier to construct ground truth flow vectors for known rigid scenes, as discussed above. A more contrived scenario is to evaluate motion methods on stereo datasets, which has been extensively used before to indicate some level of performance [HD07, BSL+07]. An obvious drawback of these "tricks" is that many interesting scenarios (like deformable objects) cannot be handled, while they are arguably the most interesting ones that really need general flow computation methods.

- Use some superior, but more expensive or not everywhere applicable technique to get good estimates and declare them as the ground truth. While a good choice of alternatives in the form of structured light and range scanners exist for stereo [SS02], obtaining optical ground truth even for the 2D image flow is much more challenging. Nevertheless, a method using a special dye that is visible in the ultraviolet spectrum only to colour the scene in order to make motion recovery close-to-trivial has been proposed [BM04]. Still, no method has yet been developed to create good quality 3D range motion ground truth – although it seems plausible to develop one by fusing the ideas from stereo and motion evaluation technologies.

7.3.3 Level of performance

Since recovery of general 3D flow is less constrained and arguably harder than the rigid motion problem, meaningful solutions appeared later, when real data was relatively easy to capture. This observation largely explains why every method covered in Chapter 5 (except possibly [Sze88]) shows example results on some real scene (although every paper uses different scenes). At the same time, the overwhelming majority do not have ground truth, and thus present no quantitative analysis.

To overcome the latter limitations, synthetic sequences were considered in a number of works to quantify the proposed improvements over previous formulations [YO97, SJB00, SJB02, HD07, LS08, SAS08, WRV+08]. Importantly, a few methods make use of ground truth constructed for simple scenes and mostly for rigid motions [Mut86, ZF92, YO97, DMC90, SJB00, SJB02]. In most recent works, low angular errors on the order of only a few degrees are reported.

Again, evaluation of structure estimates is given secondary importance and no clear conclusions on this matter can be reached. Indeed, it should not come as a surprise, because many methods interested in 3D range flow treat the stereo and motion problem in a disjoint fashion and merely use the 3D estimates from pure spatial stereo matching, performance of which is not affected by the presence of the motion cue.

7.4 Spatiotemporal stereo

7.4.1 Performance assessment

All methods discussed in Chapter 6 put the major emphasis on 3D structure, and more specifically stereo. Thus, evaluation of these methods should be particularly clear, as it boils down to conventional stereo evaluation. In turn, current evaluation technology for spatial stereo methods is in quite good shape, because easy and cheap methods for ground truth computations are available, as discussed in Chapter 2.

Indeed, all methods show processing results on various real images. Very early work [Nev76] deals with simple objects, like coffee cups, and demonstrates the meaningfulness of the recovered 3D structure by fitting circles (as a part of a cylinder). Methods designed for new view synthesis applications perform quantitative evaluation with respect to novel view generation using recovered disparity maps [Sha02, LMPF07]. Finally, [WIM05, DRR05, SW09b] operate on scenes with ground truth, which is either hand-constructed for a very simple moving box scene [WIM05], or range values are obtained by a superior method, as in [DRR05], which demonstrate a useful tradeoff between local aggregation in space and time.

Considering that spatiotemporal stereo in spirit is meant to outperform purely spatial stereo, it is important to note that all methods of Chapter 6 (except [NA02]) compare (mostly visually) their proposed algorithms to their analogues that do not take the temporal dimension into account – in most cases the improvement is visible, and does not strictly require quantitative reasoning to prove the point. At the same time, only one effort ([SW09b]) compares alternative spatiotemporal algorithms per se and shows general superiority of the proposed method to one alternative. In moving forward, intra-algorithm comparison should be accompanied by inter-algorithm comparison for more clear understanding of the needs of spatiotemporal stereo.

Chapter 8

Open Problems

8.1 Outline

This chapter will conclude the present document by summarizing what needs to be done in the future for productive research in stereomotion processing. Specifically, we discuss modeling theory, extended goals for stereomotion and effective evaluation methodologies to keep forthcoming innovations under suitable empirical control.

8.2 Modeling theory

The field of stereomotion has come a long way from being just separate stereo and motion processing to a unique trend that is associated with its own problems and solutions. Deeper understanding of these issues and the search for better solutions require appropriate modeling. Looking back at the previous chapters a number of points can be highlighted

- Ideally, we should strive for tighter integration of stereo and motion computation in order to benefit situations when separate stereo and motion processing are insufficient.

- It is important to consider special cases when stereo is of prime importance, while motion may be left unrecovered. The alternative scenario may also be of interest: recovery of ego-motion without concern for the exact structure of the surrounding 3D world. The general point is to recover only the needed properties of the system, while marginalizing over the remaining latent quantities.

- It is worthwhile to explore alternative data descriptors, aside from pixels, that are suitable specifically for the spatiotemporal problem, as they might result in computationally beneficial methods. One such example is the P-field [BS91], but this representation has yet to see much development. Another may be spatiotemporal orientation set forth as stequel [SW09b], which seems quite promising.

- It is useful to consider measurements tuned to specific applications, e.g. [GT95] considered time to impact measurement instead of distance and velocity measurements for obstacle avoidance.

8.3 New avenues of spatiotemporal stereo

Most extant methods sought for stereo and motion cooperation as a more reliable alternative to structure-from-stereo and structure-from-motion per se. Typically, each method can separately solve the problems at hand (modulo certain inherent ambiguities discussed in Chapter 2) and the effect of combination can be perceived as simply fusion of estimation results with overall outcome being of better quality.

However, it is important to find situations that are inherently difficult or even impossible for each modality separately, but are resolvable, once joint stereo and motion processing is considered.

Examples follow:

- Lambertian scenes with dynamically changing lighting conditions. A situation well-defined for stereo processing, but complicated for motion, as the temporal brightness constancy assumption is fundamentally false.

- Scenes with specularities. Stereo processing results in matching of virtual features, producing completely false depth estimation; at the same time, specular motion is quite distinctive from real scene motion – the corresponding trajectories have significantly different characteristics [ON96].

- Generalized surface reflectance modeling. There is a strong possibility to perform depth and motion recovery for non-Lambertian surfaces, e.g., ones that can be expressed as a sum of Lambertian and specular components, as more data measurements are available in stereomotion.

- Scenes with transparencies. Static stereo processing of scenes with transparencies is an open, largely unresolved, problem; however, multi-layer motion processing is better developed and has successfully demonstrated the ability to recover multiple motions at single points.

- Scenes with repetitive textures. Stereo extensively relies on local matching, which is inherently difficult to conduct in the presence of repetitive texture, or when foreground and background are very similar, a so-called camouflage effect. Motion, on the other hand, is predominantly based on differential descriptors and is very successful in breaking stereo camouflage. It is important to acknowledge that this case is most important when the number of cameras is small, e.g., two as in the ubiquitous case of binocular stereo. When the number of cameras is very large, the complete lightfield is adequately sampled and the only ambiguity of stereo is purely textureless regions [BSK01].

- Temporally coherent 3D structure estimates. Spatiotemporal stereo must exhibit continuity in structure (and motion) estimation from one time instance to another. Ideally, it should be able to construct some analog of 3D panoramas. Note that certain attempts have already been demonstrated [GM04], where the authors treat the spatiotemporal surface reconstruction problem as isosurface estimation in 4D space.

8.4 Experimental evaluation

Adequate experimental evaluation procedures can be derived from the extant techniques used for stereo and motion evaluation, which Chapter 7 extensively covered.

Significant improvement has been recently achieved in quantitative evaluation of stereo processing, as robust and reliable methods for ground truth collection were proposed, e.g. [SS03, SCD$^+$06].

Such techniques were already reused for evaluation of temporally-consistent structure estimation [ZN05, SW09b].

In terms of motion, quantitative evaluation is more challenging and initially consisted mostly of synthetic scenes, e.g. Yosemite in [BFB94]. Recent technology allows for real, non-trivial setups [BSL$^+$07]. Still, in stereomotion, most methods presented so far considered quantitative evaluation of synthetic scenes only, e.g. [HD07, WRV$^+$08]. Non-trivial real scenes with ground truth are still to be introduced.

More generally, it is important to develop datasets that will stress the importance of joint stereomotion processing by considering the hard cases described in the previous section, such as camouflage, or changing lighting conditions. A reasonable attempt in this direction has been made by [SW09b] where the authors obtain lab scenes with disparity ground truth only, but the scenes themselves contain binocular camouflage, epipolar-aligned texture and non-Lambertian surfaces. Ideally, datasets covering all cases outlined in Sec. 8.3 must be designed to boost research in stereomotion.

Finally, it is important to say a word of caution about extensive evaluation on standard datasets. The advantage of rigorous comparison and existence of quantitative performance characterization is obvious. At the same time, this creates the danger of putting too much emphasis on behaviour on particular datasets – in turn, solutions that underperform to the current state-of-the-art automatically get less significance and may be prematurely disregarded in future research, though they may have potential.

Bibliography

[AB85] Edward H. Adelson and James R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, 2(2):284–299, 1985.

[AB06] A. Agarwal and A. Blake. The Panum Proxy algorithm for dense stereo matching over a volume of interest. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2339–2346, 2006.

[ABG89] Nicolas Alvertos, Dragana Brzakovic, and Rafael C. Gonzalez. Camera geometries for image matching in 3-D machine vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 11(9):897–915, 1989.

[AH90] Yiannis Aloimonos and Jean-Yves Herve. Correspondenceless stereo and motion: Planar surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 12:504–510, 1990.

[AK06] T. Amiaz and N. Kiryati. Piecewise-smooth dense optical flow via level sets. *International Journal of Computer Vision (IJCV)*, 68:111–124, 2006.

[AKK05] Motilal Agrawal, Kurt Konolige, and Luca Kurt. Real-time detection of independent motion using stereo. In *Workshop on Applications of Computer Vision (WACV)*, pages 207–214, 2005.

[Ana89] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision (IJCV)*, 2:283–301, 1989.

[AP95] Ali Azarbayejiani and Alex P. Pentland. Recursive estimation of motion, structure and focal length. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 17(6), 1995.

[AW87] J. Aggarwal and Y. Wang. Analysis of a sequence of images using point and line correspondences. In *International Conference on Robotics and Automation (ICRA)*, volume 4, pages 1275–1280, 1987.

[BA83] Peter J. Burt and Edward H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.

[BA96] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding (CVIU)*, 61(1):75–104, 1996.

[BAHH91] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 5–10, 1991.

[BBH03] Myron Z. Brown, Darius Burschka, and Gregory D. Hager. Advances in computational stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(8):993–1008, 2003.

[BBHP92] James R. Bergen, Peter J. Burt, Rajesh Hingorani, and Shmeul Peleg. A three-frame algorithm for estimating two-component image motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 14(9):886–896, 1992.

[BC91] Ted J. Broida and Rama Chellappa. Estimating the kinematics and structure of a rigid object from a sequence of monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 13(6):497–513, 1991.

[BE95] J. L. Barron and R. Eagleson. Binocular estimation of motion and structure from long sequences using optical flow without correspondence. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 2, pages 193–196, 1995.

[Bes86] J. Besag. On the statistical analysis of dirty pictures. *Journal of Royal Statistical Society, Series B*, B-48(3):259–302, 1986.

[BFB94] J. L. Barron, D. J. Fleet, and S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision (IJCV)*, 12:43–77, 1994.

[BGRR09] Michael Bleyer, Margrit Gelautz, Carsten Rother, and Christoph Rhemann. A stereo approach that handles the matting problem via image warping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[Big98] J. Bigun. *Vision with Direction.* Springer, 1998.

[BJ80] P. Burt and B. Julesz. A disparity gradient limit for binocular fusion. *Nature*, 208:615–617, 1980.

[BK83] Dana H. Ballard and O. A. Kimball. Rigid body motion from depth and optical flow. *Computer Vision, Graphics and Image Processing*, 22:95–115, 1983.

[BKY99] Peter N. Belhumeur, David J. Kriegman, and Alan L. Yuille. The bas-relief ambiguity. *International Journal of Computer Vision (IJCV)*, 35(1):33–44, 1999.

[BM04] Simon Baker and Ian Matthews. Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision (IJCV)*, 56(3):221–255, 2004.

[BN98] Dinkar N. Bhat and Shree K. Nayar. Ordinal measures for image correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(4):415–423, April 1998.

[BR96] M. Black and A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision (IJCV)*, 19(1):57–92, July 1996.

[BS91] P. Balasubramanyam and M. A. Snyder. P-field: a computational model for binocular motion processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 115–120, 1991.

[BSK01] Simon Baker, T. Sim, and Takeo Kanade. A characterization of inherent stereo ambiguities. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 428–435, 2001.

[BSL$^+$07] Simon Baker, Daniel Scharstein, J.P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski. A database and evaluation methodology for optical flow. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[BT98] Stan Birchfield and Carlo Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(4):401–406, 1998.

[Bur45] H. Burton. The optics of Euclid. *Journal of Optical Society of America*, 35:357–372, 1945.

[BVZ01] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1222–1239, 2001.

[BWF$^+$05] Andrs Bruhn, Joachim Weickert, Christian Feddern, Timo Kohlberger, and Christoph Schnorr. Variational optical flow computation in real time. *IEEE Transactions on Image Processing*, 14(5):608–615, 2005.

[BWS05] Andres Bruhn, Joachim Weickert, and Christoph Schnorr. Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. *International Journal of Computer Vision (IJCV)*, 63:211–231, 2005.

[BZ87] Andrew Blake and Andrew Zisserman. *Visual Reconstruction.* The MIT Press, Cambridge, MA, USA, 1987.

[CK02] Rodrigo L. Carceroni and Kiriakos N. Kutulakos. Multi-view scene capture by surfel sampling: From video streams to non-rigid 3D motion, shape and reflectance. *International Journal of Computer Vision (IJCV)*, 49(2-3):175–214, 2002.

[CLTS97] S. Crossley, A. J Lacey, N. A. Thacker, and N. L. Seed. Robust stereo via temporal consistency. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 659–668, 1997.

[CMR07] A.I. Comport, E. Malis, and P. Rives. Accurate quadrifocal tracking for robust 3D visual odometry. In *International Conference on Robotics and Automation (ICRA)*, pages 40–45, 2007.

[CR76] Ciro Cafforio and Fabio Rocca. Methods for measuring small displacements of television images. *IEEE Transactions of Information Theory*, 22(5):573–579, 1976.

[CRZ00] Antonio Criminisi, Ian D. Reid, and Andrew Zisserman. Single view metrology. *International Journal of Computer Vision (IJCV)*, 40(2), 2000.

[CW07] Kevin Cannons and Richard P. Wildes. Spatiotemporal oriented energy features for visual tracking. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 532–543, 2007.

[DA89] Umesh R. Dhond and J. K. Aggarwal. Structure from stereo - A review. *IEEE Transactions on System, Man and Cybernetics*, 19(6):1489–1510, 1989.

[DD02] D. Demirdjian and T. Darrell. Using multiple-hypothesis disparity maps and image velocity for 3D motion estimation. *International Journal of Computer Vision (IJCV)*, 47:219–228, 2002.

[Der06] Konstantinos G. Derpanis. Characterizing image motion. Technical Report CS-2006-06, York University, 4700 Keele street, Toronto, Ontario, Canada, 2006.

[DG05] K. G. Derpanis and J.M. Gryn. Three-dimensional n-th derivative of Gaussian separable steerable filters. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 3, pages 553–556, 2005.

[DH00] David Demirdjian and Radu Horaud. Motion-egomotion discrimination and motion segmentation from image-pair streams. *Computer Vision and Image Understanding (CVIU)*, 78:53–68, 2000.

[DM92] Ernst D. Dickmanns and Birger D. Mysliwetz. Recursive 3-D road and relative egostate recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 14:199–213, 1992.

[DMC90] Ernst D. Dickmanns, Birger D. Mysliwetz, and Th. Christians. An integrated spatiotemporal approach to automatic visual guidance of autonomous vehicles. *IEEE Transactions of systems, man, and cybernetics*, 20(6):1273–1284, 1990.

[DRR05] James Davis, Ravi Ramamoorthi, and Szymon Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(2):296–302, 2005.

[DW09] K. G. Derpanis and R. P. Wildes. Early spatiotemporal grouping with a distributed oriented energy representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[Egn00] Geoffrey Egnal. Mutual information as a stereo correspondence measure. Technical Report MS-CIS-00-20, University of Pennsylvania, 2000.

[ER04] Michael J. Evans and Jeffrey S. Rosenthal. *Probability and statistics: the science of uncertainty.* W. H. Freeman and Company, New York, 2004.

[EW02] Geoffrey Egnal and Richard P. Wildes. Detecting binocular half-occlusions: Empirical comparisons of five approaches. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24(8):1127–1133, August 2002.

[FA91] William T. Freeman and Edward H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 13(9):891–906, 1991.

[Fau95] Olivier Faugeras. Stratification of 3-D visoin: projective, affine, and metric representations. *JOSA*, 12(3):465–484, 1995.

[FCSS09] Yasutaka Furukawa, Brian Curless, Steven M. Seitz, and Richard Szeliski. Manhattan-world stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[FEG87] W. Forstner, E., and Gulch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data*, pages 281–305, 1987.

[FH01] U. Frese and G. Hirzinger. Simultaneous localization and mapping - A discussion. In *Proc. of the IJCAI Workshop on Reasongin with Uncertainty in Robotics*, 2001.

[FH04] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient belief propagation for early vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 261–268, 2004.

[FJ90] David J. Fleet and Alan D. Jepson. Computation of component image velocity from local phase information. *International Journal of Computer Vision (IJCV)*, 5(1):77–104, 1990.

[FJJ91] David J. Fleet, Allan D. Jepson, and Michael R. M. Jenkin. Phase-based disparity measurement. *Computer Vision, Graphics and Image Processing*, 53(2):198–210, 1991.

[Fra08] Frank Dellaert et. al. 4D cities project, 2008.

[FRBG05] Uwe Franke, Clemens Rabe, Hernan Badino, and Stefan Gehrig. 6d-vision: Fusion of stereo and motion for robust environment perception. In *DAGM*, pages 216–223, 2005.

[FT79] Claude L. Fennema and William B. Thompson. Velocity determination in scenes containing several moving objects. *Computer Vision, Graphics and Image Processing*, 9:301–315, 1979.

[Gab46] D. Gabor. Theory of communication. *Journal of the Institution of Electrical Engineers*, 93:429–457, 1946.

[GCS06] Michael Goesele, Brian Curless, and Steven M. Seitz. Multi-view stereo revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2402–2409, 2006.

[GG84] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 6:721–741, 1984.

[GK95] G. Granlund and H. Knutsson. *Signal Processing for Computer Vision.* Kluwer, 1995.

[GKR94] V. Granville, M. Krivanek, and J.-P. Rasson. Simulated annealing: a proof of convergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 16:652–656, 1994.

[GM04] Bastian Goldluecke and Marcus Magnor. Space-time isosurface evolution for temporally coherent 3D reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 350–355, 2004.

[Gon06] Minglun Gong. Enforcing temporal consistency in real-time stereo estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 564–577, 2006.

[Gov01] Venu Madhav Govindu. Combining two-view constraints for motion estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 218–225, 2001.

[Gri81] Eric L. W. Grimson. *From Images to Surfaces: A Computational Study of the Human Early Visual System.* MIT Press Cambridge, MA, USA, 1981.

[GSS93] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F*, 140(2):107–113, 1993.

[GST89] Enrico Grosso, Giulio Sandini, and Massimo Tistarelli. 3-D object reconstruction using stereo and motion. *IEEE Trans. on System, Man, and Cybernetics*, 19(6):1465–1476, 1989.

[GT95] Enrico Grosso and Massimo Tistarelli. Active/dynamic stereo vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 17(9):868–879, 1995.

[GVT89] F. Girosi, A. Verri, and V. Torre. Constraints for the computation of optical flow. In *Proceedings of the IEEE Workshop on Visual Motion*, pages 116–124, 1989.

[GY01] Munglun Gong and Yee-Hong Yang. Multi-resolution stereo matching using genetic algorithm. *Proceedings of the IEEE Workshop on Stereo and Multi-Baseline Vision*, pages 21–29, 2001.

[Han91] K. Hanna. Direct multi-resolution estimation of ego-motion and structure from motion. In *Proceedings of the IEEE Workshop on Visual Motion*, pages 156–162, 1991.

[Has74] B. Haskell. Frame-to-frame coding of television pictures using two-dimensional Fourier transforms. *IEEE Transactions of Information Theory*, 20(1):119–120, 1974.

[HB85] T.S. Huang and S.D. Blostein. Robust algorithms for motion estimation based on two sequential stereo image pairs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 518–523, 1985.

[HC96] Radu Horaud and Gabriella Csurka. Self-calibration and euclidean reconstruction using motions of a stereo rig. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 96–103, 1996.

[HC00] Pui-Kuen Ho and Ronald Chung. Stereo-motion with stereo and motion in complement. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(2):215–220, 2000.

[HD07] Frederic Huguet and Frederic Devernay. A variational method for scene flow estimation from stereo sequences. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–7, 2007.

[Hee88] David J. Heeger. Optical flow using spatiotemporal filters. *International Journal of Computer Vision (IJCV)*, 1(4):279–302, 1988.

[Hee90] Joachim Heel. Direct dynamic motion vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1142–1147, 1990.

[HF01] Horst W. Haussecker and David J. Fleet. Computing optical flow with physical models of brightness variation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(6):661–673, 2001.

[HH91] Berthold K. P. Horn and John G. Harris. Rigid body motion from range image sequences. *Computer Vision, Graphics and Image Processing*, 53(1):1–13, 1991.

[Hir05] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 807–814, 2005.

[HJ92] David J. Heeger and Allan D. Jepson. Subspace methods for recovering rigid motion i: Algorithm and implementation. *International Journal of Computer Vision (IJCV)*, 7(2):95–117, 1992.

[HO93] Keith J. Hanna and Neil E. Okamoto. Combining stereo and motion analysis for direct estimation of scene structure. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 357–365, 1993.

[Hog] Andrew Hogue. Simultaneous localization and mapping techniques.

[HP94] Anthony Yuk-Kwan Ho and Ting-Chuen Pong. Cooperative fusion of stereo and motion. In *International Symposium on Speech, Image Processing and Neural Networks*, pages 292–295, 1994.

[HR02] I. Howard and B. Rogers. *Seing in Depth.* I. Porteus, Thornhill, Ontario, Canada, 2002.

[HRD$^+$99] M. Harville, A. Rahimi, T. Darrell, G. Gordon, and J. Woodfill. 3D pose tracking with linear depth and brightness constraints. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 206–213, 1999.

[HS81] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, pages 185–203, 1981.

[HS88] C. Harris and M.J. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–152, 1988.

[HS08] Heiko Hirschmuller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[HW88] Berthold K. P. Horn and E. J. Weldon Jr. Direct methods for recovering motion. *International Journal of Computer Vision (IJCV)*, 2:51–76, 1988.

[HZ04] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision. Second Edition.* Cambridge University Press, Cambridge, UK, 2004.

[IA99] Michal Irani and P. Anandan. All about direct methods. In *LNCC 1883*, pages 267–277, 1999.

[IB98] Michael Isard and Andrew Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision (IJCV)*, 29(1):5–28, 1998.

[IG06] Hiroshi Ishikawa and Davi Geiger. Rethinking the prior model from stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 3, pages 526–537, 2006.

[IM06] Michael Isard and John MacCormick. Dense motion and disparity estimation via loopy belief propagation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, volume 2, pages 32–41, 2006.

[Ish03] Hiroshi Ishikawa. Exact optimization for markov random fields with convex priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(10):1333–1336, 2003.

[JT86] Michael Jenkin and John K. Tsotsos. Applying temporal constraints to the dynamic stereo problem. *Computer Vision, Graphics and Image Processing*, 33:16–32, 1986.

[KA87] Yeon Kim and J. Aggarwal. Determining object motion in a sequence of stereo images. *IEEE Journal of Robotics and Automation (RA)*, 3(6):599–614, 1987.

[Kac04] Per-Jonny Kack. Robust stereo correspondence using graph cuts. Master's thesis, School of Computer Science and Engineering, Royal Institute of Technology, Stockholm, 2004.

[Kal60] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):3545, 1960.

[KBD04] Zia Khan, Tucker Balch, and Frank Dellaert. A rao-blackwellized particle filter for eigentracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 980–986, 2004.

[KGV83] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[Kit96] Genshiro Kitagawa. Monte carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

[KKZ03] Junhwan Kim, Vladimir Kolmogorov, and Ramin Zabih. Visual correspondence using energy minimization and mutual information. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1033–1040, October 2003.

[KN03] Ali Khamene and Shahriar Negahdaripour. Motion and structure from multiple cues; image motion, shading flow, and stereo disparity. *Computer Vision and Image Understanding (CVIU)*, pages 99–127, 2003.

[Koc95] R. Koch. 3-D surface reconstruction from stereoscopic image sequences. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 109–114, 1995.

[Kol06] Vladimir Kolmogorov. Convergent tree-reweighted message passign for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28:1568–1583, October 2006.

[Kos93] Andreas Koschan. What is new in computational stereo since 1989: A survey on current stereo papers. Technical Report 93-22, Technical University of Berlin, 1993.

[KP81] G. Konecny and D. Pape. Correlation techniques and devices. *Photogrammetric Engineering and Remote Sensing*, pages 323–333, 1981.

[KS00] Kiriakos N. Kutulakos and Steven M. Seitz. A theory of spape by space carving. *International Journal of Computer Vision (IJCV)*, 38(3):199–218, 2000.

[Kv80] J. Krol and W. van der Grind. The double nail illusion. *Perception*, 9:651–659, 1980.

[Kv97] Jan J. Koenderink and Andrea J. van Doorn. The generic bilinear calibration-estimation problem. *International Journal of Computer Vision (IJCV)*, 23(3):217–234, 1997.

[LALS04] Carlos Leung, Ben Appleton, Brian C. Lovell, and Changming Sun. An energy minimisation approach to stereo-temporal dense reconstruction. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, pages 72–75, 2004.

[LFAW08] C. Liu, W. T. Freeman, E. H. Adelson, and Y. Weiss. Human-assisted motion annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[LH91] Mun K. Leung and Thomas S. Huang. An integrated approach to 3-D motion analysis and object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 13(10):1075–1084, 1991.

[LK81] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, page 674679, 1981.

[LK90] S. Lee and Y. Kay. A kalman filter approach for accurate 3d motion estimation from asequence of stereo images. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, volume 1, pages 104–108, 1990.

[LM67] R. B. Lawson and D. C. Mount. Minimum condition for stereopsis and anomalous contour. *Science*, 158:802–804, November 1967.

[LM75] J. Limb and J. Murphy. Estimating velocity of moving images in television signals. *Computer Vision, Graphics and Image Processing*, 4(3):311–327, 1975.

[LMPF07] E. Scott Larsen, Philippos Mordohai, Marc Pollefeys, and Henry Fuchs. Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[Low04] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[LRR08] Victor Lempitsky, Stephen Roth, and Carter Rother. FusionFlow: Discrete-continuous optimization for optical flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[LS93] Jin Liu and Robert Skerjanc. Stereo and motion correspondence in a sequence of stereo images. *Signal Processing: Image Communication*, 5:305–318, 1993.

[LS08] Rui Li and Stan Sclaroff. Multi-scale 3d scene flow from binocular stereo sequences. *Computer Vision and Image Understanding (CVIU)*, 110(1):75–90, 2008.

[LSY06] Cheng Lei, J. Selzer, and Yee-Hong Yang. Region-tree based stereo using dynamic programming optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2378–2385, 2006.

[LX06] Zongqing Lu and Weixin Xie. A PDE-based method for optical flow estimation. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, pages 846–849, 2006.

[LZ03] Gand Li and Steven W. Zucker. A differential geometrical model for contour-based stereo correspondence. In *Proceedings of the IEEE Workshop on Variational, Geometric, and Level Set Methods in Computer Vision, Nice, France*, 2003.

[MA78] W. N. Martin and J. K Aggarwal. Dynamic scene analysis. *Computer Vision, Graphics and Image Processing*, 7:356–374, 1978.

[MAK$^+$06] S.P. Mallick, S. Agarwal, D. J. Kriegman, S. J. Belongie, B. Carragher, and C. S. Potter. Structure and view estimation for tomographic reconstruction: A bayesian approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2253–2260, 2006.

[MB00] Nicholas Molton and Michael Brady. Practical structure and motion from stereo when motion is unconstrained. *International Journal of Computer Vision (IJCV)*, 39(1):5–23, 2000.

[Mid08] Middlebury College Stereo Vision Page. http://www.middlebury.edu/stereo/, 2008.

[MKC00] Mikhail Mozerov, Vitaly Kober, and Tae-Sun Choi. Improved motion stereo matching based on a modified dynamic programming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2501–2505, 2000.

[MKS05] Dongbo Min, Hansung Kim, and Kwanghoon Sohn. Edge-preserving joint motion-disparity estimation in stereo image sequences. *Signal Processing: Image Communication*, 21:252–271, 2005.

[Mor81] Hans Moravec. Rover visual obstacle avoidance. In *Proceedings of the IEEE International Joint Conference on Artificial Intelligence*, pages 785–790, 1981.

[Mou69] F. Mounts. A video encoding system with conditional picture-element replenishment. *The Bell Systems Technical Journal*, 48(7):2545–2554, 1969.

[MP76] D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194:283–287, 1976.

[MS97] Sotiris Malassiotis and Michael G. Strintzis. Model-based joint motion and structure estimation from stereo images. *Computer Vision and Image Understanding (CVIU)*, 65(1):79–94, 1997.

[MS06] Dongbo Min and Kwanghoon Sohn. Edge-preserving simultaneous joint motion-disparity estimation. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, pages 74–77, 2006.

[MSS99] R. Mandelbaum, G. Salgian, and H. Sawhney. Correlation-based estimation of ego-motion and structure from motion and stereo. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 544–550, 1999.

[Mut86] Kathleen M. Mutch. Determining object translation information using stereoscopic motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 8(6):750–755, 1986.

[MYW05] Talya Meltzer, Chen Yanover, and Yair Weiss. Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 428–435, 2005.

[NA02] Jan Neumann and Yiannis Aloimonos. Spatio-temporal stereo using multi-resolution subdivision surfaces. *International Journal of Computer Vision (IJCV)*, 47(1-3):181–193, 2002.

[Nag83] H. H. Nagel. Displacement vectors derived from second-order intensity variations in image sequences. *Computer Vision, Graphics and Image Processing*, 21:85–117, 1983.

[NDF90] N. Navab, R. Deriche, and O. D. Faugeras. Recovering 3D motion and structure from stereo and 2D token tracking cooperation. In *ICCV*, pages 513–516, 1990.

[Neg98] Shahriar Negahdaripour. Revised definition of optical flow: Integration of radiometric and geometric cues for dynamic scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20:961–979, 1998.

[Nev76] Ramakant Nevatia. Depth measurement by motion stereo. *Computer Vision, Graphics and Image Processing*, 5:203–214, 1976.

[NF03] O. Nestares and D. J. Fleet. Error-in-variables likelihood functions for motion estimation. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 3, page 7780, 2003.

[NFH00] O. Nestares, D. J. Fleet, and D. Heeger. Likelihood functions and confidence bounds for total-least-squares problems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 523–530, 2000.

[NMSO96] Y. Nakamura, T. Matsuura, K. Satoh, and Y. Ohta. Occlusion detectable stereo - occlusion patterns in camera matrix. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 371–378, 1996.

[NNB04] David Nister, Oleg Naroditsky, and James R. Bergen. Visual odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 652–659, 2004.

[NS02] Jean-Christophe Nebel and Alexander Sibiryakov. Range flow from stereo-temporal matching: Application to skinning. In *IASTED Int. Conf. on Visualization, Imaging, and Image Processing*, 2002.

[OMSM00] Clark F. Olson, Larry H. Matthies, Marcel Schoppers, and Mark W. Maimone. Robust stereo ego-motion for long distance navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 453–458, 2000.

[ON94] M. Otte and H.-H. Nagel. Optical flow estimation: advances and comparisons. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 1, pages 51–60, 1994.

[ON96] Michael Oren and Schree K. Nayar. A theory of specular surface geometry. *International Journal of Computer Vision (IJCV)*, 24:105–124, 1996.

[OWF08] I. Reid O. Woodford, P. Torr and A. Fitzgibbon. Global stereo reconstruction under second order smoothness priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[PAT96] I. Patras, N. Alvertos, and G. Tziritas. Joint disparity and motion field estimation in stereoscopic imagesequences. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, volume 1, pages 359–363, 1996.

[PBB$^+$06] Nils Papenberg, Andres Bruhn, Thomas Brox, Stephan Didas, and Joachim Weickert. Highly accurate optic flow computation with theoretically justified warping. *International Journal of Computer Vision (IJCV)*, 67(2):141–158, 2006.

[PK94] C. Poelman and T. Kanade. A paraperspective factorization method for shape and motion recovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 2, pages 97–108, 1994.

[PKF05] Jean-Philippe Pons, Renaud Keriven, and Olivier Faugeras. Modelling dynamic scenes by registering multi-view image sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 822–827, 2005.

[PKFH03] Jean-Philippe Pons, Renaud Keriven, Olivier D. Faugeras, and Gerardo Hermosillo. Variational stereovision and 3D scene flow estimation with statistical similarity measures. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 597–602, 2003.

[PKRC00] M. S. Park, D. Y. Kim, K. S. Roh, and T. S. Choi. Motion stereo based on fourier local phase adaptive matching. *Optical Engineering*, 39(4):866–871, 2000.

[PMF85] S. Pollard, J. Mayhew, and J. Frisby. PMF: A stereo correspondence algorithm using a disparity gradient limit. *Perception*, 14:449–470, 1985.

[Poi08] Point Grey Research. http://www.ptgrey.com, 2008.

[Pra83] K. Prazdny. On the information in optical flows. *Computer Vision, Graphics and Image Processing*, 22:239–259, 1983.

[PvP96] Marc Pollefeys, Luc J. van Gool, and Marc Proesmans. Euclidean 3D reconstruction from image sequences with variable focal lenghts. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 1, pages 31–42, 1996.

[RAP06] Alex Rav-Acha and Shmuel Peleg. Lucas-Kanade without iterative warping. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 1097–1100, 2006.

[RB05] Stefan Roth and Michael J. Black. Fields of experts: A framework for learning image priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–867, 2005.

[RB06] Stefan Roth and Michael J. Black. Specular flow and the recovery of surface structure. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1869–1876, 2006.

[RFG07] C. Rab, U. Franke, and S. K. Gehrig. Fast detection of moving objects in complex scenarios. In *IEEE Intelligent Vehicle Symposium*, 2007.

[RG00] S. Roy and V. Govindu. MRF solutions for probabilistic optical flow formulations. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, volume 3, pages 1041–1047, 2000.

[Ric77] J. Richter (Ed.). *Selections from the Notebooks of Leonardo da Vinci.* Oxford, U.K.: Oxford University Press, 1977.

[Ric85] Whitman Richards. Structure from stereo and motion. *JOSA*, 2:343–349, 1985.

[RKLS07] C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer. Optimizing binary MRFs via extended roof duality. In *CVPR*, 2007.

[RN03] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall, Upper Saddle River, NJ, USA, 2003.

[SA91] Minas Spetsakis and John Yiannis Aloimonos. A multi-frame approach to visual motion perception. *International Journal of Computer Vision (IJCV)*, 6(3):1573–1405, 1991.

[SAS08] Tobias Schuchert, Til Aach, and Hanno Scharr. Range flow for varying illumination. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 1, pages 509–522, 2008.

[SBBB95] G. Sudhir, Subhashis Baneerjee, K. K. Biswas, and R. Bahl. Cooperative integration of stereopsis and optic flow computation. *JOSA-A*, 12(12):2564–2572, 1995.

[SCD$^+$06] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 519–528, 2006.

[SG99] Richard Szeliski and Polina Golland. Stereo matching with transparency and matting. *International Journal of Computer Vision (IJCV)*, 32:45–61, 1999.

[Sha02] Juliang Shao. Generation of temporally consistent multiple virtual camera views from stereoscopic image sequences. *International Journal of Computer Vision (IJCV)*, 47:171–180, 2002.

[SI07] Eli Shechtman and Michal Irani. Space-time behavior-based correlation - or - how to tell if two underlying motion fields are similar without computing them? *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(11):2045–2056, 2007.

[Sin90] A. Singh. An estimation-theoretic framework for image-flow computation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 168–177, 1990.

[Siz08] Mikhail Sizintsev. Hierarchical stereo with thin structures and transparency. In *Canadian Conference on Computer and Robot Vision (CRV)*, pages 97–104, 2008.

[SJ05] Stephen Se and Piotr Jasiobedzki. Instant scene modeler for crime scene reconstruction. In *Proceedings of the IEEE Workshop on Advanced 3D Imaging for Safety and Security (A3DISS), San Diego, USA*, june 2005.

[SJB00] Hagen Spies, Bernd Jahne, and John L. Barron. Dense range flow from depth and intensity data. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, pages 131–134, 2000.

[SJB02] Hagen Spies, Bernd Jahne, and John L. Barron. Range flow estimation. *Computer Vision and Image Understanding (CVIU)*, 85:209–231, 2002.

[SJW07] S. Se, P. Jasiobedzki, and R. Wildes. Stereo-vision based 3D modeling of space structures. In *In Proceedings of SPIE*, volume 6555, 2007.

[SK02] Steven M. Seitz and Jiwon Kim. The space of all stereo images. *International Journal of Computer Vision (IJCV)*, 48(1):21–38, 2002.

[SLKS05] Jian Sun, Yin Li, Sing Bing Kang, and Heung-Yeung Shum. Symmetric stereo matching for occlusion handling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 399–406, June 2005.

[SRLB08] Deqing Sun, Stefan Roth, J. P. Lewis, and Michael J. Black. Learning optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 3, pages 83–97, 2008.

[SS98a] Daniel Scharstein and Richard Szeliski. Stereo matching with nonlinear diffusion. *International Journal of Computer Vision (IJCV)*, 28(2):155–174, 1998.

[SS98b] Gideon P. Stein and Amnon Shashua. Direct estimation of motion and extended scene structure from a moving stereo rig. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 211–218, 1998.

[SS00] Gideon P. Stein and Amnon Shashua. Model-based brightness constrains: On direct estimation of structure and motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(9):992–1015, 2000.

[SS02] D. Scharstein and R. Szeliski. Taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision (IJCV)*, 47:7–42, 2002.

[SS03] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 195–202, 2003.

[SS04] Richard Szeliski and Daniel Scharstein. Sampling the disparity space image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(3):419–425, 2004.

[SS06] Hanno Scharr and Tobias Schuchert. Simultaneous motion, depth and slope estimation with a camera-grid. In *Vision, Modelling and Visualization*, pages 81–88, 2006.

[SSCB03] Silvio P. Sabatini, Fabio Solari, Paolo Cavalleri, and Giacomo Mario Bisio. Phase-based binocular perception of motion in depth: cortical-like operators and analog VLSI architectures. *EURASIP Journal on Applied Signal Processing*, 2003:690–702, 2003.

[ST94] Jianbo Shi and Carlo Tomasi. Good features to track. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, 1994.

[ST96] Peter Sturm and William Triggs. A factorization based algorithm for multi-image projective structure and motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 709–720, 1996.

[Str86] Gilbert Strang. *Introduction to applied mathematics.* Wellesley-Cambridge Press, 1986.

[Sv02] Christoph Strecha and Luc van Gool. Motion-stereo integration for depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 170–185, 2002.

[SW00] Amnon Shashua and Lior Wolf. On the structure and properties of the quadrifocal tensor. In *ECCV*, pages 710–724, 2000.

[SW01] Amnon Shashua and Yonatan Wexler. Q-warping: Direct computation of quadratic reference surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(8):920–925, 2001.

[SW06] Mikhail Sizintsev and Richard Wildes. Coarse-to-fine stereo vision with accurate 3D boundaries. Technical Report CS-2006-07, York University, Toronto, Canada, 2006.

[SW09a] Mikhail Sizintsev and Richard P. Wildes. Coarse-to-fine stereo vision with accurate 3D boundaries. *Image and Vision Computing*, 2009.

[SW09b] Mikhail Sizintsev and Richard P. Wildes. Spatiotemporal stereo via spatiotemporal quadric element (stequel) matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[Sze88] R. Szeliski. Estimating motion from sparse range data without correspondence. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2007–215, 1988.

[SZS94] Jen-Yu Shieh, Hanqi Zhuang, and R. Sudhakar. Motion estimation from a sequence of stereo images: a direct method. *IEEE Trans. on System, Man, and Cybernetics*, 24(7):1044–1053, 1994.

[SZS03] Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(7):787–800, 2003.

[SZS$^+$06] Richard Szeliski, Ramin Zabih, Daniel Scharstein, Olga Veskler, Vladimir Kolmogorov, Aseem Agarwala, Marshall Tappen, and Carsten Rother. Comparative study of energy minimization methods for markov random fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 2, pages 16–29, 2006.

[SZS$^+$08] Richard Szeliski, Ramin Zabih, Daniel Scharstein, Olga Veskler, Vladimir Kolmogorov, Aseem Agarwala, Marshall Tappen, and Carsten Rother. Comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(6):1068–1080, 2008.

[TBF05] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic robotics.* MIT Press, 2005.

[Thr02] Sebastian Thrun. *Robotic Mapping: A survey*, chapter Exploring Artifical Intelligence in the New Millenium. Morgan Kaufmann, 2002.

[TK91] Carlo Tomasi and Takeo Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie-Mellon University, Pittsburg, USA, 1991.

[TK92] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: A factorization approach. *International Journal of Computer Vision (IJCV)*, 9(2):137–154, 1992.

[TKS06] Yanghai Tsin, Sing Bing Kang, and Richard Szeliski. Stereo matching with linear superposition of layers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(2):290–301, 2006.

[TSJ92] A. P. Tirumalai, B. G. Schunk, and R. C. Jain. Dynamic stereo with self-calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 14:1184–1189, 1992.

[TV98] Emanuelle Trucco and Alessandro Verri. *Introductory techniques for 3D computer vision.* Prentice Hall, Upper Saddle River, NJ, USA, 1998.

[TYZ09] TYZX, Inc. http://www.tyzx.com/, 2009.

[UGVT88] S. Uras, F. Girosi, A. Verri, and V. Torre. A computational approach to motion perception. *Biological Cybernetics*, 60(2):79–87, 1988.

[VBR$^+$05] Sundar Vedula, Simon Baker, Peter Rander, Robert T. Collins, and Takeo Kanade. Three-dimensional scene flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(3):475–480, 2005.

[VBSK00] Sundar Vedula, Simon Baker, Steven M. Seitz, and Takeo Kanade. Shape and motion carving in 6D. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2592–2598, 2000.

[Vek05] Olga Veksler. Stereo correspondence by dynamic programming on a tree. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 384–390, 2005.

[WB00] Richard P. Wildes and James Bergen. Qualitative spatiotemporal analysis using an oriented energy representation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 786–784, 2000.

[WCR92] J. Weng, P. Cohen, and N. Rebibo. Motion and structure estimation from stereo image sequences. *IEEE Journal of Robotics and Automation (RA)*, 8:362–382, 1992.

[WD86] Allen M. Waxman and James H. Duncan. Binocular image flows: Steps toward stereo-motion fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 8(6):715–729, 1986.

[WD96] Wendong Wang and James H. Duncan. Recovering the tree-dimensional motion and structure of multiple moving objects from binocular image flows. *Computer Vision and Image Understanding (CVIU)*, 63(3):430–446, 1996.

[WIM05] Oliver Williams, Michael Isard, and John MacCormick. Estimating disparity and occlusions in stereo video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 250–257, 2005.

[WJW05] M.J. Wainwright, T.S. Jaakkola, and A.S. Willsky. MAP estimation via agreement on (hyper)trees: Message-passing and linear-programming approaches. *IEEE Transactions of Information Theory*, 51(11):3697–3717, 2005.

[WM95] Joseph Weber and Jitendra Malik. Robust computation of optical flow in a multi-scale differential framework. *International Journal of Computer Vision (IJCV)*, 14(1):67–81, 1995.

[WRV$^+$08] Andreas Wedel, Clemens Rabe, Tobi Vaudrey, Thomas Brox, Uwe Franke, and Daniel Cremers. Efficient dense scene flow from sparse or dense stereo data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 1, pages 739–751, 2008.

[XJ07] Wei Xiong and Jiaya Jia. Stereo matching on objects with fractional boundary. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[YBBR93] Masnobu Yamamoto, Pierre Boulanger, J.-Angelo Beraldin, and Marc Rioux. Direct estimation of rangle flow on deformable shape from video rate range camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(1):82–89, 1993.

[YC90] G.-S. J. Young and R. Chellappa. 3-D motion estimation using a sequence of noisy stereo images: Models, estimation, and uniqueness results. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 12(8):735–759, 1990.

[YK05] Kuk-Jin Yoon and In-So Kweon. Locally adaptive support-weight correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 924–931, 2005.

[YO97] Jae-Woong Yi and Jun-Ho Oh. Recursive resolving algorithm for multiple stereo and motion matches. *Image and Vision Computing*, 15:181–196, 1997.

[YSD03] Weichuan Yu, Gerald Sommer, and Kostas Daniilidis. Multiple motion analysis: in spatial or in spectral domain? *Computer Vision and Image Understanding (CVIU)*, 90(2):129–152, 2003.

[ZCS03] Li Zhang, Brian Curless, and Steven M. Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 367–374, 2003.

[ZF92] Zhengyou Zhang and Olivier D. Faugeras. Three-dimensional motion computation and object segmentation in a long sequence of stereo frames. *International Journal of Computer Vision (IJCV)*, 7(3):211–241, 1992.

[Zha95] Zhengyou Zhang. Motion and structure of four points from one motion of a stereo rig with unknown extrinsic parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 17(12):1222–1227, 1995.

[ZHZ88] X. Zhuang, R. M. Haralick, and Y. Zhao. From depth and optical flow to rigid body motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 393–397, 1988.

[ZK00] L. Zitnick and T. Kanade. A cooperative algorithm for stereo matching with occlusion detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22:675–684, 2000.

[ZK01] Ye Zhang and Chandra Kambhamettu. On 3D scene flow and structure estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 778–785, 2001.

[ZLF96] Zhengyou Zhang, Quang-Tuan Luong, and Olivier Faugeras. Motion of an uncalibrated stereo rig: self-calibration and metric reconstruction. *IEEE Journal of Robotics and Automation (RA)*, 12(1):103–113, 1996.

[ZN04] Hongsheng Zhang and Shahriar Negahdaripour. Improved temporal correspondences in stereo-vision by RANSAC. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, pages 52–55, 2004.

[ZN05] Hongsheng Zhang and Shahriar Negahdaripour. Epiflow quadruplet matching: Enforcing epipolar geometry for spatio-temporal stereo correspondences. In *Workshop on Applications of Computer Vision (WACV)*, pages 481–486, 2005.

[ZOS$^+$07] Zhiwei Zhu, T. Oskiper, S. Samarasekera, R. Kumar, and H. S. Sawhney. Ten-fold improvement in visual odometry using landmark matching. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2007.

[ZSCS04] Li Zhang, Noah Snavely, Brian Curless, and Stephen M. Seitz. Spacetime faces: High-resolution capture for modeling and animation. In *ACM SIGGRAPH Proceedings*, pages 548–558, 2004.

[ZTCS99] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape from shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 21(8):690–706, 1999.

[ZW94] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 151–158, 1994.