Ingenieurfakultät Bau Geo Umwelt

Lehrstuhl für Methodik der Fernerkundung

Dense Stereo Matching with Robust Cost Functions and Confidence-based Surface Prior

Ke Zhu

Vollständiger Abdruck der von der Ingenieurfakultät Bau Geo Umwelt Bauingenieur- und Vermessungswesen
der Technischen Universität München zur Erlangung des akademischen Grades eines
Doktor-Ingenieurs
genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. -Ing. Uwe Stilla

Prüfer der Dissertation:

1. Univ.-Prof. Dr. -Ing. Richard Bamler
2. Hon.-Prof. Dr. -Ing. Peter Reinartz,
   Universität Osnabrück

Die Dissertation wurde am 15.10.2013 bei der Technischen Universität München eingereicht und durch die Ingenieurfakultät Bau Geo Umwelt am 17.02.2014 angenommen.

I would like to dedicate this thesis to my loving parents, my wife and our child ...

Acknowledgements

I am grateful to Prof. Dr. Richard Bamler and Prof. Dr. Peter Reinartz for giving me the opportunity to carry out the thesis at TU-München and in German Aerospace Center (DLR). I also appreciate all their contributions of time, ideas, and suggestions for completing this thesis.

And I would like to express my heartfelt gratitude to Dr. Pablo d'Angelo for his continued encouragement and invaluable suggestion throughout my doctoral research study. I would like to acknowledge Dr. Daniel Neilson for the inspired and effective cooperation during my IGSSE exchange at the University of Saskatchewan. I would like to thank Dr. Matthias Butenuth for his goal-oriented supervision.

Abstract

The goal of dense stereo matching is to estimate the distance, or depth to an imaged object in every pixel of an input image; this is achieved by finding pixel correspondences between the source image and one or more matching images. Applications that make dense stereopsis active come from areas such as photogrammetry, remote sensing, mobile robotics, and intelligent vehicles. For instance, many remote sensing applications require depth maps to generate digital elevation models from airborne and satellite imagery. There are hundreds, if not thousands, of approaches seeking to solve stereopsis. A number of factors that make computational stereopsis quite challenging become apparent once one begins to use it in real-world applications; non-Lambertian reflectance, complex scene and radiometric changes, among other factors, are usually present in real-world data.

Dense stereo algorithms can be categorized into two methodologies based on the way the problem is solved: Local and global. Local algorithms aim to solve the problem via a local analysis at each input-image pixel, whereas global algorithms formulate the stereopsis problem as one of finding an optimal solution to a global energy, or probability function. Developing a global stereo method considers three main factors – calculation reliable observation to measure matching similarity, formulation of energy, or probability, function using additional priors, and optimization of the global function to find the global extremum.

In this dissertation two methodical novelties are contributed – the *merging strategy of match costs* and the *confidence-based surface prior* incorporating a semiglobal optimization framework. All dense stereo matching algorithms use match cost functions to measure the similarity between two pixels. In a real-world scenario, good radiometric conditions are often disrupted by complicated and dynamic lighting sources, inappropriate camera configuration, and non-Lambertian reflectance of objects. We investigate the interdependencies among matching performance, cost functions, and observation conditions using both close-range and remote-sensing data. Our cost-merging strategy combines the advantages of different match cost functions and gives consideration to imagery configurations. In addition, a novel probabilistic surface prior is introduced incorporating a new energy optimization method, called *iSGM3*. Our approach builds a probabilistic surface prior over the disparity space using confidences on a set of reliably matched correspondences. Unlike many region-based methods, our method defines an energy formulation over pixels, instead of regions in a segmentation; this results in a decreased sensitivity to the quality of the initial segmentation. This dissertation suggests the way to developing robust stereo methods is on the level of obtaining costs and suitable energy formulation, and not only the energy optimization. Both costs merging and the surface prior are generally applicable for almost all extended stereo methods.

Contents

Nomenclature

$\mathcal{I}$	An image.
$\mathcal{I}_s, \mathcal{I}_m$	Source (reference) and match image[1].
$\mathcal{O}$	Observation of a stereo pair, $\mathcal{I}_s$ and $\mathcal{I}_m$.
$p \in \mathcal{I}$	Pixel p of image $\mathcal{I}$.
$\mathcal{I}^{3\to1}$	An intensity image with merged color channels.
$\bigtriangledown_e \mathcal{I}$	A gradient image derived in epiplar direction.
Seg	Segmentation of a image.
δ_p	Disparity at pixel p in a disparity map.
d_p	Disparity at pixel p in a plane map.
D	Disparity range.
Δ	Disparity map.
Δ^{pl}	Plane-fit disparity map.
Π^m	A disparity plane with index m.
$\mathrm{P}(X)$	The prior probability of X.
$\mathrm{P}(X\|Y)$	The conditional probability of X, given Y.
$\mathrm{P}(X,Y)$	The joint probability of X and Y.
α, β	Scalar value.
ω	Weight factor.
λ, κ	Penalty factor.

[1] Without loss of generality, the left image is considered as the reference image. $\mathcal{I}_s = \mathcal{I}_L$ and $\mathcal{I}_m = \mathcal{I}_R$.

Chapter 1

Introduction

The goal of dense stereo matching is to estimate the distance, or depth, to an imaged object in every pixel of an input image; this is achieved by finding pixel correspondences between the source image and one or more matching images. The displacement between corresponding pixels is then used to calculate depth, along with the relative position of the two cameras. Early stereo algorithms used to extract features in different views, such as edges or corners, and identify the similarities with features in the input images. In contrast, modern dense stereo algorithms reconstruct scenes even with little texture where no features can be extracted.

Although it seems we can infer depth with our own eyes effortlessly, human depth perception is not able to estimate the distance of objects metrically. Computational dense stereo methods perform the task by assigning a concrete depth to each pixel when enough parallax is present. Depth information is useful to understand scenes, and it enables higher-level image processing and information extraction. However, several factors make computational dense stereo hard in practice. Neighboring pixels may have the same intensity/color/local structure, leading to matching ambiguities. Changing observation views causes occlusions that block part of one image from being seen in the other, and no depth can be estimated for these areas.

Broadly, dense stereo algorithms can be categorized into two methodologies based on the way the problem is solved: local and global algorithms [Scharstein and Szeliski, 2002]. The most fundamental component of every stereo algorithm is the match cost [Hirschmüller and Scharstein, 2009]. All stereo methods include a cost function to measure the fitness of possible correspondences in a stereo pair. The match cost is an observation directly calculated from input images, and also called data term in an energy formulation. Local algorithms independently estimate correspondences for each pixel separately, without considering neighboring correspondences. In contrast, global methods consider the neighborhood of each pixel using a regularization term, such as a smoothness assumption, or its belonging to a segment. Such an additional term adds prior knowledge and expectations to a depth map – for example, smooth surfaces, sharp edges, and scene visibility [Bleyer and Gelautz, 2005; Kolmogorov and Zabih, 2001; Szeliski et al., 2008]. In general, stereo matching is an ill-posed problem, so some regularization is required to achieve a meaningful solution. Moreover, the assumptions are often contradictory. In addition to smooth surfaces within objects, large discontinuities should remain at object boundaries. The problem is represented in a disparity map that leads to a global extreme of an energy function consisting of data and regularization terms defined over the whole image. This generally leads to computationally NP-hard [1] problems [Boykov et al., 2001; Kolmogorov and Zabih, 2001; Neilson and Yang, 2008]. Different optimization methods are used to find the (approximated) extremum.

[1] **N**on-deterministic **P**olynomial-time hard

In summary, developing a computational stereo method considers three main factors as shown in Figure 1.1: a) gaining reliable observation to measure matching similarity (cost computation); b) modelling goal function for the unknown solution with some prior assumptions (energy-function formulation); c) achieving/approximating the global minimum (energy optimization).

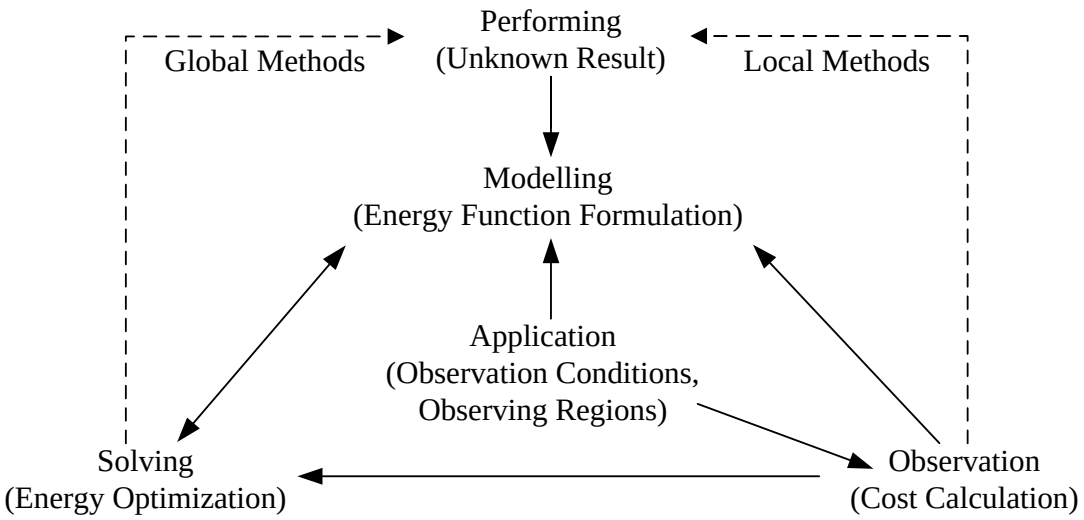


Figure 1.1: **Overview of the computational dense stereopsis.** The three main factors for developing a dense stereo method include obtaining reliable match costs, formulating solvable energy function, and achieving/approximating the global extreme.

Applications that make the dense stereo active come from areas such as photogrammetry, mobile robotics, intelligent vehicles and remote sensing. For instance, many remote sensing applications require depth maps to generate Digital Elevation Models (DEMs) from airborne and satellite imagery. DEMs are a fundamental dataset required for applications such as mobile phone network planning, flood prediction, and 3D city modeling and analysis [Jin et al., 2010; Ju et al., 2009; Leberl et al., 2010; Zhang et al., 2011]. LiDAR [1] is often used for 3D reconstruction both in close range and remote-sensing applications, but advanced dense stereopsis can be used in place of LiDAR, resulting in models with higher resolution, improved application flexibility and lower operating cost [Kurz et al., 2012; Reinartz et al., 2010]. One application for fast, airborne 3D reconstruction is damage evaluation and rescue support in the case of natural disasters, such as land slides, as shown in Figure 1.2.

Unfortunately, the most recent stereo matching methods are developed using data sets with relatively good radiometric configurations and small base lines. The observed scenes typically contain simple geometric figures. In contrast, real-world applications are more complex and difficult than the widely-used benchmarks due to challenges in observation conditions and in the features of the imaged objects. In remote-sensing applications, urban areas include high buildings, slanted roofs and large homogenous regions. The light source is complex and dynamic – the sun instead of artificial ambient light sources. Most surfaces do not show a Lambertian-reflectance behavior, which leads to various image intensities of the same object point from different camera viewpoints. Moving cars, shadows of buildings and changing weather make data dynamic, even over a short period. Despite intense investigations by the research community in the last decades, these challenges are not resolved and open problems remain.

This work introduces two methodological novelties which aim to be generally applicable

[1] **Li**ght **D**etection **A**nd **R**anging

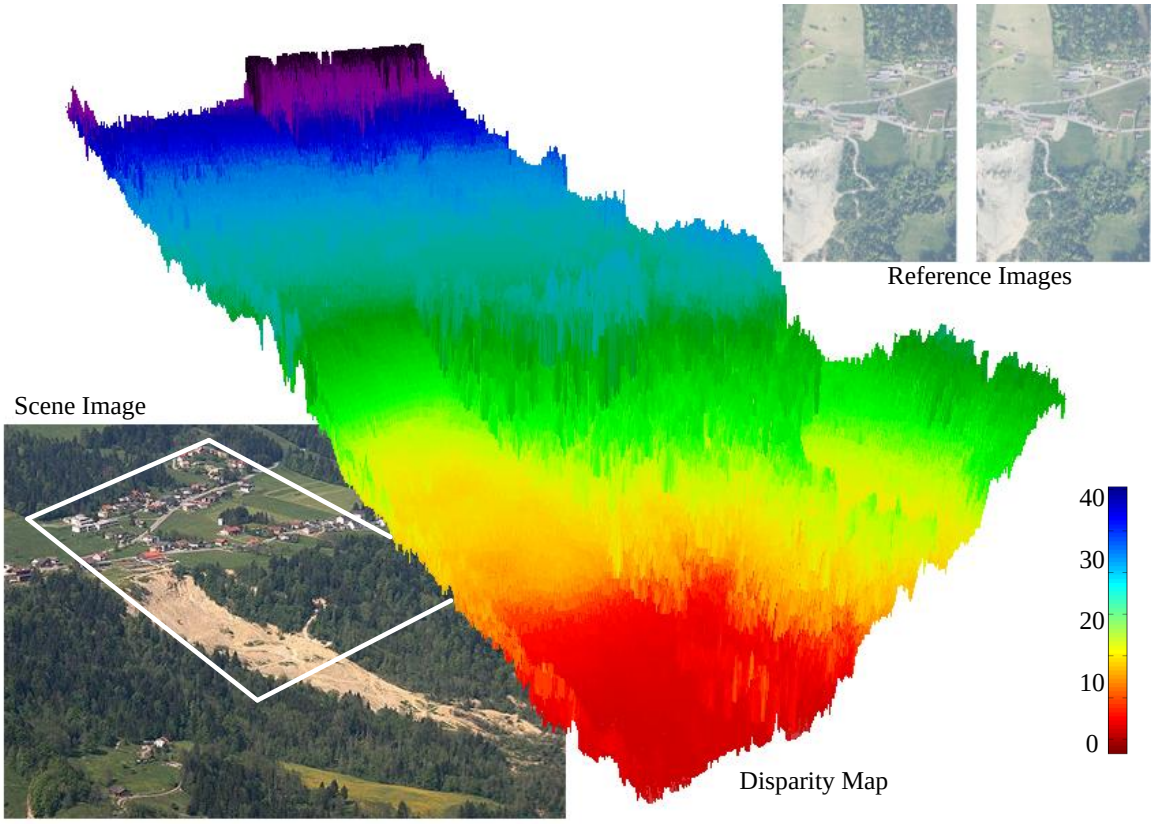


Figure 1.2: **Airborne 3D Reconstruction after landslide in Doren, Austria. Top-right**: Airborne stereo pair taken by DLR[a] 3K camera system using Canon EOS 1D Mark II cameras. **Center**: Disparity map generated by **Top-right** and colored with a heat scale such that colder (glue/violet) colors are closer to the camera than hotter (red/yellow) colors. **Bottom-left**: a scene image [b] from the similar view point of the visualization. Reconstructed region is assigned by a white frame.

[a] German Aerospace Center
[b] Image source: Gemeindedaten von Doren [www.statistik.at]

for almost all global optimization methods. 1) Based on the performance study of match cost functions using both standard computer vision benchmarks and remote-sensing data, a merging strategy is introduced to design robust match costs. 2) A novel confidence-based surface prior is developed within a probabilistic framework. Depending on the reliability of the prior (the confidence) from a previous matching, the introduced surface prior is modeled as a Gaussian distribution, which can be probabilistically fused with the current approach. Moreover, we introduce a new optimization framework in the energy space for iteratively updating/using the confidence. Three data sets are used in this dissertation for benchmarking and evaluation of the developed methods: the Middlebury benchmark with radiometric changes, airborne image sequences with an increasing baseline and satellite stereo pairs with large stereo angles.

This dissertation is structured as follows. In Chapter 2 we describe the dense stereo problem and existing methods for solving it. In addition, the influence and challenge of applied data on matching performance are briefly discussed. The two main contributions of this dissertation are then methodically formulated. In Chapter 3, we introduce different match cost functions

and the merging strategy to obtain robust costs. Chapter 4 presents the novel confidence-based surface prior and its relaxation in the global energy formulation. In Chapter 5, according to the evaluation of different cost functions, we investigate the interdependencies among match cost, matching performance, applied region and observation constraints. Moreover, we compare the results with and without the proposed surface prior under a global energy formulation and demonstrate that our approach has superior performance on all data sets used. Finally, we conclude our work in Chapter 6.

1.1 Challenging Real-world Data

There are hundreds approaches seeking to solve the stereopsis. A number of factors that make it quite challenging become apparent once one begins to use it in real-world applications like mobile robotics and remote sensing. Non-Lambertian reflectance, complex scene, and radiometric changes, among other factors, are usually present in real-world data. This leads to low performance of stereo algorithms that perform well on some de-facto standard benchmarks in the computer vision community. With respect to real-world data, we have summarized the challenges for stereo matching methods as follows:

Complicated radiometric conditions. Lighting sources in the real world including the sun and man-made sources, are complicated and dynamic. Even for close-up images, radiometric changes can appear – for example, a cloud is covering the sun, the headlight of a car is turned on, and so on. In contrast, dense stereo algorithms are typically developed and evaluated using data with a small baseline configuration as well as artificial and often ambient light sources. Radiometric changes due to, for example, vignetting and gamma changes, and so forth, are often simulated by modifying small baseline images [Hirschmüller and Scharstein, 2009; Scharstein and Szeliski, 2002]; but, these simulations do not capture all effects such as non-Lambertian reflectance and changing observation views.

View-dependent effects. Large stereo angles and baseline lengths cause *occlusions* that prevent the visibility of objects. For remote sensing data, regions under building shadows are often underexposed. Pixelwise match costs in occlusion and shadow areas are quite random. Using global algorithms, objects can be dilated into such weakly-matching regions due to over-smoothing (foreground fattening problem). Moreover, local structures on object boundaries can be changed by observations from different view points. Encoding local structures is limited by the fronto-parallel sampling mechanism of window-based methods. A result is that instead of slanted surfaces in the reality, fronto-parallel surfaces are preferred in most global methods.

Sensitivity to the parameter configuration. The matching performance of a stereo method is always dependent on the parameters used. For most global methods, changing the strength of the smoothness term leads to different results. Using a small smoothing constraint, the result tends to perform sharp edges, but many outliers may appear on smooth surfaces and planes. A large smoothing constraint can reduce mismatching on continual regions, but at the cost of over-smoothed object boundaries. Moreover, small discontinuities are often over-smoothed during energy propagation such that small-scale details on objects are often erased in a reconstructed scene. One of the most important reasons for the only slight performance differences of many state-of-the-art methods is that parameter configuration is often tuned according to ground truth. However, the sensitivity to the

parameter configuration is quite individual for each stereo method. The performances of these benchmarking results on real-world data are difficult to extrapolate.

Absent semantic information. In texture-less areas, parametric match costs are unreliable. Nonparametric, mostly window-based, match costs such as census transformation can only partially overcome this problem and lead to dilated edges of object boundaries. Global methods using smoothness terms are not able to handle large homogeneous regions due to absent semantic information at the object level. The thought of dealing with such regions as inseparable segments is very sensitive to a given segmentation, which is mostly extracted based on the color consistency within a two dimensional image.

Figure 1.3 demonstrates a reconstructed urban scene using an airborne stereo pair with $15cm$ ground resolution. One of the input image contains 15 million pixels. The 3D scene includes building roofs (slanted surfaces), high buildings (large discontinuities on boundaries), regions with less textures (homogeneity) and narrow streets under shadows (weakly matching areas). Dense stereo matching becomes more challenging for such data in contrast to the standard benchmarks.

Figure 1.3: **Reconstructed urban scene in Munich, Germany.** High buildings, moving passengers, shadows, and texture-less regions are included.

1.2 Contributions

The performance of a dense stereo matching method depends on all components including energy function formulation, match costs, energy minimization, disparity optimization, and post-processing. These components are interdependent upon each other. In this dissertation we focus on the technical aspects of these components and contribute two main methodical novelties. The *merging strategy of match costs* combines advantages of different match cost functions and gives

consideration to imagery configurations. In addition, a novel *probabilistic surface prior* is introduced incorporating a new energy optimization method, called iSGM3 [1]. Both novelties are generally applicable for almost all extended stereo methods and only limited adaptation is necessary. To evaluate our results, we apply not only the de-facto standard benchmarks from the computer vision community, but also remote-sensing data. The challenges as well as the influences on matching performance of the the real-world data are discussed during our evaluation. An overview of our contributions are listed as follows:

1. investigation on match cost functions in context of imagery conditions;
2. comparison between the de-facto standard benchmarks and remote sensing data;
3. match-costs merging according to data in order to develop robust match observations; and
4. the confidence-based surface prior incorporating with
5. a modified semi-global optimization framework to minimize energy function with the additional surface prior.

All dense stereo matching algorithms use match cost functions to measure the similarity between two pixels. In a real-world scenario, good radiometric conditions are prevented by complicated and dynamic lighting sources, inappropriate camera configuration, and non-Lambertian reflectance of objects. We study the interdependencies among matching performance, cost functions, and observation conditions using both close-range and remote-sensing data. Three typical match costs including a parametric cost (absolute difference), a non-parametric cost (census transformation) and the mutual information are investigated. We characterize the performing variations of the cost functions with respect to the image features such as homogeneity and discontinuity. The investigation indicates that non-parametric match costs perform well on real-world data but result in dilation of object boundaries. Based on this study, a merging strategy of different costs is introduced to obtain reliable match costs. The performance study on cost functions can be guided by researcher and developers for robust real-world applications.

We present a novel formulation for pixel-wise surface stereo matching. Our approach builds a probabilistic surface prior over the disparity space using confidences on a set of reliably matched correspondences. We minimize the proposed energy formulation using a modified semi-global optimization framework. Since the confidences are derived with respect to image matching likelihood and smooth prior probability, our approach is less sensitive to an initial image segmentation in comparison to existing segment-based stereo methods. Our method is iterative. Given a dense disparity estimation we fit planes, in disparity space, to regions of the image. We then recalculate a new disparity estimation with the addition of our novel confidence-based surface prior that constrains disparities to lie near the fit planes. The process is then repeated. Unlike many region-based methods, our method defines an energy formulation over pixels, instead of regions in a segmentation; this results in a decreased sensitivity to the quality of the initial segmentation. Our surface prior differs from existing soft surface constraints in that it varies the per-pixel strength of the constraint to be proportional to the confidence in our given disparity estimation. The addition of our surface prior has three main benefits: sharp object-boundary edges in areas of depth discontinuity; accurate disparity in surface regions; and low sensitivity to segmentation. Our results demonstrate that our approach has superior performance on all data sets used.

Our results include evaluations using data sets with different properties. The benchmarks from the computer vision community contain close-up data with ambient light sources and small

[1] **i**terative **S**emi-**G**lobal **M**atching with **3** terms

base lines. The observed scenes typically contain simple geometric shapes. Their ground truths are highly precise and have the same resolution as the input images. The evaluation using these benchmarks provides us with an accurate analysis. Compared with remote sensing data, we also show the limitations of the standard benchmarks for developing stereo methods. A continually recorded airborne image sequence provides stereo pairs observing the same location with increasing baseline length. This allows us to study the stereo matching performance depending on the baseline of the stereo pairs. The airborne images cover urban areas including shadows, high buildings, large homogenous roofs and small streets. Moreover, in contrast to Middlebury data, the satellite stereo pairs with very large stereo angles show the challenges in many real-world applications, especially when considering remote sensing. Our comparison and discussion demonstrate the key for developing robust stereo methods is on the level of obtaining costs and suitable energy formulation, and not only on the way of energy optimization. Our contributions, merging costs and confidence-based surface prior, can be used by almost all stereo methods.

Chapter 2

Dense Stereo Matching

Depth information is lost when a 3D scene is optically projected onto an image plane. Passive stereo vision aims to obtain distances of objects seen by two or more cameras from different viewpoints. The parallax between different view positions causes a relative displacement of corresponding features in the input images. The relative displacement, called *disparity*, encodes the depth information lost during projection onto the image plane. Given a binocular stereo pair, *dense stereo matching* assigns for each pixel in one of the input images a disparity to indicate its corresponding pixel in another input image. The assigned disparity can be used to calculate metric depth, if the relative orientation of the two images is known. The key problem, which is difficult to solve and computationally expensive, is to find the correspondences between almost all pixels in two images. Dense stereo matching, which leads to a depth map containing every pixel, is generally used for 3D reconstruction. This differs from common feature-based methods, which can only generate sparse depth maps [Haralick and Shapiro, 1992]. Combinations of dense stereo matching with feature-based matching have been developed; they mostly use the feature-based matching to stabilize or initialize the dense matching [Lowe, 2004; Sadeghi et al., 2008; Taylor, 2003].

Stereo pairs are mostly rectified before matching. The purpose of image rectification is to limit the computation of stereo correspondences in one dimension according to epipolar geometry, which is the intrinsic projective geometry between two views and is obtained using the intrinsic and extrinsic parameters of the two cameras. Given a calibrated stereo pair, in which the camera parameters, relative position and orientation of two cameras are known – rectification transforms the two image planes such that the conjugate epipolar lines become collinear and parallel to one of the image axes [Fusiello et al., 2000]. Once the correspondences are found, the 3D scene can be reconstructed using triangulation.

The performance of a dense stereo matching method depends on all components, which can be generally presented in four steps [Scharstein, 1999; Scharstein and Szeliski, 1998, 2002]:

1. Match cost computation;
2. Spatial aggregation of match costs;
3. Disparity calculation with or without optimization; and
4. Disparity map refinement.

A match cost refers to the similarity of image locations and can be calculated if the displacement of two corresponding pixels is given. The per-pixel match cost can be spatially aggregated

over a support region in order to reduce imagery outliers from sensors. In the disparity calculation phase (step 3), optimization methods can be used to achieve a solution for expected forms of the result. Here, most methods can be classified as *local* or *global*, depending on how to solve the problem. Local methods tend to use only the aggregated costs from step 1 and step 2 and to select the disparities locally. Global methods typically make assumptions about the smoothness of the disparity map and consider the disparity selection within a neighboring context. Most global methods are formulated under the energy function frameworks based on Markov Random Fields and are solved using different optimization methods to find the global minimum [Lempitsky et al., 2007; Szeliski et al., 2008]. Finally, a calculated disparity map can be refined by removing some outliers and filling a small number of quantized disparities in step 4.

Computational stereopsis is one of the most active research topics in computer vision. Hundreds, if not thousands, of different algorithms have been developed in the past decades. In this chapter, we discuss the computational stereopsis problem and different methods for solving it. In section 2.1 the triangulation between object depth and pixel disparity is described. An overview of match cost functions is given in section 2.2. Section 2.3 introduces local stereo algorithms. In section 2.4 we derive the energy function formulation for stereopsis from the probabilistic formulation. In addition, a few of global optimization methods is presented.

2.1 Binocular Reconstruction

Given a binocular pair of stereo images, $\mathcal{I}_s$ and $\mathcal{I}_m$, the goal of stereopsis is to infer the distance from the camera to the 3D objects visible in these images. However, rather than inferring the depth, d_p of pixel p , computational stereo algorithms typically calculate the disparity, δ_p, at each *visible pixel* which can be observed both in $\mathcal{I}_s$ and $\mathcal{I}_m$.

As shown in Figure 2.1, the spatial point P is captured as p and p' in the source and match images respectively. The camera centers C_s and C_m are co-planar with P and its projections p and p'. The line connecting two camera centers is referred to the baseline, and we denote its length in pixel units as b. Assuming that $\mathcal{I}_s$ and $\mathcal{I}_m$ are rectified, thus all potential correspondences lie on the same epipolar line. Correspondence searching is then limited to one dimension – the epipolar line is identical to one of the image axes [Fusiello et al., 2000]. We define *disparity* as the displacement between pixel $p(x, y) \in \mathcal{I}_s$ and its correspondence $p'(x', y') \in \mathcal{I}_m$ along the x-axis of the rectified image coordinate system as $\delta_p = |x - x'|$ [1] and $y = y'$. Throughout this dissertation we use this definition of disparity. Disparity is inversely proportional to object depth, shown as the triangulation in Figure 2.1:

$$d_p = f\frac{b}{\delta_p} \tag{2.1}$$

where f represents the focal length of the cameras, which are assumed identical.

Thus, if the correspondence is known, we can use the coordinates of two matched pixels to calculate the displacement δ, further to estimate the depth of a 3D scene point. If all correspondences, or at least most, are found, we refer to the result containing the displacements with image locations as a *disparity map*. The process of finding a disparity map for a binocular stereo pair is referred as *dense stereo matching*.

[1] The disparity values of a stereo pair can either ≥ 0 or ≤ 0. In this dissertation, we assume $\delta_p \geq 0$.

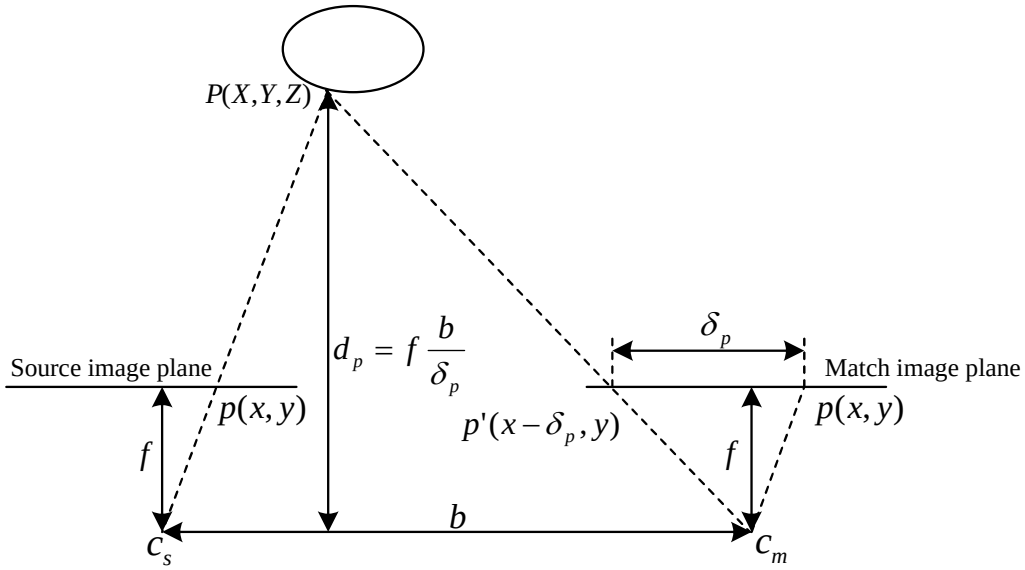


Figure 2.1: Binocular triangulation of a rectified stereo pair.

2.2 Match Costs

Match cost is the most fundamental component for correspondence computations. All stereo methods use a cost function to compute the similarity of possible correspondences, but might combine the cost function with different aggregation, optimization, and post-processing approaches. There are many ways to define a match cost function. One can use methods ranging from the simple Euclidean distance metric to more complex spatial aggregation. In this section we describe match costs briefly. A detailed description of the cost functions investigated in this dissertation is presented in Chapter 3.

2.2.1 Match Cost Functions

Cost function measures the fitness of assigning a disparity value, δ, to a pixel $p(x, y)$, where (x, y) is the coordinate of p defined in the source image, $\mathcal{I}_s$. The cost, $C(p, \delta)$ is calculated by warping p at $p' = (x + \delta, y)$ in the match image, $\mathcal{I}_m$, and comparing the observations between $\mathcal{I}_s(p)$ and $\mathcal{I}_m(p')$[1]. Within a one dimensional integral searching range, $|D| = |\delta_{\max} - \delta_{\min}|$ match costs can be calculated for each p in $\mathcal{I}_s$. Thus, we can generate a three dimensional cost cube or so-called *disparity space image* (DSI) for $\mathcal{I}_s$. Each element in the cube, $\mathrm{DSI}_s(p, \delta)$, stores the quantifiable information for two corresponding pixels calculated using a match cost function.

The simplest and most intuitive cost functions assume constant intensities or colors of corresponding pixels. However assumptions based on such radiometric consistencies can only work under ideal imagery conditions in texture-rich areas. The difficulty to measure a robust match cost arises from two independent factors according to Šára [2002]:

1. *Data uncertainty due to insufficient signal-to-noise ratio in images or mis-calibrated cameras of weakly textured objects;* and

2. *Structural ambiguity due to the presence of periodic structures combined with the inability of a small number of cameras to capture the entire light field.*

[1] Throughout this dissertation we assume, $\mathcal{I}_s$ and $\mathcal{I}_m$ are rectified.

Furthermore, we would add two factors that make match costs particularly difficult for challenging real-world data.

1. Non-Lambertian bi-directional reflectance which causes some corresponding pixels to have drastically different colors in different images; and

2. Occlusion of scene components due to differing viewpoints such that assigning depth to such pixels must rely on mechanisms other than the match cost function self; such as the smoothness assumptions made by global stereopsis algorithms.

There are different ways to define a function to calculate a robust matching cost. One pixel-wise way is to statistically model the radiometric changes of corresponding intensities/colors. Another local way is to consider the neighboring observations within a region, using for instance, Euclidean distance metric. In subsection 2.2.2, we focus on the spatially aggregated match cost functions, which can be optionally used. The formula definitions of cost functions investigated in this dissertation are introduced in Chapter 3 with our cost-merging strategy.

2.2.2 Spatial Cost Aggregation

Spatial aggregation is a computationally efficient way to improve match costs and is often applied by local methods, like the winner-takes-all disparity selection for real-time applications [Gong et al., 2007; Zhu et al., 2010]. We summarize different aggregation strategies into two categories according to the spatial and weight definitions. Both approaches can be used separately as well as in combination.

We define a local region, $\tilde{W}$, which can be a square window or a more generic shape of a neighborhood. $\tilde{W}$ is centered at p of $\mathcal{I}_s$ and p' of $\mathcal{I}_m$ given a disparity δ_p. We define a function $\mathcal{O}(p)$ to observe the local features at pixel p – intensity, color, an so forth. Additionally, we denote a function, C to measure the cost between the individual observations, $\mathcal{O}_s(p)$ and $\mathcal{O}_m(p')$ with respect to a given disparity of δ_p. The function $Z^{\tilde{W}}$ is used to normalize the cost. We define a general function for spatial cost aggregation over a region, $C^{\tilde{W}}$, as following:

$$C^{\tilde{W}} = Z^{\tilde{W}} \sum_{p,p' \in \tilde{W}, \delta(p) \in D} w^p C(\mathcal{O}_s(p), \mathcal{O}_m(p'), \delta_p) \tag{2.2}$$

where w^p is an optional weight function with respect to the distance of pixels from the local center. This model is based on the assumption that the whole window $\tilde{W}$ at position p can be assigned a single disparity value δ_p.

Spatial definitions. The basic spatial aggregation is defined within a square window and achieved by a box filter. Bobick and Intille [1999] introduce a shiftable window approach using a separable sliding min-filter. A rather complicated approach is the oriented-rod aggregation, which classifies pixels into heterogeneous groups and applies a shiftable-window filter to homogeneous pixels [Kim et al., 2005]. In contrast to cost aggregations using rectangular windows, the adaptive approaches considers shapes of the input images [Kanade and Okutomi, 1994a]. The quality of aggregated match costs relies on the definition of the support region. Typically color segmentation and gradients are used to obtain an adaptive region [Gong and Yang, 2005b].

Weight definitions. The spatial costs in a support region can be typically weighted with respect to their spatial Euclidian distance from the local center, which is normally located at the pixel to be aggregated. Yoon and Kweon [2005] introduce an adaptive-weight approach,

which computes the weighted average of adjacent match costs with the weights generated using both input images.

However, according to a great number of different approaches in the literature, spatial cost aggregations are typically incapable of computing accurate disparity maps. The size of the patch determines the resolution of the output disparity. Often, neighboring pixels are assigned with the same disparity value. While a small window results a noisy disparity map, a large window might over-smooth the result. Also weighting approaches can generally not handle large discontinuities at object boundaries. Therefore a fixed window size suited for the whole input image is difficult to choose. Another limitation of cost aggregations is that fronto-parallel planes are generally preferred, because a constant disparity for all pixels within the support region is assumed. Moreover, the visibility of local structures can be prevented by large parallax. Adaptive strategies can improve matching performances, only if support regions are found correctly. Thus, whatever the kind of cost aggregation, their performances rely on the windows or regions used.

2.3 Local Stereo Algorithms

Local stereo algorithms independently compute the disparity for each pixel; typically the match costs are computed using window-based cost aggregations. Fixed window size leads to either over-smoothing or noisy results. However, while adaptive windows can improve matching performance, poorly-textured surfaces cannot be matched consistently. The main strength of local methods is low computational cost, in terms of memory and time.

There are many local methods, and most of them differ in the way they aggregate costs. In this section we introduce only a very limited sampling of them.

2.3.1 The Winner-takes-all Algorithm

The most intuitive and widely used local method is the *winner-takes-all* (WTA) algorithm, which is introduced very early in the computer vision area [Pollard et al., 1985; Rosenfeld et al., 1976; Zucker et al., 1981]. As its name suggests, WTA assigns pixels with a disparity level, which has the lowest cost calculated from a match cost function. More formulary, given match costs from a function, $C(p, \delta)$ for image $\mathcal{I}_s$, WTA finds a disparity map, Δ as

$$\Delta = \arg\min_{\delta}(C(p, \delta)), \ \forall p \in \mathcal{I}_s \tag{2.3}$$

where δ refers to a disparity value within the searching range.

The results of WTA strongly rely on the cost function used. In areas with poor to no texture, the choice of WTA is effectively random due to highly ambiguous observations. Despite these problems, the WTA algorithm is used in most local stereo algorithms, and many global methods use WTA to compute an initial solution. The main strength of the WTA algorithm is its speed; on account of its simple implementation, disparity maps can be computed quickly even for large images with a big disparity search range. In some global frameworks, such as Semi Global Matching (SGM), WTA is executed after optimization to select the disparities.

2.3.2 Other Algorithms

Li [1990] proposes the *loser-takes-nothing* method that calculates a disparity map through iterative eliminations of matches with high costs. The process suppresses the most likely candidate

at each iteration until only one candidate is left to achieve an unambiguous set for both images. This method executes as a slowest-descent approach, and is computationally very inefficient.

Zhang et al. [1995] introduce the *some-winners-take-all* method by updating the matching to retain a symmetric one-to-one matching between two images. They measure the matching strength between points using a correlation operator and sort all strengths in a decreasing order. In addition, a table to describe the ambigity/non-unambigity between candidate matches is created in decreasing order. The potential matches, which are among both the first percentage of matches in both tables, are selected as correct matches. Thus, the ambiguous potential matches are not selected even where there is a high matching strength, and the unambiguous matches with low strength are also rejected.

2.4 Global Stereo Algorithms

Recently, a great number of stereo methods are framed as global optimization problems. As many modern methods in the computer vision area, their origin goes back to the work of German and German [1984], who first incorporated a Bayesian interpretation of energy functions based on Markov Random Field (MRF) for image restoration. Li [1994] has introduced a more unified approach for MRF modeling in low- and high-level vision problems. In this section we present the general Bayesian approach based on MRF for the stereopsis problem and derive the energy function formulation from the probabilistic formulation according to Li [1994].

2.4.1 MAP-MRF model for Stereo Matching

Global stereo models based on Markov Random Field (MRF) are formulated under the Bayesian framework. The optimal solution for this formulation is defined as the maximum *a posteriori* (MAP) probability estimation. The MAP-MRF framework enables developing vision algorithms using sound principles rather than *ad hoc* heuristics [Li et al., 1997]:

- The posterior probability can be derived from a prior probability and a likelihood model using Bayes' rule.
- Contextual constraints can be expressed as prior probability under the Markovian Property.

Bayesian MAP. Depth estimation is one of the labeling problems in computer vision. Given a stereo pair $(\mathcal{I}_s, \mathcal{I}_m)$ as the observation $\mathcal{O}$, we define a finite set of sites referring to $W \times H$ pixels on a regular 2D-grid[1]:

$$S = \{p \mid p = (x, y), x \in [1, W], y \in [1, H]\} \tag{2.4}$$

where p is the coordinate defined in an image. Let D be a set of disparity labels. A random variable assigns a disparity level δ_p to site p, $\Delta(p) = \delta_p$ with $\delta_p \in D$. Thus, a resolution Δ for $\mathcal{I}_s$ is a labeling configuration over S: $\Delta \in D^{W \times H} = \underbrace{D \times D \ldots \times D}_{W \times H}$. Our target is to calculate a disparity map Δ for $\mathcal{I}_s$, such that the probability of Δ maximizes the posterior probability $\mathrm{P}(\Delta \mid \mathcal{O})$:

$$\Delta = \arg \max_{\Delta \in D^{W \times H}} \mathrm{P}(\Delta \mid \mathcal{O}) \tag{2.5}$$

[1] We assume the reference and matching images have the same size.

Using Bayes' rule the posterior can be written as:

$$\mathrm{P}(\Delta \mid \mathcal{O}) = \frac{\mathrm{P}(\mathcal{O} \mid \Delta)\mathrm{P}(\Delta)}{\mathrm{P}(\mathcal{O})} \tag{2.6}$$

Given $\mathcal{I}_s$ and $\mathcal{I}_m$, observation $\mathcal{O}$ is a constant. The posterior can be derived as product of the likelihood probability, $\mathrm{P}(\mathcal{O})$ and the prior model, $\mathrm{P}(O \mid \Delta)$:

$$\mathrm{P}(\Delta \mid \mathcal{O}) \propto \mathrm{P}(\mathcal{O} \mid \Delta)\mathrm{P}(\Delta) \tag{2.7}$$

The likelihood $\mathrm{P}(\mathcal{O} \mid \Delta)$ of the observation $\mathcal{O}$ expresses how probable the observed data is for a labeling configuration over $\mathcal{O}$ under the condition of Δ. The *a priori* joint probability $\mathrm{P}(\Delta)$ expresses the contextual constraint inferred from truth, which is in general difficult to know. That is the reason using MRF modeling – in order to specify the *a priori* probability.

MRF Prior. We define a neighborhood system for S, $\mathcal{N} = \{\mathcal{N}_p \mid p \in S\}$. $\mathcal{N}$ contains all neighbors of p and satisfies (1) $p \notin \mathcal{N}_p$ and (2) $p \in \mathcal{N}_q \iff q \in \mathcal{N}_p$, where q is a neighbor pixel of p. We use Δ to denote a family of random variables with $\Delta = \{\Delta(p) \mid p \in S\}$. Δ is a Markov Random Field over S with respect to the neighborhood system $\mathcal{N}$ if and only if the two conditions are satisfied:

(i) Positivity: $\mathrm{P}(\Delta(p) = \delta) > 0, \forall p \in S$

(ii) **Markovianity**: $\mathrm{P}(\Delta(p) \mid \Delta(q), q \in S, p \neq q) = \mathrm{P}(\Delta(p) \mid \Delta(q), q \in \mathcal{N}_p)$

The first condition deems Δ to be a random field. The second condition states the local interaction that the probability of an assignment at p is conditioned only on the results of its neighborhood.

Gibbs-Markov Equivalences. According to the Hammersley-Clifford theorem [Besag, 1974], Δ is a MRF of S with respect to $\mathcal{N}$ if, and only if the joint probability $\mathrm{P}(\Delta)$ is a Gibbs distribution:

$$\mathrm{P}(\Delta) = Z^{-1} \times e^{-\frac{1}{T}U(\Delta)} \tag{2.8}$$

where Z serves as a normalization constant and T is a control parameter. Furthermore, a clique is defined by either a single neighbor or a set of neighbors. We denote a clique c for a graph $(S, \mathcal{N})$ as a subset of S with $c \in \mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k, \ldots\}$. The ensemble of cliques is relative to $\mathcal{N}$ depending on the neighborhood definition and a clique, $c = \mathcal{C}_k$, consists of k neighbors. $U(\Delta)$ is defined as the prior energy of a configuration Δ over the whole sites S. The global energy can be written as the sum over all potentials [Besag, 1974]:

$$U(\Delta) = \sum_{c \in \mathcal{C}} V_c(\Delta) = \sum_{\{p\} \in \mathcal{C}_1} V_1(\Delta(p)) + \sum_{\{p,q\} \in \mathcal{C}_2} V_2(\Delta(p), \Delta(q)) + \ldots \tag{2.9}$$

As a result, we express the local interactions between sites (the joint prior probabilities) using a set of potential functions V_c under a corresponding clique system $\mathcal{C}$.

Posterior Energy. We assume that the likelihood function is in an exponential form:

$$\mathrm{P}(\mathcal{O} \mid \Delta) = Z_{\mathcal{O}}^{-1} \times e^{-U(\mathcal{O}|\Delta)} \tag{2.10}$$

From 2.8 and 2.10 the posterior probability is also a Gibbs distribution, which has the following form:

$$\mathrm{P}(\Delta \mid \mathcal{O}) = Z_E^{-1} \times e^{-U(\Delta|\mathcal{O})} \tag{2.11}$$

Then the posterior probability can be restated now as a posterior energy.

$$E(\Delta) = U(\Delta \mid \mathcal{O}) = U(\mathcal{O} \mid \Delta) + \frac{1}{T}U(\Delta) \tag{2.12}$$

The MAP-MRF is now derived as an energy minimization problem for 2.12 that

$$\hat{\Delta} = \arg \min_{\Delta \in D^{W \times H}} U(\Delta \mid \mathcal{O}) \tag{2.13}$$

with

$$\mathrm{P}(\Delta \mid \mathcal{O}) \propto e^{-\sum_{p \in S} V_p(O(p)|\Delta(p)) - \sum_{(p,q) \in \mathcal{N}} V_{p,q}(\Delta(p),\Delta(q))} \tag{2.14}$$

where the functions V_p and $V_{p,q}$ are defined in Equation 2.9.

2.4.2 Energy Function Formulations

We summary that the energy function formulations are based on MRFs. In this subsection we discuss only the global energy formulations including a basic smoothness prior. The energy functions can be defined over pixels as well as segments.

Energy Function Formulation over Pixels. Given a stereo image pair, $\{\mathcal{I}_s, \mathcal{I}_m\}$, a disparity map, $\Delta : \mathcal{I}_s \to \mathbb{Z}^{>0}$ for $\mathcal{I}_s$, can be expressed as the function that minimizes the energy equation:

$$E(\Delta) = \sum_{p \in \mathcal{I}_s} C(p, \Delta(p)) + \lambda \sum_{\{p,q\} \in \mathcal{N}} V_{p,q}(\Delta(p), \Delta(q)) \tag{2.15}$$

where p and q denote pixels in $\mathcal{I}_s$, $\mathcal{N}$ is the set of neighboring pixel pairs in $\mathcal{I}_s$, $C : \mathcal{I}_s \times \mathbb{Z}^{>0} \to \mathbb{R}$ is a match cost function that provides a measure of fitness for assigning disparity values to pixels, and $V_{p,q}(\delta_1, \delta_2) : (\mathbb{Z}^{>0})^2 \to \mathbb{R}$ is a smoothing term that encodes a smoothness constraint on the resulting disparity map. Finding optimal minima of these equations is generally NP-hard; thus, algorithms that utilize this framework find approximations.

Energy Function Formulation over Segments. There are global formulations defined over labels, instead of over pixels [Bleyer and Gelautz, 2005; Klaus et al., 2006]. Such formulations force pixels belonging to disparity planes, which are typically obtained by plane fitting using a pre-matched disparity map and a segmentation. Because each pixel must be a part of some plane, the surface assumption is a hard-constraint.

Given a segmentation Seg_s of the reference image $\mathcal{I}_s$, the labeling function τ assigns each segment $s \in Seg_s$ a corresponding plane $\tau(s)$. We minimize the energy function defined as:

$$E(\tau) = \sum_{s \in Seg_s} C(s, \tau(s)) + \lambda \sum_{\{s_i,s_j\} \in \mathcal{N} | \tau(s_i) \neq \tau(s_j)} V_{s_i,s_j}(\tau(s_i), \tau(s_j)) \tag{2.16}$$

where $\mathcal{N}$ is the set of all adjacent segments. V_{s_i,s_j} is a penalty function to indicate the smoothness between planes, which can be incorporated by the common border lengths or mean color similarity.

2.4.3 Methods for Optimization

Since optimization of MRF-based energy functions is generally NP-hard [Kolmogorov and Zabih, 2001], various approximation methods have been proposed in the past few decades – for example, graph cuts, dynamic programming, belief propagation, and region-tree. Szeliski and his collogues have investigated different energy minimization methods for Markov Random Fields with smoothness-based priors [Szeliski et al., 2008]. Applications such as stereo matching, photomontage, binary segmentation, and denoising are demonstrated incorporating test data [Szeliski et al., Online]. An up-to-date and more comprehensive study is presented by Kappes et al. [2013].

Early researchers used simulated annealing [Barnard, 1989; German and German, 1984] to find an approximate solution to their global formulations [Lee et al., 1998]. Since simulated annealing is an energy minimization method, probability maximization formulations need be converted to an energy minimization formulation as shown in subsection 2.4.1 before an approximate solution can be found; this is usually done by assuming that the energy follows a Gibbs distribution. However, simulated annealing can require tens, if not hundreds, of thousands of iterations to find an approximate solution to an energy equation.

The graph cuts technique for minimizing certain energy functions was first introduced for the stereopsis problem by Boykov and Kolmogorov [2004]. Given a binary energy function, a special weighted graph where the vertices correspond to the variables in the energy function can be constructed such that the location of the min-cut on the graph will tell us the variable assignment of the global minimum of the energy function. Boykov et al. leverage this feature of the graph cuts algorithm to find the global minimum of an energy function on binary variables in order to construct two iterative algorithms for finding a local minimum of energy functions over non-binary variables. Boykov et al. also prove that both of these algorithms will find a solution with energy that is within a constant multiple of the global minimum energy. The value of this particular result is questionable in the light of the works of Barbu and Zhu. [2005]; Tappen and Freeman [2003], which show, among other things, that the energy of the true disparity map is usually higher than that of the solution found by graph cuts. However, this surprising effect is due to the definition of the energy function, not to the solver of the energy function.

Loopy belief propagation [Parzen, 1962] has also been used by a number of authors [Brunton et al., 2006; Forstmann et al., 2004; Klaus et al., 2006; Li and Zucker, 2006; Sun et al., 2003; Yang et al., 2006] to great effect. Currently, seven of the top ten reported algorithms in the Middlebury stereo rankings use belief propagation to find an approximate solution of their stereo correspondence formulations; though, whether that is a function of the effectiveness of belief propagation or the energy/probability formulations used has not been tested. Li and Zucker [2006] build geometric constraints in belief propagation to improve non-frontal parallel scenes.

In this subsection we formally introduce two optimization methods in detail. The Mean Field Approximation is to achieve 2.5 for probability formulation; the Semi-Global Matching is addressed to achieve 2.13 for energy function formulation over pixels.

2.4.3.1 Mean Field Approximation

Mean filed approximation [Chandler and Percus, 1988; Parisi, 1988; Peterson and Anderson, 1987] was first used in statistical physics and has been widely used to solve computer vision problems defined on MRFs – for instance, segmentation [Forbes and Fort, 2007], tracking [Medrano et al., 2009], and stereo disparity estimation [Strecha et al., 2006; Yuille et al., 1990].

In mean field approximation, the probability density function (pdf) of a random variable is approximated by some tractable factorized pdfs; the Kullback-Leibler divergence between the

approximating pdf and the true pdf is minimized; the minimization can be achieved by iterative message passing [Riegler et al., 2012]. In the stereopsis problem, the posterior pdf, $\mathrm{P}(\Delta|\mathcal{O})$ of 2.5 is used as the prior $\mathrm{Q}(\Delta)$, in the next iteration. We assume the pdf, $\mathrm{Q}(\Delta)$, can be fully factorized over all sites (pixels) $p \in \mathcal{S}$ as follows:

$$\mathrm{Q}(\Delta) = \prod_{p \in \mathcal{S}} \mathrm{Q}_p(\delta_p) \tag{2.17}$$

where $\mathcal{S}$ is a finite set including all pixels of an image. $\mathrm{Q}_p(\delta_p)$ is a distribution over $|D|$ possible disparity values of $\delta_p \in D$ at the site p.

The solution to achieve a factorized variational distribution, $\mathrm{Q}(\Delta)$ is equivalent to minimize Kullback-Leibler (KL) divergence between the true posterior distribution and the approximate distribution [Yedidia et al., 2003]:

$$KL(\mathrm{Q}(\Delta)\|\mathrm{P}(\Delta|\mathcal{O})) = \sum_{\Delta \in D^{W \times H}} \mathrm{Q}(\Delta) \log \frac{\mathrm{Q}(\Delta)}{\mathrm{P}(\Delta|\mathcal{O})} \tag{2.18}$$

with $KL > 0$; KL is convergent at zero only if $\mathrm{Q}(\Delta)$ is equal to $\mathrm{P}(\Delta|\mathcal{O})$. Putting 2.9 , 2.15 and 2.17 together, we have:

$$\begin{aligned} KL(\mathrm{Q}(\Delta)\|\mathrm{P}(\Delta|\mathcal{O})) = & -\sum_{p} \sum_{\delta_p \in D} \mathrm{Q}(\delta_p) C(p, \delta_p) \\ & -\sum_{p} \sum_{\{p,q\} \in \mathcal{N}} \sum_{\delta_p, \delta_q \in D} \mathrm{Q}_p(\delta_p) \mathrm{Q}_q(\delta_q) V_{p,q}(\delta_p, \delta_q) \\ & +\sum_{p} \sum_{\delta_p \in D} \mathrm{Q}_p(\delta_p) \log \mathrm{Q}_p(\delta_p) \end{aligned} \tag{2.19}$$

$C(p, \delta_p)$ is the match cost function; $V_{p,q}(\delta_p, \delta_q)$ is the smoothness term defined in Equation 2.15. Finding the minimum of 2.19 with respect to Q is a pixel-wise update by averaging the neighbors and can be verified as:

$$\mathrm{Q}_p(\delta_p) \leftarrow \frac{1}{Z} \exp\left(-\left(C\left(p, \delta_p\right) + \sum_{q \in \mathcal{N}_p} \sum_{\delta_q \in D} \mathrm{Q}_q\left(\delta_q\right) V_{p,q}\left(\delta_p, \delta_q\right) \right) \right) \tag{2.20}$$

where Z guarantees $\sum_{\delta_p \in D} \mathrm{Q}(\delta_p) = 1$ after each iteration.

The main defect of Mean Field Approximation is the computational inefficient information propagation. The iterative cost update for each pixel is based only on the local neighborhood, which consists of four or eight directly connected pixels. Once a block of mismatching appearers in a previous iteration, it is difficult to redress such region in following steps.

2.4.3.2 Semi-Global Matching

Dynamic programming [Bensrhair et al., 1996; Birchfield and Tomasi, 1999; Ohta and Kanade, 1985] is widely used to achieve approximate solutions for global stereo formulations [Chen, 2007; Forstmann et al., 2004; Gehrig and Franke, 2007; Gong and Yang, 2005b], due to its simple implementation and fast computation. Dynamic programming solves the stereo problem by minimizing the energy equation 2.15 for each scanline independently.

As no regularization is performed across scanlines, discrepancies occur between the scanlines, this effect is also known as streaking [Neilson, 2009]. Dynamic programming cannot be

generalized to two dimensions, and can only be applied along the epipolar lines. Hirschmüller [2008] proposes a semi-global optimization (SGM) algorithm that performs multiple passes of cost aggregation, in different directions, to eliminate the streaking problem while maintaining a linear complexity of the algorithm. The passes are performed in either the four, eight, or sixteen cardinal directions. The global energy for a disparity map Δ is defined as $E(\Delta)$:

$$E(\Delta) = \sum_{p \in \mathcal{I}_s} (C(p, \Delta(p)) + \sum_{\{p,q\} \in \mathcal{N}} P_1 \left[|\Delta(p) - \Delta(q)| = 1 \right] + \sum_{\{p,q\} \in \mathcal{N}} P_2 \left[|\Delta(p) - \Delta(q)| > 1 \right] \tag{2.21}$$

The first term sums the costs of all pixels in the image with their particular disparities $\Delta(p)$. The second term penalizes a disparity change of 1 with a penalty of P_1. The third term adds a larger penalty of P_2 for disparity differences bigger than 1 pixel. In each direction r, a cost function, $L_r : \mathcal{I}_s \times \mathbb{Z}^{>0} \rightarrow \mathbb{R}$, is computed, such that $L_r(p, \Delta(p))$ provides the cost of assigning a disparity value of δ to pixel p. The SGM algorithm sums the costs of the different directions L_r into a single cost function $A(p, \Delta(p)) : \mathcal{I}_s \times \mathbb{Z}^{>0} \rightarrow \mathbb{R}$:

$$A(p, \Delta(p)) = \sum_r L_r(p, \Delta(p)) \tag{2.22}$$

$L_r(p, \Delta(p))$ in Eq 2.22 represents the cost of pixel p with disparity $\Delta(p)$ along one direction r. It is calculated as following:

$$\begin{aligned} L_r(p, \Delta(p)) =& C(p, \Delta(p)) + \\ & min \begin{cases} L_r(p - r, \Delta(p)) \\ L_r(p - r, \Delta(p) - 1) + P_1 \\ L_r(p - r, \Delta(p) + 1) + P_1 \\ \min\limits_i L_r(p - r, i) + P_2 \end{cases} \\ & - \min\limits_i L_r(p - r, i). \end{aligned} \tag{2.23}$$

The SGM algorithm then calculates the disparity map, $\Delta : \mathcal{I}_s \rightarrow \mathbb{Z}^{>0}$, using winner-take-all as $\Delta(p) = \arg \min_\Delta \{A(p, \Delta(p))\}$.

Semi-Global Matching can be combined with different match cost functions like Mutual Information and Census [Hermann et al., 2011; Hirschmüller, 2008; Humenberger et al., 2010]. Different combinations are evaluated using data sets with small baseline configurations and simulated radiometric changes [Hirschmüller and Scharstein, 2009].

Matching performance is influenced by the penalty values P_1 and P_2. The optimum values depend on the input images. To improve the robustness, the penalty P_2 for large discontinuities can be adapted with respect to the local intensity gradient:

$$P_2' = \frac{P_2}{1 + |\mathcal{I}_s(p) - \mathcal{I}_s(p-1)| / \mathcal{T}}. \tag{2.24}$$

where $\mathcal{T}$ is a user-defined parameter to control the reduction of the penalty. Banz et al. [2012] investigate different penalty functions and found indicate such inverse proportional adaptive functions like 2.24 can perform more robust than constant functions under difficult imaging conditions. In addition, different filtering techniques are compared for matching in the presence of sub-pixel calibration errors [Hirschmüller and Gehrig, 2009]. Sub-pixel refinement using tuned interpolation functions is introduced by Haller et al. [2010]. Humenberger et al. [2010] generate a disparity map using conventional SGM and fit planes according to segments in the post-processing.

Hermann and Klette [2012] introduce an iterative semi-global approach in order to reduce the memory requirement. Using a disparity map in low resolution, a homogenous map is generated to indicate connecting pixels, whose disparities vary more than a user-defined threshold; these pixels are arranged on possible object boundaries. For each aggregation direction, a disparity distance map is generated according to the homogenous map; the distance vectors of a pixel indicate the closest opposite category – surface or edge. The disparity range of surface pixels can be imitated in a small range using the homogenous map; Finding the smallest and largest disparity value using the disparity distance map can limit searching range for edge pixels.

Furthermore, it has recently been shown that SGM can be implemented in real-time on a variety of platforms, like on CPUs, FPGA and GPUs [Banz et al., 2010; Gehrig and Rabe, 2010; Zhu et al., 2010]. The real-time capability, simple implementation and high accuracy of SGM leads to a domination of SGM for a wide range of real-world applications in mobile robotics and remote sensing [Gehrig et al., 2009; Hirschmüller et al., 2012; Zhu et al., 2011].

2.5 Feature based Stereo Algorithms

Feature based stereo algorithms extract primitives like zero-crossing points, edges, line segments, etc. and compare their attributes to find the corresponding features in the other images. In such methods, only significant feature pixels are detected and matched. Textureless regions remain unmatched. Feature-based algorithms were especially popular in the early days of computer vision because pixel-wise information is not reliable to measure the likelihood everywhere.

Early researchers match individual features based on their position [Baker and Binford, 1981; Deriche and Faugeras, 1990; Grimson, 1985]. Matching ambiguities are often constrained by enforcing that adjacent features have similar disparities [Horaud and Skordas, 1989; Medioni and Nevatia, 1985]. The ordering constraint is used to achieve figural continuity [Goulermas and Liatsis, 2000]. Feature based stereo methods generate sparse depth maps, which can be used for dense 3D scene reconstruction by interpolation. Wei and Ngan [2005] set the corresponding edges as seeds and assign depths for the non-edge pixels by interpolating from the nearby assigned disparity values. Besides individual pixel-wise matching of edges, Veksler [2001] presents a region feature defined by a set of connecting pixels. The approaches of Bascle and Deriche [1993]; Brint and Brady [1990]; Robert and Faugeras [1991] represent linked 3D points as 3D curves.

Recently, local descriptors like SIFT [Lowe, 2004] and SURF [Tola et al., 2008] are used for stereo and multi-view matching with unknown epipolar geometry. The corresponding features are typically used to ensure environment perception for estimation camera poses, like in SLAM applications [Tomono, 2009]. Sadeghi et al. [2008] use dynamic programming to match pixels between subsequent edge points, whose disparities are obtained based on epipolar and color constraints. Taylor [2003] extents edge-based correspondences for reconstruction of surface structures. Instead of a global optimization, Geiger et al. [2010] introduce a smoothing prior by forming a triangulation on a set of reliable matching points and reducing ambiguities of the rest points according to the reliable disparities.

However a great number of approaches, common feature-based methods are limited by some factors, even under the global energy frameworks:

- Extraction of features. Almost all feature-based methods use pre-processing to detect features. This causes (1) inhomogeneous feature distribution on input images with textureless regions, where no features can be detected. (2) inaccurate location of features between two stereo images. (3) ambiguous neighbors, if using robust local descriptors, which lead to blurring.

- Interpolation from sparse feature image to dense pixel image. An interpolation for unmatched pixels within homogenous regions can be achieved successfully, only if significant structures are detected and correctly matched. Disparities of non-feature pixels – the majority of an input image – depend completely on the results of feature matching. Often observed in featureless areas of disparity maps, object boundaries are dilated.

Due to these limitations, stereo matching based on low-level features cannot generate dense accurate disparity maps for complicated scenes. However, in this dissertation we introduce a novel confidence-based surface prior for global energy formulations incorporating a dense pixel-wise semi-global optimization method. For a methodical comparison, we introduce in Subsection 2.5.1 two early edge-based match approaches [Lhuillier and Quan, 2002; Wei and Ngan, 2005]. In addition, Subsection 2.5.2 presents the work of Geiger et al. [2010], who have developed a smoothing prior by forming a triangulation on a set of reliable matching points.

2.5.1 Edge-based Match Propagation

Lhuillier and Quan [2002] proposed a quasi-dense matching method based on region growing. The zero-mean normalized cross-correlation is used to match points of interest of two images. The seeds, reliable corresponding features, are selected by restricting their costs and satisfying the consistencies. We denote $\mathcal{N}_s(p)$ and $\mathcal{N}_m(p')$ as neighborhoods of a corresponding seed $(p, p') \in \text{SEED}^{i-1}$ from $(i-1)$ iteration with $p \in \mathcal{I}_s$ and $p' \in \mathcal{I}_m$. Fix two small windows over p and p', the potential matches within this region are limited by the discrete 3D disparity gradient as:

$$\mathcal{N}(p, p') = \{(q, q') | q \in \mathcal{N}_s(p), q' \in \mathcal{N}_m(p'), \|(p - p') - (q - q')\| \leq \varepsilon\} \tag{2.25}$$

where ε is a user defined threshold. All potential matches are stored by decreasing costs that the match with lowest cost ist selected into seed in this iteration. Repeating this processing a quasi-dense depth map can be obtained.

However, textureless regions cannot be matched using neighboring propagation. Wei and Ngan [2005] introduce a Gaussian weighted spatial interpolation using color constraints to fill the sparseness in this disparity map. A density function is used for an iterative interpolation.

2.5.2 Efficient Large-Scale Stereo Matching

Energy using a global optimization method propagate computational inefficient using iterative optimization methods. Early edged-based stereo methods generate only sparse/semi-dense disparity maps and requite interpolation. Geiger et al. [2010] introduce a smoothing prior by forming a triangulation on a set of reliable matching points and reducing ambiguities of the rest points according to the reliable disparities. Their approach leads to an efficient exploitation of energy without a global optimization and is named *Efficient Large-Scale Stereo Matching* (ELAS). The introduced probabilistic generative model enforces smoothness replying on support feature points within a small window. They uses a prior distribution estimated from support points into a local formulation for low computation effort. However, compared to global formulations, the typical challenge for such formulations is poorly-textured regions.

Given a stereo pair $(\mathcal{I}_s, \mathcal{I}_m)$, a set $S = \{s_1, ..., s_m, ..., s_M\}$ contains support feature points of $\mathcal{I}_s$. Each s_m is located at $(u_m, v_m)^T$ in $\mathcal{I}_s$ with matched correspondence Δ_m. The target is to calculate a disparity map Δ_s, which maximizes the probability, $\mathrm{P}(\Delta_s, \mathcal{I}_s, \mathcal{I}_m, S)$. Assuming $\mathcal{I}_s$ and $\mathcal{I}_m$ are conditionally independent to S given Δ_s, the joint distribution factorizes

$$\mathrm{P}(\Delta_s, \mathcal{I}_s, \mathcal{I}_m, S) \propto \mathrm{P}(\mathcal{I}_s, \mathcal{I}_m \mid \Delta_s)\mathrm{P}(\Delta_s \mid S) \tag{2.26}$$

with the image likelihood, $\mathrm{P}(\mathcal{I}_s, \mathcal{I}_m \mid \Delta_s)$ and the prior $\mathrm{P}(\Delta_s \mid S)$.

2.6 Summary

Stereo matching, or more precisely calibrated correspondence searching in 1D, is a multi-labeling problem in the computer vision area, among others, such as flow for dynamic scenes, structure from motion with unknown camera parameters, segmentation, and denoising. The stereopsis problem is generally ill-posed due to its task formulation – infinitely many scenes can be projected on the same image. Many approaches for dense stereo matching have been developed in different contexts; they can be grouped as local/global methods, dense/sparse reconstruction, or probabilistic/energy approaches. However, almost all dense stereo matching algorithms can be decomposed into four main steps Scharstein and Szeliski [2002]: cost computation, spatial aggregation, optimization and refinement. Each step can be solved by many variations of methods – sometimes quite differently. Performance of a stereo method usually depends on the characteristics of the input data. Robust methods that behave well in many application areas are especially important.

Despite intensive research, the computational stereopsis problem remains challenging. When developing global stereo matching methods, there are three passing concerns:

- How can we obtain reliable likelihood? Challenges such as the ambiguity in textureless regions, radiometric changes, and occlusions disrupt computation of a reliable observation, which is the most important component of every stereo algorithm;
- Which properties should the priors have? Smoothness and discontinuity are generally conflicting requirements. Over-smoothing leads to loss of object boundaries. In contrast, less smoothing leads to more noise on the resulting disparity map. As well as the smoothness term, global formulations should express other object properties – for example, the surface constraint; and
- Is the target energy formulation solvable, and how? At the very least the formulations presented in Section 2.4 are NP-hard to find a global solution for, and in some cases finding a global solution is even NP-complete [Kolmogorov and Zabih, 2001].

These problems with which we still struggle today are investigated in this dissertation. We apply the cost-merging strategy to obtain reliable match cost for real-wold applications. In addition, a confidence-based surface prior incorporating with a semi-global optimization is proposed.

Chapter 3

Robust Match Cost Functions for Dense Stereopsis

Match cost functions measure the fitness of assigning a disparity value $\delta \in D$ to a pixel $p \in \mathcal{I}_s$. This fitness is evaluated by warping p at disparity δ into the other given image and comparing the information (such as intensity, color, or color gradient) at p with the information at the warped-to pixel $q \in \mathcal{I}_m$. The cost function helps to determine the likelihood of pixels p and q being projected from the same object point.

Currently, almost all of the best stereo-matching algorithms are framed as global energy minimizations, which aim to solve the stereopsis using smoothing assumptions. However, match cost is always included, whether as the likelihood in a probabilistic formulation of Eq. 2.7 or the data term in a energy formulation of Eq. 2.15. In contrast to the smoothness prior, which is based on empirical assumptions for the unknown result, match cost is measured directly from the input images.

Match cost functions can be grouped into parametric costs, mutual information, and non-parametric costs[Scharstein and Szeliski, 2002]. The common parametric costs include absolute difference (AD), the sum of absolute difference (SAD), Birchfield and Thomasi (BT), normalized cross correlation (NCC), and other extensions based on these [Birchfield and Tomasi, 1998; Heo et al., 2011; Ryan et al., 1980]. Mutual Information is introduced by Viola and Wells [1997] and enables the registering of images with complex radiometric relationships in a stereo pair [Chrastek and Jan, 1998]. Non-parametric costs like rank and census transformations detect local structures within a support window and are therefore invariant to many radiometric changes [Humenberger et al., 2010; Zabih and Woodfill, 1994].

As introduced in subsection 2.2.1, match costs are limited by data uncertainty, structural ambiguity, and a lack of visibility. Data uncertainty is caused by an insufficient signal-to-noise ratio in images of weakly textured objects. Structural ambiguity refers to the presence of periodic structures combined with the inability of a small number of cameras to capture the entire light field. The lack of visibility that blocks part of one image from being seen in the other results in erroneous being estimated for such areas. These difficulties are caused by the following effects:

- Observation conditions contain all imagery configurations when the images are captured – for example, exposure time, illumination, baseline length, and stereo angle. Different imagery configurations cause over-/under-exposure, radiometric changes, and occlusions.
- The regions in which match cost is applied present the features of imagery objects, which are quite different. We extend the definition of land cover from the remote-sensing area,

in order to describe the imagery features of homogeneity, continuity, and the texture of covering regions.

The performance of a global stereo matching method depends on components including match cost, cost aggregation, optimization, and disparity-map refinement. The performance of match cost is often separately investigated, independently of other components. The works of Hirschmüller and Scharstein [2009]; Neilson and Yang [2011]; Scharstein and Szeliski [2002] indicate that match costs perform very differently with close-range data sets. Zhu et al. [2011] evaluate match costs on remote-sensing data and show the challenges of real-world applications. However, a systematic investigation on match costs incorporating different data is absent. In this chapter we introduce a sample formulation of match cost functions and address their advantages and disadvantages. Then we present the cost-merging strategy to obtain robust match costs. The evaluations using the Middlebury data and the remote-sensing data are finally presented in Chapter 5.

3.1 Related Work

Dense stereo algorithms are typically evaluated using data sets with a small baseline configuration as well as artificial and often ambient light sources. Radiometric changes due to vignetting, gamma changes, and so forth, have been analyzed for modifying small baseline images [Hirschmüller and Scharstein, 2009; Scharstein and Szeliski, 2002], but these might not capture the effects caused by large baselines, especially for objects with strongly non-Lambertian reflectance behaviors. Real exposure and light-source changes in an indoor environment are used in the work of [Hirschmüller and Scharstein, 2009]. However, data sets under natural light sources such as the sun are not included. In these previous evaluations, census shows the best and most robust overall performance. Mutual information performs very well with global methods. On radiometrically distorted Middlebury data sets and data sets with varying illumination, census and mutual information clearly outperform absolute difference.

Neilson and Yang [2011] introduce a cluster-ranking-based statistical evaluation method for constructible matching measures. The Middlebury data sets and synthetically generated image pairs with simulated noise are evaluated using different global stereopsis frameworks. Their analysis indicates that no single match cost function is perfect for any situation. Non-parametric match costs like census are not included. In the work of Hermann et al. [2011], census, the absolute difference using gradient images, and the sum of absolute difference (SAD) are evaluated on driving straight frames for urban scenarios and the Middlebury 2005 sets. The gradient-based match cost seems to outperform census slightly, as the illumination differences are not strong.

The work of Gong et al. [2007]; Min et al. [2011]; Tombari et al. [2008] focuses on the aggregation techniques by offline global optimization algorithms. Their evaluations are limited using only the four data sets of the Middlebury online benchmark. Furthermore, some researchers have demonstrated their results using remote-sensing data, but qualitative evaluations are not given, probably due to the absence of groundtruth [Heo et al., 2008; Pock et al., 2008].

Thus, an investigation on matching performances of different cost functions including both close-range and remote-sensing data was not performed in the literature. In the work of Zhu et al. [2011], mutual information and census using intensity images are evaluated on an airborne image sequence with increasing baseline length and satellite stereo pairs with large view angles. In this dissertation we extend our investigation to gradient images and close-range data with radiometric changes are also included. We study interdependencies among match performance,

cost functions, and data sets used. Based on these studies, we contribute a cost-merging strategy, which considers the advantages of different match-cost functions in the context of imagery conditions and data features.

3.2 Intensity, Color and Gradient

In the human visual system, both chromatic (color) and achromatic (luminance/brightness) mechanisms are used for stereopsis [Jordan et al., 1990]. However, because it is still unclear exactly how these mechanisms are applied in concert, it is not possible to use human-intuitive measures for computational stereopsis. The human visual system seems to be able to determine whether the achromatic or chromatic mechanism is more accurate [de Dios and Garcia, 2003]. In contrast, almost all recent computational dense match measurements are generated using some low-level cost functions, which use image intensity, color or/and gradient.

There are many different representations of color that can be used, each with their own benefits and drawbacks. One of the standard color spaces is the vector space using RGB representation, which is most familiarly applied for computer screens, digital camera sensors, etc. A color in a RGB space is represented as a triple $\vec{c} = (r, g, b)$. The components represent the amount of the primary colors (r for red, g for green, b for blue) that are combined to create the color. The hue, saturation, value (HSV) color space is commonly used by digital artists to simplify the process of selecting colors. The H component of an HSV color is an angular representation of the hue of the color, the S component indicates how saturated the color is, and the V component encodes the luminance. The HSV representation of a RGB color , (r,g,b), is defined as follows:

$$V = \max(r, g, b) \tag{3.1}$$

$$S = \begin{cases} \frac{V - \min(r,g,b)}{V} & if V \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

$$H = \begin{cases} 0 & S = 0 \\ \frac{60(g-b)}{S} & V = r, S \neq 0 \\ 120 + \frac{60(b-r)}{S} & V = g, S \neq 0 \\ 240 + \frac{60(r-g)}{S} & \text{otherwise} \end{cases} \tag{3.3}$$

Then, if $H < 0$ it is placed into the range [0, 360) by adding 360 degrees.

For gray-scale images, the pixel value is represented by the brightness of the pixel in single or multiple color channels. Depending on the scaling range, the pixel value is often stored as an 8 to 32-bit integer. For an 8-bit image, zero is typically taken to be black, and 244 is taken to be white. Moreover, the image gradient indicates the directional changes in color or intensity of an image. The gradient at each pixel is a 2D vector with components given by derivatives in the horizontal and vertical directions, $\bigtriangledown(p) = (\bigtriangledown_x(p), \bigtriangledown_y(p))$. Pixels where the edges are located are assigned the largest magnitude along the normal of the edge-tangent line. Image gradients have already been used as features for stereo matching by Pollard et al. [1985].

As a 3D point is projected onto different camera planes, these corresponding pixels may have the same presentation (color/intensity), only if the illumination conditions and reflection properties are ideal. In practise, pixels projected from the same object point often have different

colour/intensity values in two images. Pixel-wise brightness constancy is disrupted by radiometric changes and noise. With the exception of underexposure or overexposure, color gradients are, due to significant edge locations, normally less sensitive to radiometric changes. However, match measures using gradient images are limited in homogenous regions, uncertain localization of features and the gradient-derivation direction[1]. In homogenous regions, the image gradients are frequently similar. The gradients can be observed not only on object boundaries, but also on surfaces, depending on the image textures. Even edges of object boundaries can be shifted between two input images due to changing view points, especially for stereo pairs with a large baseline. Moreover, a gradient image is often derived in the x and/or y direction, which causes some lines parallel to these directions having very low gradient values.

3.3 Parametric Match Costs

Parametric match costs are commonly pixel-based and include absolute differences, squared differences, and the truncated versions using intensities, colors or gradients. The basic assumption of most parametric cost functions is the brightness/gradient constancy between two corresponding pixels.

The simplest cost function is the absolute difference (AD), which can only work on data with the Lambertian reflectance and good camera configurations. In Equation 3.4, $\mathcal{I}_s(p)$ and $\mathcal{I}_m(p-\delta)$ denote the intensities of pixel p in the source image and its corresponding pixel with disparity δ in the match image separately.

$$C_{AD}(p,\delta) = |\mathcal{I}_s(p) - \mathcal{I}_m(p-\delta)| \tag{3.4}$$

A user-defined threshold is often used to truncate the maximal intensity difference between two pixels in order to suppress outliers. The truncated absolute difference (TAD) is then defined as:

$$C_{TAD}(p,\delta) = \begin{cases} T & |\mathcal{I}_s(p) - \mathcal{I}_m(p-\delta)| > T \\ |\mathcal{I}_s(p) - \mathcal{I}_m(p-\delta)| & \text{otherwise} \end{cases} \tag{3.5}$$

For multi-channels images, the intensity defined in Equation 3.4 and 3.5 can be replaced by colors or color gradients.

Spatial aggregation is optionally used to limit the influence of mismatches. The basic assumption of the spatial aggregation is that the disparities within a small region are almost constant. Typical spatially aggregated cost functions include the sum of absolute/squared differences (SAD/SSD) [Kanade, 1994; Matthies et al., 1989] and normalized cross correlation (NCC) [Ryan et al., 1980]. These cost functions measure the compatibilities between source and match images with a candidate shift at every pixel. A constant offset (bias) of pixel intensity values is compensated by the zero-mean versions ZSAD and ZSSD. Such methods are often applied with winner-takes-all optimization as local stereo algorithms.

The sum of absolute differences (SAD) method is done by summing of intensity/color differences over a rectangle window:

$$C_{SAD}(p,\delta) = \frac{1}{m \times n} \sum_{p \in W, \delta \in D} |\mathcal{I}_s(p) - \mathcal{I}_m(p,\delta)| \tag{3.6}$$

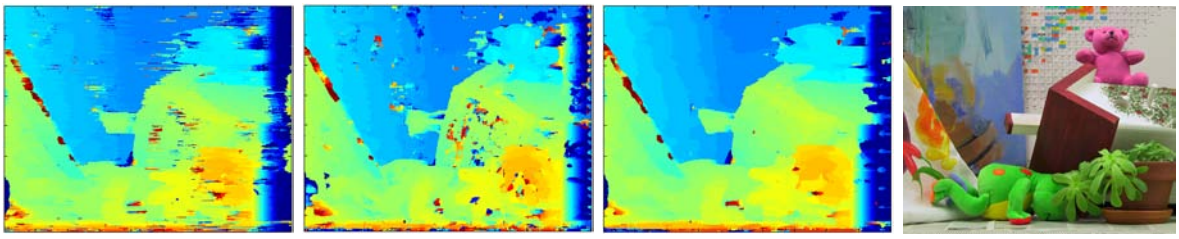
[1] Generally, the gradients should be derived in the diagonal direction of the epipolar line.

Similarly, the normalized cross correlation (NCC) is defined as the product of the two intensity vectors normalized over a window:

$$C_{NCC}(p,\delta) = \frac{1}{n} \frac{\sum_{q\in\mathcal{N}_p} \mathcal{I}_s(q)\mathcal{I}_m(q-\delta)}{\sqrt{\sum_{q\in\mathcal{N}_p} \mathcal{I}_s(q)^2 \sum_{q\in\mathcal{N}_p} \mathcal{I}_m(q-\delta)^2}} \tag{3.7}$$

where n is the number of pixels in $\mathcal{N}_p$.

The matching performance of spatial aggregations depends on the window size used, which is typically fixed. Finding an optimal size for the whole image is impossible. A large window size includes enough colour/intensity variations for robust matching on object surfaces, but at the cost of over-smoothing. Figure 3.1 demonstrates the resulting differences of SAD using various window sizes. Some adaptive-window approaches select the optimal window sizes and/or shapes automatically according to local structures such like color gradients [Boykov et al., 1998; Kanade and Okutomi, 1994b; Veksler, 2003]. Segmentation-based approaches select the aggregation shapes using given segments of the input source image [Gong and Yang, 2005b; Tao and Sawhney, 2000; Wang et al., 2004]. Such approaches can improve matching performance, only if the pre-processing – feature detection or segmentation, works well. Moreover, cost aggregation prefers fronto-parallel surfaces such that all pixels within a support window are assigned the same disparity.



(a) 3 × 31 (b) 7 × 13 (c) 9 × 31 (d) Reference image

Figure 3.1: WTA results of SAD using various window sizes. Large window size reduces noises in cost of over-smoothing.

The advantage of cost aggregations with a fixed window size is that the calculation can be implemented on parallel architecture – for example, modern graphics cards (GPUs). The computation time of cost aggregations using regular window increases proportionally to the window size [Zhu et al., 2010]. For instance, Figure 3.2 shows the run-time comparison of SAD with different window sizes on a 480 × 375 pixels stereo pair with 64 pixels disparity range using a NVIDIA GeForce GTX 295 graphics card.

$W_{m\times n}$	n=3	5	7	9	11	13	15	17	21	25	29
m=3	4.9	8.4	10.0	14.3	18.6	121.0	59.5	67.1	82.2	97.3	113.2
5	9.5	13.8	18.1	24.6	30.1	37.3	122.5	138.4	170.2	201.9	234.6
7	14.6	22.7	29.3	38.3	48.1	59.5	194.6	220.0	270.6	322.0	371.9
9	17.3	26.2	35.9	45.6	56.9	69.7	235.4	266.3	327.7	389.9	451.4

Figure 3.2: The run-time analysis of the SAD method with different fixed window sizes m × n using CUDA on GPUs.

3.4 Mutual Information

Mutual information (MI) is introduced by Shannon [1948] and popularized for image registration problems by Viola and Wells [1997]. It is used to measure the similarity between images and is realized by some variant of gradient descent [Kim et al., 2003; Viola and Wells, 1997]. Egnal [2000] has introduced applying mutual information for correlation based dense stereo matching. The key advantage of mutual information is the ability of handling complex radiometric relationships between intensities of images. In stereopsis, mutual information can be insensitive to recording and illumination changes [Hirschmüller and Scharstein, 2009]. In order to obtain a reliable initial disparity map, Hirschmüller [2008] presents a hierarchical strategy to calculate joint entropy iteratively and efficiently.

Give a stereo pair $\mathcal{I}_s, \mathcal{I}_m$ and an initial disparity map Δ, let $\mathcal{I}_m^{'} = \Delta(\mathcal{I}_m)$ be the warped image according to Δ. The mutual information is defined by the individual entropies of input images and their joint entropy, $H_{\mathcal{I}_s,\mathcal{I}_m^{'}}$:

$$MI_{\mathcal{I}_s,\mathcal{I}_m^{'}} = H_{\mathcal{I}_s} + H_{\mathcal{I}_m^{'}} - H_{\mathcal{I}_s,\mathcal{I}_m^{'}} \tag{3.8}$$

where the entropies are calculated by the probability distribution of intensities from $\mathcal{I}_s$ and associated $\mathcal{I}_m^{'}$:

$$H_{\mathcal{I}} = -\int_0^1 P_{\mathcal{I}}(i) \log P_{\mathcal{I}}(i) \mathrm{d}i \tag{3.9}$$

$$H_{\mathcal{I}_s,\mathcal{I}_m^{'}} = -\int_0^1 \int_0^1 P_{\mathcal{I}_s,\mathcal{I}_m^{'}}(i_s, i_m) \log P_{\mathcal{I}_s,\mathcal{I}_m^{'}}(i_s, i_m) \mathrm{d}i_s \mathrm{d}i_m \tag{3.10}$$

with $\mathcal{I} \in \{\mathcal{I}_s, \mathcal{I}_m^{'}\}$ and $i \in [0, 255]$. This definition of mutual information is over the full images and requires the disparity map a priori to calculate the joint entropy. To allow pixel-wise matching, Kim et al. [2003] introduce a transformation of 3.10 into a sum over overlapping pixels using Taylor expansion:

$$H_{\mathcal{I}_s,\mathcal{I}_m^{'}} = \sum_p h_{\mathcal{I}_s,\mathcal{I}_m^{'}}(\mathcal{I}_s(p), \mathcal{I}_m^{'}(p)) \tag{3.11}$$

where $h_{\mathcal{I}_s,\mathcal{I}_m^{'}}$ is calculated by the joint probability distribution $P_{\mathcal{I}_s,\mathcal{I}_m^{'}}$ of corresponding intensities. Under the assumption of the individual entropies are almost constant, Kim et al. [2003] define the data energy of 2.15 as the negative sum of $h_{\mathcal{I}_s,\mathcal{I}_m^{'}}$ (mutual information is maximized, while energy is minimized). However, the intensities having non-correspondence in occlusion areas should not be included in the calculation [Hirschmüller, 2008]. They suggest to consider the individual entropies in data term to improve object borders:

$$H_{\mathcal{I}} = \sum_p h_{\mathcal{I}}(\mathcal{I}(p)) \tag{3.12}$$

Finally, given a disparity map Δ, we can quantify 3.8 by the pixel-wise summing over the whole image:

$$MI_{\mathcal{I}_s,\mathcal{I}_m,\Delta} = -E_{data} = -\sum_p C_{MI}(p, \delta) \tag{3.13}$$

where $\delta = \Delta(p)$. The matching cost function using mutual information is now defined as:

$$C_{MI}(p, \delta) = -h_{\mathcal{I}_m}(\mathcal{I}_m(p)) - h_{\Delta(\mathcal{I}_s)}(\mathcal{I}_s(p, \delta)) + h_{\mathcal{I}_m,\Delta(\mathcal{I}_s)}(\mathcal{I}_m(p), \mathcal{I}_s(p, \delta)). \tag{3.14}$$

3.5 Non-Parametric Matching Costs

Zabih and Woodfill [1994] introduce non-parametric local transforms for computing optical correspondences. Their approach includes two local transforms, rank and census, which rely on the relative order of local intensity values, not the values themselves. Since all non-parametric cost functions calculate the match distance only depending on the ordering of colours/intensities and not the magnitude of intensities, they tolerate all radiometric distortions that preserve this ordering.

The implementations of rank and census transformation are filtering followed by a comparison using the absolute difference or Hamming distance. Permutations of two corresponding windows are ranked and their alignment bits are calculated to measure the feature distance. Rank transform of an image, $\mathcal{I}$, is defined as the amount of pixels within a local region, whose intensity is lower than the center pixel.

$$Rank_{\mathcal{I}}(p,q) = \|\{q \in \mathcal{N}_p | \mathcal{I}(q) < \mathcal{I}(p)\}\| \tag{3.15}$$

The rank transformation is known to be susceptible to noise in textureless areas. Thus, we define the match cost using the rank transformations of $\mathcal{I}_s$ and $\mathcal{I}_m$ as follows:

$$C_{Rank}(p,p') = |Rank_{\mathcal{I}_s}(p,q) - Rank_{\mathcal{I}_m}(p',q')| \tag{3.16}$$

where $p' \in \mathcal{I}_m$ is the wrapped pixel of $p \in \mathcal{I}_s$ with $p' = p - \delta$.

Census transformation is invariant to monotonic gray value changes, and thus, can tolerate a large class of global and local radiometric changes. It encodes the local image structure within a transform window and defines a bit string where each bit describes the relative ordering between the computing pixel and its local neighbor. A bit is set if a pixel inside the window has a lower intensity than the center pixel. The distance between two bit strings is computed using the Hamming distance. In our work, a 9×7 [1] window is used and supports the matching costs in the range of 0 to 63 [2]. ξ denotes a census transform within a window of image $\mathcal{I}$, $W(\mathcal{I})$. $\bigotimes$ computes the Hamming distance:

$$C_{Cen}(p,\delta) = \bigotimes\Big(\underset{W(\mathcal{I}_s)}{\xi}(p), \underset{W(\mathcal{I}_m)}{\xi}(p-\delta)\Big) \tag{3.17}$$

Figure 3.3 illustrates the calculation of a census match cost using a 3×3 window. Each pixel within this window is compared with the center pixel such that a binary bits code is built to describe the local structure. Doing bitwise exclusive or of two bit strings and summing all bits equal one, the local difference between two windows is calculated.

In contrast to parametric match costs and spatial cost aggregations, non-parametric match costs such as census transformation have advantages in the following ways:

- encoding local spatial structure, which is relative insensitive to radiometric changes;
- reducing effects of variations caused by the camera's gain and bias;
- increasing robustness in dealing with outliers near to depth-discontinuities; and
- tolerating to factionalism. If a minority of pixels in a local neighborhood has a highly different intensity distribution than the majority, only comparisons involving a member of the minority are affected.

[1] Small window size reduces the matching robustness. Large window size leads to blurring.
[2] 64 bits is the maximal bit size allows a fast implementation.

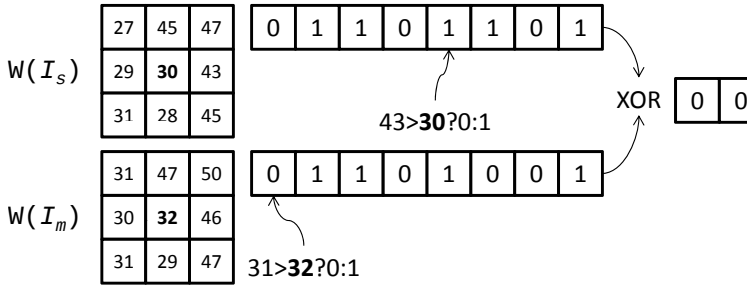


Figure 3.3: Illustration of calculating census match cost.

However, compared with parametric match costs and mutual information, non-parametric match costs have loss of information associated with the pixel due to the window-based computation. Local transformations using large window sizes lead to dilated object edges. To avoid blurring and for a fast implementation, we used to use a window size with maximal 7×9 pixels. The magnitude of intensities is therefore strongly elided in contrast to gray values. For a census window of 3×3 pixels, the variable to store the census value would be of size 23, or 8 bits; for a census of window size 5×5 the number of bits required to store the census value would be 25 bits.

3.6 Match costs merging

Same match cost can perform differently with respect to regions used, even within a same stereo pair. Our motivation of merging match costs comes from such context-dependent performances of match costs. Most stereo methods have been developed using in-lab captured stereo pairs. How far can such data differ from real-world data? Figure 3.4 (c) and (d) demonstrate the radiometric changes of a Middlebury stereo pair and an airborne stereo pair with large baseline respectively. The lightness densities of the **Flowerpots** stereo pair in HSV color space are highly similar, thus parametric costs are able to measure similarity of corresponding pixels. In contrast, the lightness distributions on building roofs of the airborne stereo pair are quite different.

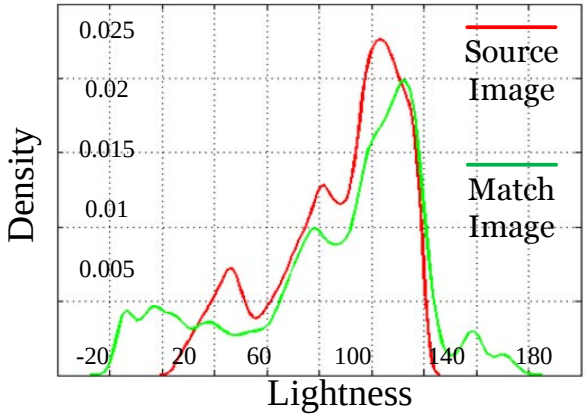
Various cost functions perform differently in context of illuminations, object type (landscapes), and stereo configurations. In fact, all factors appear concurrently such that a separate evaluation for each factor is impossible. In Table 3.1, based on the evaluation shown in Chapter 5, we summarize the performances of match costs in scale according to four criteria: how large are the radiometric changes between two stereo images; how lange is the baseline between two camera principle points; how rich is the image textured?; and how smooth is the scenario? In the first row of this table, the light colors of gray bars indicate less radiometric changes, short base line, less texture, and less smoothness of the object respectively, vise versa. Matching performances under these criteria are scaled from light (for bad match) to dark (for good and robust match) colors. For instance, non-parametric cost functions perform more robust than mutual information in weekly-textured regions. But, mutual information can generate sharper edges than non-parametric costs. As shown in this table, different match cost functions are complementary in different situations. Thus, match-costs merging according to imagery conditions can combine their advantages and improve matching performance in robustness and accuracy.

In this work we merge absolute difference (AD) or mutual information (MI) with census respectively using a weighted average::

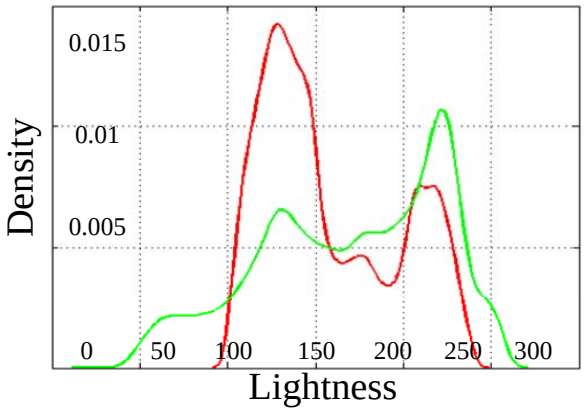
$$C_{x+Cen}(p,\delta) = w_x \times C_x(p,\delta) + (1 - w_x) \times C_{Cen}(p,\delta). \tag{3.18}$$

(a) **Flowerpots** from Middlebury stereo pairs and mask on pots surfaces

(b) Airborne stereo pair with large baseline and mask on building roofs



(c) Lightness density of (a)



(d) Lightness density of (b)

Figure 3.4: Lightness densities of two stereo pairs in HSV colour space. The analyses are limited using the masks for object surfaces and building roofs respectively. The images of **Flowerpots** have similar lightness distributions shown in (c). In contrast, the lightness densities of two airborne images are different as shown in (d).

where x is either absolute difference or mutual information. For easier combination with other costs, we rescale each match cost into a range from 0 to 1023. w_x denotes the weight parameter, which is scaled in the range $[0, 1]$.

The most important reasons for radiometric changes in a real-world application are dynamic light sources and large stereo baseline. In the case of stereo pairs with static light sources and small baseline, the radiometric changes for corresponding pixels are comparatively less presented.

Data with	Radiometric Changes - +	Baseline Length - +	Texture - +	Smoothing - +
Parametric Costs	+ - -	+ -	- - +	+ -
Mutual Information	+ -	+ -	- - +	+ -
Non-parametric Costs	+ -	+ -	- +	- +

Table 3.1: A qualitative comparison of match costs under different criteria. The light color in a gray bar in the first row indicates less radiometric changes, short base line, texture-less regions and less smoothing on result, vise versa. Matching performance is scaled from light (for bed) and deep (for good) colors.

Non-parametric match costs perform more robust than other cost functions for data sets with large radiometric changes and homogenous surface areas. However, parametric match costs and mutual information generate sharp edges, if the photogrammetric consistencies remain in a stereo pair. Thus, the weight parameter w_x should be adapted to imagery conditions in order to use the advantages of different match costs. In chapter 5, we demonstrate the parameter tuning of w_x for stereo pairs with a increasing baseline length and show the dependence between matching performance and the weight selection.

Matching cost	intensity-based	gradient-based
Absolute Difference	$C_{AD}(\mathcal{I}^{3\to1})$	$C_{AD}(\bigtriangledown_e\mathcal{I})$
Mutual Information	$C_{MI}(\mathcal{I}^{3\to1})$	$C_{MI}(\bigtriangledown_e\mathcal{I})$
Census	$C_{Cen}(\mathcal{I}^{3\to1})$	$C_{Cen}(\bigtriangledown_e\mathcal{I})$
Sum of AD and Cen	$C_{AD+Cen}(\mathcal{I}^{3\to1})$	$C_{AD+Cen}(\bigtriangledown_e\mathcal{I})$
Sum of MI and Cen	$C_{MI+Cen}(\mathcal{I}^{3\to1})$	$C_{MI+Cen}(\bigtriangledown_e\mathcal{I})$

Table 3.2: Abbreviations of different match costs. We consider both gray value and gradient based variants of matching costs. Color channels are averaged, if they are available. The gradient images are created using the partial derivative with respect to the epipolar direction.

3.7 Summary

Match cost is one of the most important components of stereo processing. All stereopsis algorithms need match costs to measure the similarity of two corresponding pixels or regions. The robustness of a match cost is decisive for the matching performance, also for global methods.

In this chapter we describe parametric, non-parametric match costs, and the match cost based on mutual information. We analyzed their advantages and disadvantages from their mathematical definitions. Their matching performances are qualitatively summarized in context of data (radiometric changes, object features) and observation conditions (stereo baselines). Depending on applications, all these challenges might be present at the same time. There is no single cost function that performs best for all circumstances. Parametric match costs without spatial aggregations are pixel-wisely calculated, thus fine details, such as building edges, can be accurately reconstructed. Match costs based on mutual information expose the global radiometric relationship between two images and can reduce effects of variations caused by the camera's bias and gains. Non-parametric match costs encode image local structures and are invariant to any monotonic radiometric changes. However the window-based mechanisms of non-parametric costs cause lose of shape edges and fine details. To develop match costs for

real-world data, we introduce the cost-merging strategy with respect to the imagery contexts like stereo baseline length.

In this dissertation absolute difference, mutual information, and census are evaluated using intensity as well as gradient images. In addition, the linearly merged match costs using AD and MI with census respectively, are compared with the conventional match costs. Table 3.2 summarizes the abbreviations of all ten match costs used in this work. For a representative evaluation, both the Middlebury data sets with/without radiometric changes and the remote-sensing data with short/long baselines are used in Chapter 5.

Chapter 4

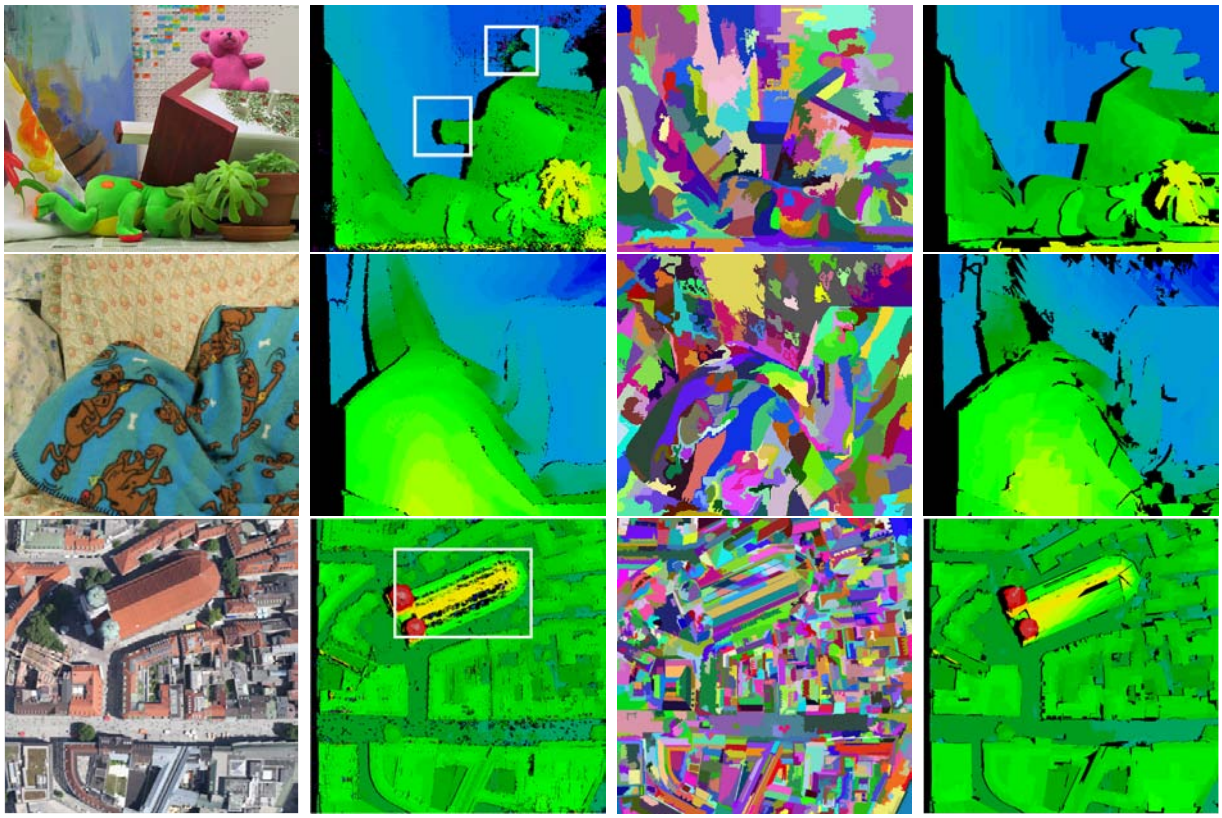
Probabilistic Pixel-wise Surface Stereo

In this chapter we introduce a probabilistic model for pixel-wise surface stereo matching. This approach builds a surface prior over the disparity space using the confidence based on a set of reliably matched correspondences. For each pixel, the probability of its depth lying on an object plane is modeled as a Gaussian distribution, whose variance is determined using the confidence from a previous matching. We minimize our global formulation in energy space using iSGM3 – a modified semi-global optimization for three terms. This method is iterative. Since the confidences are derived with respect to image likelihood and smooth prior probability, our approach has decreased sensitivity to a given image segmentation in comparison to existing segment-based stereo methods.

Global energy minimization algorithms that are defined over image regions, rather than pixels, achieve the top few ranks in the de-facto standard Middlebury online benchmark when sorted according to depth discontinuities (disc) [Bleyer et al., 2010, 2011; Klaus et al., 2006; Scharstein and Szeliski, 2002; Taguchi et al., 2008; Wang and Zheng, 2008]. Our own experience and the Middlebury benchmark indicate that these region-based methods are preferable around object boundaries and in large homogeneous areas; the use of regions helps propagate strong matches into sub-regions with poor matches. However, defining an energy minimization over regions, rather than pixels, imposes a hard constraint that forces depths to lie on the smooth surface associated with a region; removing fine-level details from the depth map in the process. In our method, we relax the hard region-surface constraint imposed by these methods and continue to gain the benefits of region-based methods by defining an energy minimization over pixels that incorporates a probabilistic constraint that depths lie near, rather than on, a pre-calculated smooth surface. We adjust the strength of our constraint using a measure of the confidence we have that the smooth surface is correct; relaxing the constraint when we believe that it will not benefit the disparity calculation.

As shown in Figure 4.1, the introduced confidence-based surface prior for global energy minimization formulation addresses these problems:

- *Foreground/background fattening*: Often observed in the occlusion areas of Middlebury benchmark sets Scharstein and Szeliski [2002] and in the shadow regions of remote sensing data Kurz et al. [2012], objects are dilated into weakly matching areas. This problem is demonstrated in Figure 4.1: the pink teddy bear, church towers, and the edges of buildings that are near shadowed streets. The matching costs are ambiguous, causing the energy minimization to dilate the objects beyond their borders. Robust matching cost like



(a) Source images (b) Results using SGM (c) Segmentation of (a) (d) Results using hard constraint

Figure 4.1: Segment-based stereo matching is very sensitive to the given segmentation. As this hard plane constraint performs well on **Teddy** (less blurring within the white maskers), segmentation artifacts are observed on the texture-rich areas of **Cloth3**. The result of the airborne image including a large homogenous roof has generally sharper building shapes using the hard constraint. However, a lot of shadowing streets are enforced as a disparity-plane part of the building roofs.

Mutual Information Viola and Wells [1997] can improve the matching performance only partially Hirschmüller and Scharstein [2009]; Zhu et al. [2011].

- *Sensitivity to image segmentation*: As shown in Figure 4.1, segmentation-based methods can perform very well, if the given segments are correlated with the object boundaries Bleyer et al. [2010]; Klaus et al. [2006]; Taguchi et al. [2008]; Wang and Zheng [2008]. However, the results are very sensitive to the given segmentations. Oft over-segmentation is required. In texture-rich regions, artifacts from segmentation can be appeared.

- *Incorrect matching in large homogeneous areas*: In a large low-texture or homogeneous area, parametric match costs can be unreliable; even when using a global energy minimization algorithm. This is demonstrated in the church roof in Figure 4.1. Non-parametric, window-based, matching costs like census Zabih and Woodfill [1994] can overcome this problem, but can lead to dilated edges Zhu et al. [2011].

In this chapter, the formulation and the implementation of the proposed probabilistic surface stereo method are introduced. The performance of this approach using both the Middlebury benchmark data sets and the real-world remote sensing data is discussed in Chapter 5. The

remainder of this chapter is organized as follows. In Section 4.1 the related works are outlined. Section 4.2 introduces the hard surface constraint used for stereo matching and points out its limitations. Then, we describe the confidence-based surface prior for energy optimization in Section 4.3. Finally, the iSGM3 optimization for an energy formulation with data term, smoothness prior, and surface constraint is introduced in Section 4.4.

4.1 Related Work

One of the main defects of pixel-wise matching is the absence of semantic information. Pixel-wise observation is limited within a small and often a regular window, even using some adaptive strategy. Image features like edges, corners and line segments are sparse and heterogeneously distributed. Combining such feature primitives with dense matching methods is prevented in texture-less regions such like occlusions and shadow regions, because the data costs are ambiguous. In contrast, color segment allows to trade similar pixels as a whole unit. In this section, we throw a brief overview about the methods developed in the last years.

The works of Koschan et al. [1996]; Wei and Quan [2004] assume that pixels within a segment have the same disparity. Over-segmentation and region splitting are required to relax the constraint. Tao and Sawhney [2000] enforce the depth of each homogeneous region as a nominal plane with allowable additional smooth depth variations. Cai et al. [2005] employ matching segments on scanlines using fuzzy set theory to handle ambiguities.

More relevant to this dissertation are methods formulated as a global energy function. Many of the best ranked methods in the Middlebury online benchmark use a region-based, rather than pixel-based, global approach [Klaus et al., 2006; Taguchi et al., 2008; Wang and Zheng, 2008]. These approaches define Equation 2.15 in a way that assigns a smooth surface (usually a plane) to image regions, rather than depth to pixels. The performance of this hard constraint, that all depths in a region be on the same smooth surface, is highly influenced by the quality of the subdivision of the image into regions. If depth discontinuities are not coincident with the border between regions, then they will be lost in the resulting disparity map. Thus, these methods rely on over-segmentation of the image into small regions to maintain good accuracy. By incorporating a soft, rather than a hard, constraint the proposed method is very robust to the color segmentation of the image into regions. Saygili [2012] initiate a disparity map by matching with SURF key points. Each image segment is assigned with a disparity plane using graph cuts. Aydin and Akgul [2010] introduces a synchronous energy optimization with segment based regularization.

Sun et al. [2005], Bleyer et al. [2010], and Woodford et al. [2009] have all proposed different soft constraints for region-based stereo algorithms that are different from our proposed constraint. The soft constraint proposed by Sun et al. is most similar to the proposed method in that they introduce the addition of a single soft constraint term to the energy minimization formulation in Equation 2.15 that encourages the disparity of a pixel to lie near a plane calculated from a given disparity estimation. However, their soft region constraint does not incorporate confidence; they assume that the provided disparity estimation is trustworthy. The algorithms proposed by Bleyer et al. [2010, 2011] split a given segmentation into overlapping subsegments, and add a term to their formulation that softly constrains overlapping segments to contain a single contiguous surface. Their energy formulations with seven optimization terms could lead to parochial usability depending on applied scenes. Unlike the proposed method, Woodford et al. utilize an over-segmentation of the image into many small regions, and propose a weighting of their smoothness term that discourages disparity edges from cutting through regions.

Common methods for deriving an approximate minimum of Equation 2.15 include loopy

belief propagation [Pearl, 1988; Sun et al., 2003], graph cuts [Boykov and Kolmogorov, 2004], iterated condition modes [Jodoin and Mignotte, 2004], fusion moves [Lempitsky et al., 2007], and dynamic programming [Gong and Yang, 2003; Hirschmüller, 2008]. The semi-global matching method proposed by Hirschmüller [Hirschmüller, 2008] is widely used in the photogrammetry, remote sensing, and intelligent vehicle application areas due to being orders of magnitude faster than the other methods of optimization while still producing high quality results.

4.2 Hard Surface Constraints

Several factors drive us to add surface prior into global energy formulations, softly. A global method enforces smoothness of results by minimizing a MRF-based energy function, which typically can be decomposed into two parts, an image likelihood and a smoothing prior. Often the smoothness assumption has to be compromised with the match costs in order to allow sharp edges in large discontinuity areas. However, a suitable smoothing degree is difficult to be found. Choosing a large smoothing strength leads to over-smoothing and foreground fattening. In contrast, less smoothing can result in noise on object surfaces. Adapting smoothing penalty according to edge locations relies strongly on edge extraction [Banz et al., 2012; Gong and Yang, 2005a]. Moreover, stereo methods without surface prior favor fronto-parallel planes instead of slanted surfaces.

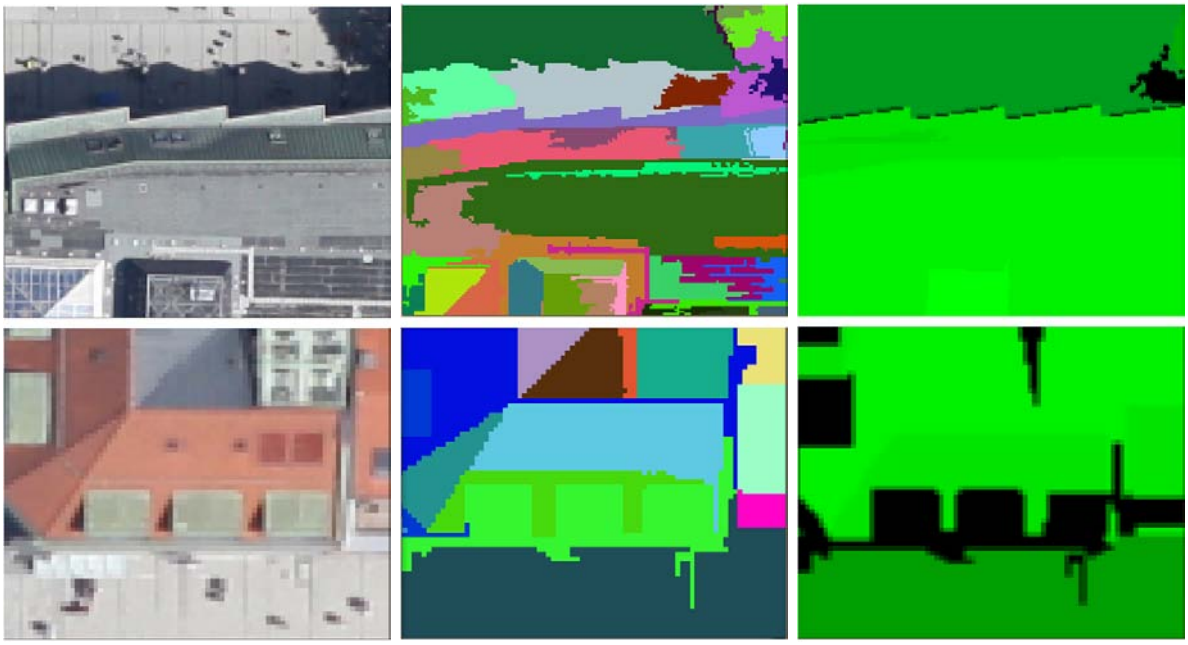
Object-surface prior can potentially contribute energy optimization for slanted surfaces by remaining discontinuity of object boundaries. But object surfaces are unknown. What we can obtain directly from input images are segments based on color/intesity similarity. Segments are limited in the 2D image domain and their boundaries can differ from the real object separation. Moreover, the segmentations of source and matching images are often not the same. In order to obtain object surfaces, pixel-wise disparities are fitted to segments such that many disparity planes are calculated. However, the main limitation of fitting planes is its high sensitivity to the segmentation initialization.

Figure 4.2 shows both positive and negative influences of initial subdivision by fitting disparity planes. Once a segment includes pixels only within an object surface, a plane-fit region can be reconstructed smoothly with sharp edges in discontinuity area (see the triangular building boundary). If segment contains pixels belonging to different object surfaces, any hard constrained region is completely failed (roof windows segmented with building facades). Methods [Klaus et al., 2006; Saygili, 2012; Taguchi et al., 2008; Wang and Zheng, 2008] employ energy minimization in a surface space as shown in Equation 2.16 are able to correct mis-matches once the color segmentation is wrong. Thus, almost all existing segment-based stereo matching methods require over-segmentation, which leads to very small segments. This limitation is contradictory to obtain reliable disparities within a segment for a robust plane fitting.

4.3 Confidence-Based Surface Prior

We calculate a disparity map Δ for $\mathcal{I}_s$ and denote the probability of $\Delta(p) = \delta_p$ being correct as $\mathrm{P}(\delta_p)$ where $\mathrm{P}(\delta_p) = \max\{\mathrm{P}(\delta) : \forall\delta\}$. Then we segment $\mathcal{I}_s$ into contiguous regions using the well-known mean-shift segmentation algorithm Comaniciu and Meer [2002] [1]. Using the disparity map Δ and the segmentation Seg_s, we fit a slanted plane, in disparity space, to each region in the segmentation. The plane fitting results in a dense disparity plane map for $\mathcal{I}_s$ that we will denote $\Delta^{pl} : \mathcal{I}_s \rightarrow \mathbb{Z}^{>0}$. Each pixel $p \in Seg_s^m$ has to belong to a disparity plane Π^m with

[1] Note that we do not require that this segmentation be an over-segmentation.



(a) Reference images. (b) Segmentation of (a). (c) Plane-fit disparity maps from (b).

Figure 4.2: Problem statement of using hard surface constraint. Mis-matches are indicated by solid black regions using consistence checking. **Top**: As the building roofs are clearly separated with the streets, hard surface constrain reconstructs sharp edges on object boundaries. **Bottom**: Roofs, building facades and part of streets in **(a)** are partitioned within a same segment in **(b)**. Using hard constraint leads to failed matching shown as solid black regions in **(c)**

$m \in [0, ..., |Seg_s|)$. The disparity of pixel p after plane fitting is denoted as d_p, which can differ from δ_p calculated from the initial pixel based matching process.

The plane fitting provides a unique assignment for each pixel. The goal of our work is to use this result as an additional surface prior for global frameworks in a probabilistic way. Thus we assume:

Assumption 4.3.1. $\forall p \in \mathcal{I}_s$: $\mathrm{P}(\Delta^{pl}(p)) \sim N(\mu = d_p, \sigma^2)$

Assumption 4.3.2. $\mathrm{P}(\Delta(p) = d_p) \ll \mathrm{P}(\Delta(p) = \delta_p) \Rightarrow \Delta(p) \notin \Pi^m$

Assumption 4.3.1 indicates that the probability of $\Delta^{pl}(p)$ is normally distributed at mean $\mu = d_p$ calculated by plane fitting. Assuming a normal distribution allows us to take advantage of its properties and make inferences from a hard planar constraint to a probabilistic surface prior. Assumption 4.3.2 builds the probability of $\Delta(p) \in \Pi^m$ according to a confidence observed from the initial pixel based matching. The confidence is obtained by comparison of $\mathrm{P}(\delta_p)$ and $\mathrm{P}(d_p)$. The standard deviation is then defined as:

$$\sigma = t(\mathrm{P}(d_p), \mathrm{P}(\delta_p)) \tag{4.1}$$

where t is a function such that $\sigma^2 \propto \log(\mathrm{P}(\delta_p)/\mathrm{P}(d_p))$. Recall that a high value for $\mathrm{P}(d_p)$ indicates that d_p is a good candidate match for p. Thus, since σ^2 is the variance, it will cause a sharply peaked Gaussian distribution with its maximum at $\mu = d_p$ when we are confident

that $\mathrm{P}(d_p)$ is a good candidate for the disparity of p, and a wider distribution when we are less confident.

Through the probabilistic interpretation of Δ^{pl}, the hard planar constraint is then relaxed as a soft prior according to the confidence obtained from a previous matching computation. More formally the probabilistic surface prior is then described in the energy space in the next section.

4.4 Energy Minimization via iSGM3

In this work the initial disparity maps for source and match images are calculated using SGM [Hirschmüller, 2008] respectively. After consistence checking, the plane-fit disparity map Δ^{pl} is generated by the voting-based plane-fitting algorithm proposed by Wang and Zheng [Wang and Zheng, 2008].

In energy space, we add a new surface prior, $S : \mathcal{I}_s \times \mathbb{Z}^{>0} \rightarrow \mathbb{R}$, to Equation 2.15:

$$\mathrm{E}(\Delta) = \sum_{p \in \mathcal{I}_s} C(p, \Delta(p)) + \lambda \sum_{\{p,q\} \in \mathcal{N}} V_{p,q}(\Delta(p), \Delta(q)) + \sum_{p \in \mathcal{I}_s} \kappa S(p, \Delta(p)) \tag{4.2}$$

where λ and κ are two user-defined constants to control their term strengths, respectively. Note that κ is individually estimated for each pixel using the confidence. The additional surface prior S favors disparities close to object planes. Our method is iterative that the confidence used for the surface prior can be updated.

Equation 4.2 is solved by a modified semi-global matching method, called iSGM3. Compared to SGM Hirschmüller [2008], our iterative iSGM3 aggregates the path-wise costs in one direction with three terms including data cost, smoothness penalty an additional surface cost function. This matching process can be repeated and mostly converges after a little loops.

4.4.1 Obtaining Reliable Disparities

A pre-matched disparity map is required to obtain a plane map Δ^{pl}. Conventional SGM formulated in Equation 2.21 is used to calculate disparity maps, δ_s and δ_m for source and match images respectively. The disparity maps contain outliers caused mostly by failed matching. For a reliable plane fitting, the correspondence checking or left-right checking is oft executed to gain a reliable disparity map [Gong and Yang, 2003].

$$\delta_s^{'}(p) = \begin{cases} \delta_s(p) & |\delta_s(p) - \delta_m(p + \delta_s(p))| < T \\ \mathrm{NaN} & \text{otherwise} \end{cases} \tag{4.3}$$

where T is a user defined threshold to tolerate the difference of two disparity maps. NaN indicates the matching at pixel p is failed that there is no label assigned.

4.4.2 Robust Plane Fitting using Voting

Plane fitting is already used for weakly-textured stereo scenes [Yang et al., 2008]. Matching in homogenous areas is prevented by highly ambiguous observations, i.e. the match costs are quite similar. Only using few reliable matching points, a dense plane for a texture-less region can be reconstructed. However, there are three main ill-defined assumptions for plane fitting: 1) Obtaining reliable match points; 2) Forcing surface to plane; 3) Determination of region area. Thus, we use the plane-fit disparity map only as an intermediate result for a rough approximation of the targeted surface map.

We express a plane Π_i^m [1] as the following function:

[1] Without loss of generality, the left image is considered as the reference image.

$$\delta_i(x, y) = a \cdot x + b \cdot y + c \tag{4.4}$$

where $i \in (L, R)$. (x, y) denotes the image coordinates of pixel p. $\delta_i(x, y)$ is the disparity of p. We assume each Π_i^m contains N_i^m reliable match pixels within a segment. Three plane parameters, a, b and c, need to be estimated.

We fit disparity maps in space according to segments. The plane parameters can be solved though Singular Value Decomposition [Shi and Tomasi, 1994]. Using the reliability of each disparity as weight, the normal vector of a disparity plane is given by the eigenvector belonging the minimum eigenvalue of matrix A.

$$A = \begin{bmatrix} \sum_{t=1}^{N} w_t \cdot x_t^2 & \sum_{t=1}^{N} w_t \cdot x_t \cdot y_t & \sum_{t=1}^{N} w_t \cdot x_t \cdot \delta_t \\ \sum_{t=1}^{N} w_t \cdot x_t \cdot y_t & \sum_{t=1}^{N} w_t \cdot y_t^2 & \sum_{t=1}^{N} w_t \cdot y_t \cdot \delta_t \\ \sum_{t=1}^{N} w_t \cdot x_t \cdot \delta_t & \sum_{t=1}^{N} w_t \cdot y_t \cdot \delta_i & \sum_{t=1}^{N} w_t \cdot \delta_t^2 \end{bmatrix} \tag{4.5}$$

For robust plane fitting, reliable depthes in disparity map are required. However consistence checking[1] eliminates outliers mainly in occlusion areas, mis-matches are not avoidable. Using the RANSAC method can remove outliers [Yang and Förstner, 2010], but the results rely on the randomly selected initial points. The comparison of Wang and Zheng [2008] shows the voting-based plane-fitting algorithm is more competent as the RANSAC method for fitting disparity maps.

Given a disparity map for $\mathcal{I}_i$ and the segmentation Seg_i, we fit a slanted plane, in disparity space, to each surface in the segmentation using the voting-based plane-fitting algorithm as outlined in Figure 4.4: The slant of a plane in x direction is obtained by calculation gradients, $\frac{\partial \delta_i}{\partial x}$, of all possible pixel pairs having the same x, row by row through a segment as shown in Figure. The calculated gradients are voted in a one-dimensional histogram, whose x-coordinate denotes all options of a and y-coordinate is the count amount of each option. A gaussian filter is executed to smooth the histogram to eliminate outliers, especially at $\frac{\partial \delta_i}{\partial x} = 0$. The maximum of the histogram is regarded as the final estimation of a. Similarly, b can be estimated by calculating $\frac{\partial \delta_i}{\partial y}$. Once a and b are obtained, according to Equation 4.4, we can estimate c in the same way using the given disparities $\delta_i(p)$.

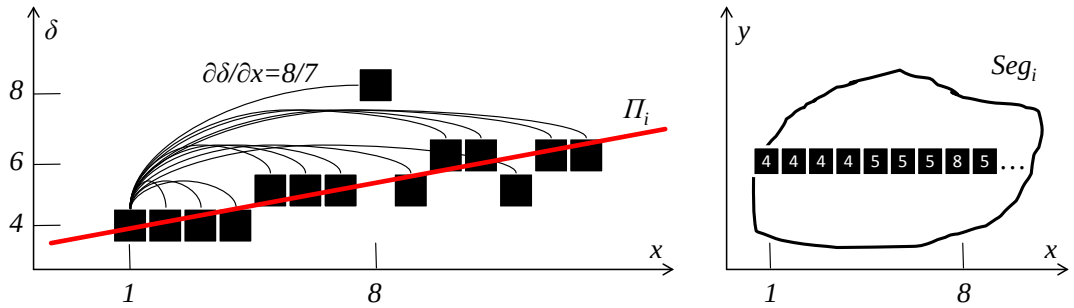


Figure 4.3: Illustration of calculating local gradients for plane fitting within a segment row. Red line denotes a cross-section of plane Π_i to be fitted. Local gradients are calculated using all possible reliable disparity pairs within the same row.

[1] Or Left-right checking eliminates outliers by comparison the disparity maps from left-right matching and right-left matching.

procedure REGIONBASEDPLANEFITTING(Seg_i, δ_i)
  repeat
    $(\{\frac{\partial\delta}{\partial x}\}, \{\frac{\partial\delta}{\partial y}\}) \leftarrow$ GradientCalculation(Seg_i^m, δ_i)
    $(h_x, h_y) \leftarrow$ HistogramVoting $(\{\frac{\partial\delta}{\partial x}\}, \{\frac{\partial\delta}{\partial y}\})$
    $(h_x^{'}, h_y^{'}) \leftarrow$ GaussianSmoothing (h_x, h_y)
    (a,b) $\leftarrow$ ParameterEstimation $(h_x^{'}, h_y^{'})$
    $h_c \leftarrow$ HistogramVoting $(a, b, \delta_i^{'})$ ▷ Eq. 4.4
    $h_c^{'} \leftarrow$ GaussianSmoothing (h_c)
    c $\leftarrow$ ParameterEstimation $(h_c^{'})$
  until M segments completed
  return $\{P_i\}$
end procedure

Figure 4.4: Process of region-based plane fitting using voting. Given a color segment and the disparity within it, Seg_i, δ_i, the proposed method estimates the horizontal and vertical slants, (a, b), of a spatial plane using histograms.

4.4.3 Iterative SGM3

Typical SGM minimization [Hirschmüller, 2008] is operated with a data term and a smoothing term. It approximates a global, 3D smoothness constraint by combining many 1D constraints from different aggregation directions for pixel-wise matching. To solve Equation 4.2, we introduce a modified semi-global matching method, called iSGM3. Compared to SGM [Hirschmüller, 2008], our iterative iSGM3 aggregates the path-wise costs in direction r with three terms including data cost, smoothness penalty and the additional surface cost function $S(p, \delta)$ as:

$$L_r(p, \delta) = C(p, \delta) + V_{p,p-r}(\delta, \delta') + S(p, \delta). \tag{4.6}$$

where $V_{p,q}$ is a truncated linear smoothing term with $V_{p,q}(\delta, \delta') = \min\{|\delta - \delta'|, \tau\}$. δ' denotes the disparity at q. The aggregated match costs in different directions, L_r with $r \in [0, 15]$, are then summed in $\text{E}_i(p, \delta)$:

$$\text{E}_i(p, \delta) = \sum_r L_r(p, \delta) \tag{4.7}$$

where i denotes the energy calculated at the i-th step. The initial energy at $i = 0$ can be calculated by Equation 2.15.

According to assumption 4.3.1, we denote the function $f(\delta) = N(\delta, \sigma(\delta))$. Using the confidence introduced in Equation 4.1, we define σ in the energy space as:

$$\sigma(\delta) = (\text{E}_{i-1}(p, \delta) - \text{E}_{i-1}(p, \delta_p) + \epsilon)^2 \tag{4.8}$$

where ϵ is a user-defined parameter to avoid a sharply peaked penalty. In the implementation, ϵ is chosen to be equal to 1.5. A quadratic function is selected over the confidence to smooth the penalties if $\text{E}_{i-1}(p, d_p)$ is very close to $\text{E}_{i-1}(p, \delta_p)$. The surface cost function is then defined as follows:

$$S(p, \delta) = (f(d_p) - f(\delta)) \cdot \frac{\text{E}_{i-1}(p, \delta_p)}{f(d_p)} \tag{4.9}$$

$S(p, \delta)$ penalizes the cost of a pixel belonging to an estimated plane according to its confidence.

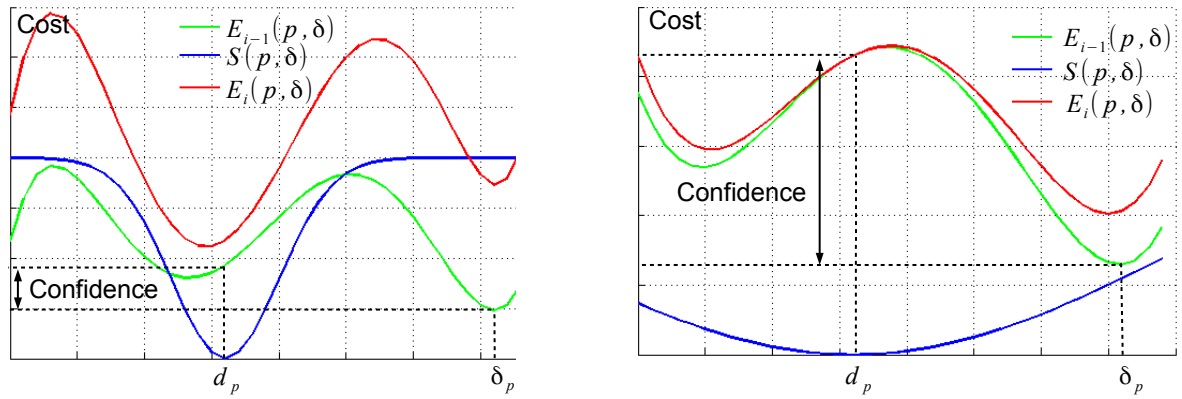


Figure 4.5: Cost fusion using surface prior: The green line denotes the previously computed costs for one within a disparity range. The blue line is the surface penalty S from the plane fitting processing. The fused cost is shown as the red line. **(a)** Good confidence: The disparity is shifted near to the plane-fit result. **(b)** Bad confidence: Surface penalties are overall similar. The disparity remains at the early position.

The proposed algorithm to employ the confidence-based surface prior is outlined in Figure 4.6. This algorithm employs an iterative feedback loop to incrementally improve the produced disparity estimates; in practice we observe convergence of the depth estimation in one to three iterations. For instance, we demonstrate our method in Figure 4.7: We fit the disparity map of (d) in space according to segments in (b). The plane-fit result is shown in (e) with observable effects of segments. The final disparity map is calculated in (f).

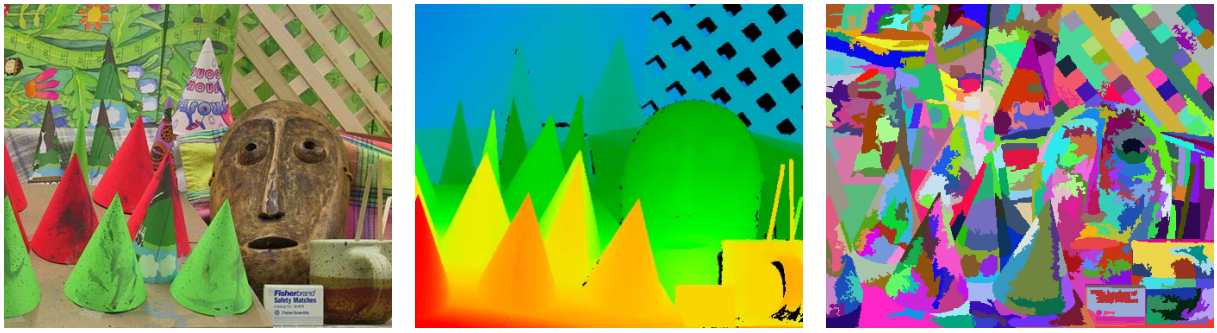
4.5 Summary

In this chapter the confidence-based surface prior for global energy minimization formulations is introduced. Given a dense disparity estimation we fit planes, in disparity space, to image segments. For each pixel, the probability of the depth of a pixel lying on an object plane is modeled as a Gaussian distribution, whose variance is determined using the confidence from a previous matching. A new disparity estimation with the addition of our confidence-based surface prior is than recalculated. This process can be then repeated. The confidence-based surface prior differs from existing surface constraints in that it varies the per-pixel strength of the constraint to be proportional to the confidence in our given disparity estimation. The global energy

procedure CALCDISPARITY($\mathcal{I}_L, \mathcal{I}_R$)
  $\text{Seg}_L \leftarrow \text{Segmentation}(\mathcal{I}_L)$
  $\text{Seg}_R \leftarrow \text{Segmentation}(\mathcal{I}_R)$
  $C_L, C_R \leftarrow \text{CalculateMatchCosts}(\mathcal{I}_L, \mathcal{I}_R)$
  $(\Delta_L, \text{E}'_L) \leftarrow \text{SGM}(C_L, \mathcal{I}_L \text{ as } \mathcal{I}_s)$ ▷ Eq. 2.15
  $(\Delta_R, \text{E}'_R) \leftarrow \text{SGM}(C_R, \mathcal{I}_R \text{ as } \mathcal{I}_s)$ ▷ Eq. 2.15
  repeat
    $\Delta'_L, \Delta'_R \leftarrow \text{LeftRightConsistency}(\Delta_L, \Delta_R)$
    $\Delta^{pl}_L \leftarrow \text{RegionBasedPlaneFitting}(\text{Seg}_L, \Delta'_L)$
    $\Delta^{pl}_R \leftarrow \text{RegionBasedPlaneFitting}(\text{Seg}_R, \Delta'_R)$
    $S_L \leftarrow \text{CalculateConfidenceConstraint}(\Delta^{pl}_L, \text{E}'_L)$
    $S_R \leftarrow \text{CalculateConfidenceConstraint}(\Delta^{pl}_R, \text{E}'_R)$
    $(\Delta_L, \text{E}'_L) \leftarrow \text{iSGM3}(C_L + S_L, \mathcal{I}_L \text{ as } \mathcal{I}_s)$ ▷ Eq. 4.2
    $(\Delta_R, \text{E}'_R) \leftarrow \text{iSGM3}(C_R + S_R, \mathcal{I}_R \text{ as } \mathcal{I}_s)$ ▷ Eq. 4.2
  until N iterations completed
  $\Delta'_L, \Delta'_R \leftarrow \text{LeftRightConsistency}(\Delta_L, \Delta_R)$
  return $\{\Delta'_L, \Delta'_R\}$
end procedure

Figure 4.6: Proposed iterative algorithm. Given a stereo image pair, $\{\mathcal{I}_L, \mathcal{I}_R\}$, the introduced method calculates an initial disparity map. The resulted match costs are used as the confidence of the plane-fit process. iSGM3 aggregates the previous match costs and generate iteratively disparity maps.

minimization with three priors – the data, smoothness and surface prior, is computationally solved by the algorithm – iterative Semi-Global Matching with 3 terms (iSGM3), which inherits the (efficient) path-wise cost aggregation from SGM and adds the confidence-based surface prior into the framework. Unlike many region-based methods, the iSGM3 method defines the energy formulation over pixels, instead of regions in a segmentation; this results in a decreased sensitivity to the quality of the initial segmentation, especially in texture-rich regions, where an object surface is oft over-segmented. In contrast, iSGM3 trends to handle homogenous regions as an integral part, where pixel-wise matching costs are very similar. Generally the introduced method has three main benefits for solving dense stereopsis: sharp object-boundary edges in areas of depth discontinuity; accurate disparity in surface regions; and low sensitivity to the initial color segmentation.



(a) Reference image (b) Ground truth. (c) Segmentation of (a)

(d) SGM result (e) Plane-fit disparity map of (d) (f) iSGM3 result

Figure 4.7: Illustration of processing steps of iSGM3: Initial disparity map (d) is fitted to segments of (c) in order to compute a plane map (e). The pixel-wise confidence is generated by comparing energies at disparity levels (d) and (e). (f) is the result using iSGM3.

Chapter 5

Results

As introduced in Scharstein and Szeliski [2002] the performance of a global dense stereo matching method depends on several factors including match costs, energy minimization and post-processing. In this chapter we evaluate match cost functions and the confidence-based surface prior respectively in order to focus on only one factor at a time. In contrast to many existing works, our evaluation does not only use the Middlebury online benchmark, but also the Middlebury data sets with radiometric changes and more challenging remote sensing data.

The evaluation of all match cost functions used is executed using the identical energy minimization framework for a consistent and fair comparison. Both intensity and gradient images are utilized as input data. In total, ten match cost functions as shown in Table 3.2 including absolute difference, mutual information, census and their combinations are investigated. Masks for occluded and discontinuity areas are applied to observe matching performances in different regions. The robustness of a match cost by changing illuminations and exposure time within a stereo pair is discussed using all 27 Middlebury data sets. The airborne image sequence allows us to compare match costs when the baseline length increases. We study the interdependencies among matching performance, cost functions, and observation conditions. Based on this study, we recommend a combination of parametric and non-parametric match costs, especially for real-world data.

The procedure to evaluate the confidence-based surface prior is developed in order to demonstrate three merits when using it in a global stereopsis method: more completeness on low texture regions, less sensitivity to a given segmentation, and sharp edges on object boundaries. We compare results with and without our confidence-based prior using the similar energy optimization framework and show that even on the regions where the segmentation is incorrect, our method does not produce any plane-fit artifacts. Both the Middlebury benchmark and the airborne stereo pairs are used. Because the lower resolution of the LiDAR groundtruth and temporal reforming of buildings of the utilized airborne data sets, we manually generate the building-edge segments and compare the disparity changes between two edge sides. This analysis shows the superior performance of our surface prior on object boundaries.

Finally, we summarize the evaluation of match costs and the surface prior together in order to state the problems of current stereo methods for challenging data. Based on the evaluations, we propose that robust match costs and additional priors are two fundamental components to improve match performance.

5.1 Data Sets Used

Three types of data are used for both a qualitative and quantitative evaluation in this dissertation. The de-factor standard Middlebury data sets including 30 stereo pairs with radiometric changes provide dense ground truth and allow comparison with other methods. The airborne image sequence considers continually captured urban areas and create several stereo pairs with an increasing baseline length. A LiDAR point cloud is applied as ground truth. The satellite stereo pairs are configured with a large stereo angle, which makes matching especially difficult. This section introduces these data sets and their observation conditions. Moreover, the advantages and limitations of each data set are highlighted.

5.1.1 Middlebury Stereo Benchmark

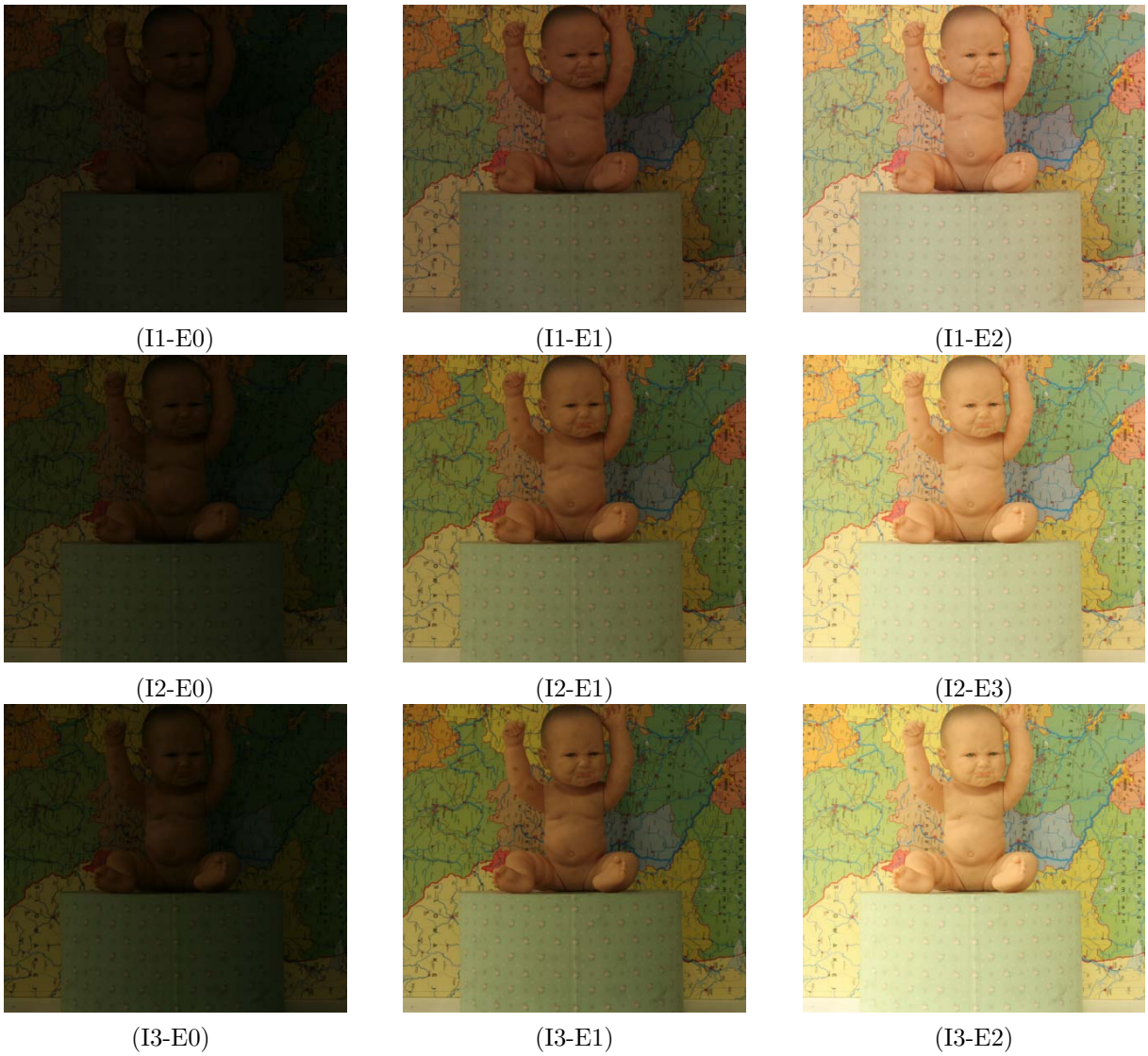
Scharstein and his colleagues have made rectified stereo pairs with ground truth available, which have become the standard stereo-vision benchmark for the computer vision community [Scharstein and Szeliski, 2002]. These data sets consist of several indoor scenes captured with regularized exposures and controlled light sources. The disparity range of a Middlebury stereo pair is between 16 to 64 pixels. As shown in Figure 5.1, each ground truth disparity map is masked to provide three different regions for quantitative analysis: discontinuities (disc), non-occluded regions (nocc), and everything (all).

(a) Reference image (b) Mask **all** (c) Mask **nocc** (d) Mask **disc**

Figure 5.1: **Teddy** from the Middlebury online benchmark. Three masks are applied for the evaluation: **all** for every available ground truth disparity; **nocc** for non-occluded areas and **disc** for discontinuous on object boundaries.

The Middlebury 2002 data sets (two stereo pairs) along with the **Venus** pair from the Middlebury 2001 data set and the **Tsukuba** data set are the four image pairs currently being used as the online benchmark which is widely used for developing stereo methods in the computer vision community. Compared with the four stereo pairs without any radiometric changes, the Middlebury 2005 and 2006 data sets including 27 stereo scenarios in total are captured in a controlled environment for structured light reconstruction [Scharstein and Szeliski, 2003]. Each scene is captured with three levels of illumination and exposure configuration, thus nine stereo pairs under different radiometric conditions are created. These data sets consider a variety of objects with large homogenous areas, slanted surfaces, geometric gadgets and large discontinuous object boundaries. However, lengths of baselines applied by the Middlebury data are relative short and the camera moves almost parallel to the observed scene. The radiometric changes are not natural – the lighting positions for each illumination level are not changed. Figure 5.2 illustrates a sample for nine exposure-illumination combinations. Figure A.1 summarizes all reference images of the 30 scenarios used in our work.

One of the advantages of the Middlebury data sets is their dense accurate ground truth. In contrast to the LiDAR point cloud used for airborne images, the ground truths of Middlebury



(I1-E0) (I1-E1) (I1-E2)

(I2-E0) (I2-E1) (I2-E3)

(I3-E0) (I3-E1) (I3-E2)

Figure 5.2: **Baby1** from the Middlebury data sets. Levels are labeled from 1 to 3 for illumination and 0 to 2 for exposure time.

have the same resolutions as their input images. A pixel-to-pixel comparison allows accurate evaluation on object boundaries and surface discontinuities. However, the data are captured in artificial environments, have short baselines and contain simple scenarios. All of these factors motivate us to use remote sensing data sets introduced in next subsections.

5.1.2 DLR 3K Data Sets

Two continuously recorded airborne optical image sequences with known geometry are used in our evaluation. Both sequences cover similar urban areas (Munich center) and have almost the same flight altitude, approximately 1.5 kilometers above ground. A 1.7-meter-resolution LiDAR 3D point cloud was acquired in 2005 and is used as ground truth.

The image sequence from the 2007 flight campaign was taken by Canon EOS 1D Mark II cameras with a 50-centimeter lens from the 3K camera system [Kurz et al., 2007]. The baseline

between following images is about 35 meters. Each sequentially recorded image is matched with the same master image so that eight stereo pairs observing the same location are built with an increasing baseline length. The largest baseline we present in this work is about 250 meters as shown in Figure 5.1. The original images are down-sampled by a factor of two to reduce the inaccuracy of camera calibration. This data set is used to evaluate match cost functions to follow the impacts of cost functions on the matching performance.

Stereo pair	M-1	M-2	M-3	M-4	M-5	M-6	M-7
Baseline length (m)	35	70	105	140	175	210	245

Table 5.1: Baseline length of 3K sequence.

An updated image sequence over the similar area is recorded by the 3K+ camera system in 2011, which consists of three Canon EOS 1D Mark III cameras. Also only images from the nadir views are used for stereo matching. The images are sharper than the 2007 data. We use this data set for the evaluation of our confidence-based surface prior with focus on the building boundaries. Building line segmentations are manually generated from the input images for each stereo pair. The reason why we have not used the 2011 data for the evaluation of match cost functions is that many buildings are undergoing reconstruction, leading to large differences between the 2005 LiDAR and 2011 optical data sets. Figure 5.3 demonstrates two stereo pairs with short and large baseline lengthes respectively.

The used airborne imagery captures an urban area in the city center of Munich, Germany. High buildings, narrow streets and homogenous roofs are covered. The images are taken in JPEG format to achieve a high frame rate. The image size is 5616×3744 pixels. RGB channels are merged to intensity for matching.

The stereo pairs with increasing baseline length impair the radiometric consistence between reference and match images. The local structures change when baseline length increases little by little. These allows us to study the challenges from real-world data for stereo matching. The evaluation using LiDAR points is limited by two factors: the low resolution of our LiDAR ground truth and the temporal new reconstructions after recording in 2005. We triangulate the LiDAR points and calculate the average distance between reconstructed DEM points onto the mesh surfaces. Line segmentations on building boundaries are manually selected and only the lines on non-facade sides are applied to observe the matching performance in large discontinuity regions.

5.1.3 Satellite Stereo Pairs

Additionally, we evaluate the matching costs on a Worldview-1 stereo image pair with a ground sampling distance of 50 cm, and a relatively large stereo angle of 35°. A 3D point cloud acquired by the Institut Cartogràfic de Catalunya (ICC) with airborne laser scanning is used as reference data. The density of the point cloud is approximately 0.5 points per square meter. The data is part of the ISPRS matching benchmark [Reinartz et al., 2010].

The challenges for the satellite stereo pair are the lower resolution and the extremely large stereo angle. These prevent both the radiometric consistence and the local structure similarity within a stereo pair. We demonstrate the importance of robust match cost for stereo matching on real-world data.

(Short-L)

(Short-R)

(Long-L)

(Long-R)

Figure 5.3: Stereo pairs of 3K airborne images with short baseline (70m) and long baseline (210m).

5.2 Evaluation of Matching Cost Functions

The performance of a dense stereo matching method depends on all components including pre-processing, matching costs, energy minimization, disparity optimization and post-processing. All dense stereo matching algorithms use matching cost functions to measure the similarity of image locations. The most intuitive matching cost is absolute difference, which assumes that corresponding pixels have the same intensity [Kanade, 1994]. This conventional technique can only work with the Lambertian assumption. In a real-world scenario, good radiometric conditions are prevented by lighting geometry, illumination, camera configuration, material of reflecting surfaces, and so forth [Kumar et al., 2011]. The influence of the applied data on the matching performance using different costs is not investigated in the previous evaluation works [Hirschmüller and Scharstein, 2009; Neilson and Yang, 2011; Scharstein and Szeliski, 2002].

In recent years, many stereo matching methods are developed and evaluated using the Middlebury stereo benchmark without radiometric changes [Scharstein and Szeliski, 2002]. This test bed includes several close-range indoor stereo pairs with ground-truth disparities and provides a quantitative evaluation. The data used for the online ranking include only four stereo pairs with relatively small baselines and the same radiometric configuration for each stereo pair [Scharstein and Szeliski, 2002]. Generally, remote sensing data are not addressed when developing stereo matching methods in the previous works, probably due to the limitation of ground-truth in the

computer vision community.

All of the above reasons motivate us to investigate the performance of match costs on both close range and remote sensing data sets, so that several general guidelines as shown in Chapter 6 can be introduced for developing robust stereo methods in real-world applications. The main highlights of the evaluation of match costs are:

- Investigation on the relationship between matching performance, matching costs and radiometric conditions like length of stereo baseline and non-Lambertian reflectance.
- Comparison of the matching performance using various matching costs in different areas like homogenous and discontinuity regions.
- Improving matching performance by merging different matching costs with depending on radiometric conditions and observation regions.

5.2.1 Methodology for Evaluation

The match cost functions in the proposed evaluation are defined on intensity and gradient images. RGB color channels are averaged, if they are available. The gradient images are computed in epipolar-line direction (flying direction). The SGM matching algorithm is used for evaluation of all match costs.

Moreover, matching performance depends also on the parameters selected. For a reliable evaluation, we tuned all parameters for each cost function. In addition, the weight parameters used in the cost merging formulations are tuned from 0 to 1 with 0.1 steps. This tuning allows concentrating on the performance of matching costs rather than the stereo method. Each data set, the parameter configuration is kept constant for all stereo pairs. We published the used parameters to allow replication of our experiments. Note that, in the 2005 and 2006 Middlebury data sets, each view of a stereo scene is captured with three different exposures and under three different illuminations, so that 9×9 depth maps can be calculated for one stereo scene.

During the evaluation on the Middlebury data sets, we observed that the same cost function can perform significantly differently on some stereo scenarios, even when the parameter setting remains constant. Thus, following the Middlebury online benchmark, we rank cost functions for each scene respectively and average all ranks together, such that the influence from outliers is minimized.

For the remote sensing data, after stereo matching, the points are projected into UTM Zone 32 North for the aerial images and UTM Zone 31 North for the satellite images respectively to generate Digital Surface Models (DSMs). Holes in the generated DSMs are filled with inverse distance weighted interpolation. We compute the Euclidean distance d between points in ground truth and the triangulated DSM. The percentage of pixels with a distance higher than a threshold is defined as bad-pixel. As the DSMs generated by stereo matching, and to a less extent the points cloud might contain outliers which violate the assumption of a normal distribution, we follow Höhle and Höhle [2009] and compute measures based on robust statistics, in addition to the classical mean and standard deviation values. This includes the normalized median absolute deviation (NMAD):

$$NMAD = 1.4826 \times median_j(|d_j - median_i(d_i)|) \tag{5.1}$$

as a robust estimate of the standard deviation.

5.2.2 Results on Middlebury Data Sets

In this subsection the results using different match cost functions using the Middlebury stereo data are evaluated. Results without and with radiometric changes are respectively analyzed and discussed. Using stereo pairs without radiometric changes, we can focus more on the influence of object features on matching performance. In contrast, the influence of observation constraints like illumination and exposure can be better discussed when applying data sets with radiometric changes.

5.2.2.1 Results on Data without Radiometric Changes

Table 5.2 shows the relative ranking of match cost functions applied on the Middlebury data sets without radiometric changes. The ranks are listed according to the average of all individual rankings of each stereo scene. The merged match cost function, mutual information and census using gradient images (MIC_{grad}), reaches the best rank. In non-occluded areas, matching using gradient images performs better than using intensity images. The ranking tables shows, match-costs merging can improve the performance in general. Under the group of individual match costs (costs without merging), census (Cen_{int}) outperforms absolute difference (AD_{int}) and mutual information (MI_{int}). The relative ranks are not in the same proper order as the average bad-pixel percentages, because different match costs perform significantly differently on stereo pairs when large homogenous areas are included. For example, on **Plastic** from Middlebury 2006, the bad-pixel percentage using gradient-based absolute difference is 27%, in contrast, census achieves only 43%. Thus the averaged bad-pixel percentage can not express the common performance on all data sets.

Matching cost	Abb.	P_1	P_2	w	rank	bad-pixel (%)
$C_{MI+Cen}(\bigtriangledown_e)$	gC_{MIC}	450	1650	0.4	1	6.36
$C_{AD+Cen}(\bigtriangledown_e)$	gC_{ADC}	250	850	0.6	2	6.44
$C_{Cen}(\bigtriangledown_e)$	gC_{Cen}	400	850	-	3	7.59
$C_{MI+Cen}(\mathcal{I}^{3\to1})$	iC_{MIC}	300	1200	0.1	4	7.65
$C_{Cen}(\mathcal{I}^{3\to1})$	iC_{Cen}	300	1500	-	5	7.60
$C_{AD+Cen}(\mathcal{I}^{3\to1})$	iC_{ADC}	200	700	0.4	6	7.72
$C_{MI}(\bigtriangledown_e)$	gC_{MI}	600	1700	-	7	9.0
$C_{AD}(\bigtriangledown_e)$	gC_{AD}	100	350	-	8	6.7
$C_{MI}(\mathcal{I}^{3\to1})$	iC_{MI}	500	1700	-	9	19.26
$C_{AD}(\mathcal{I}^{3\to1})$	iC_{AD}	10	30	-	10	22.97

Table 5.2: Table of the relative ranking of different matching cost functions with applied parameter setting. The last column presents the average bad-pixel percentages of all 30 scenes.

Matching cost	**Baby2**	**Lampshade2**	**Wood2**
iC_{AD}	19.13%	59.89%	48.37%
gC_{AD}	3.13%	7.01%	1.86%
iC_{Cen}	5.21%	6.4 %	2.62%

Table 5.3: Comparison of AD_{int}, AD_{grad} and Cen_{int} on data sets with homogenous areas. **Baby2**, **Lampshade2** and **Wood2** denote the selected stereo scenes.

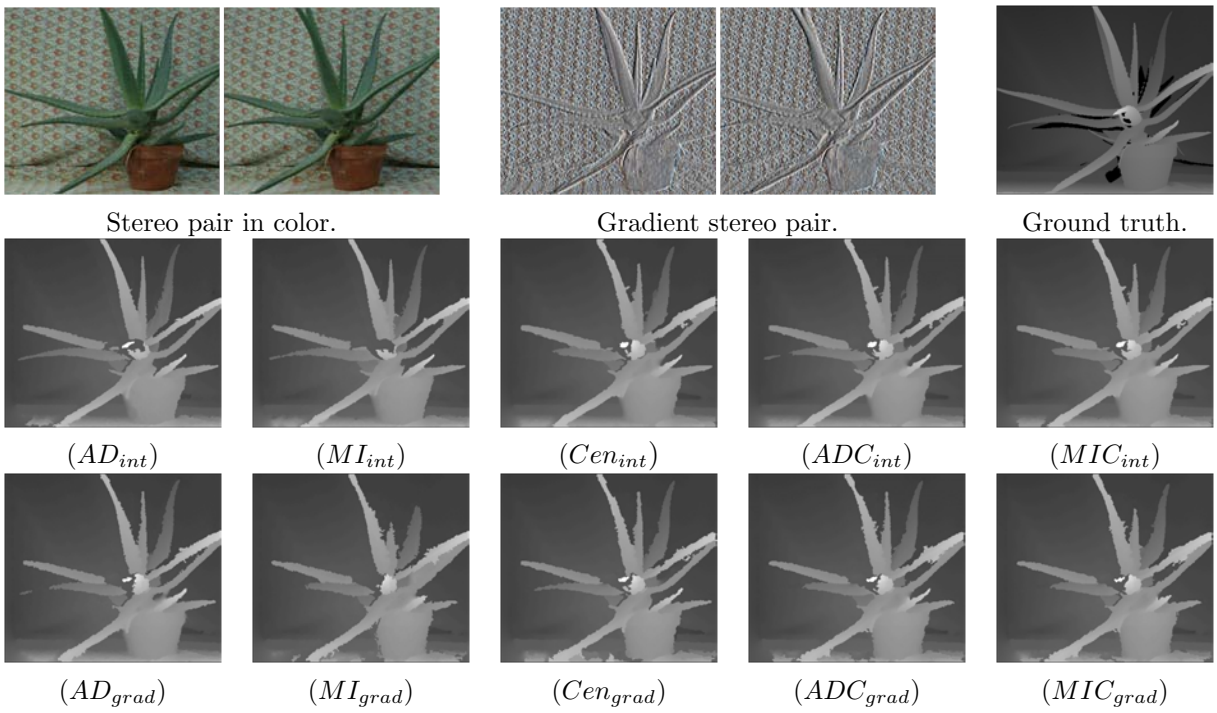


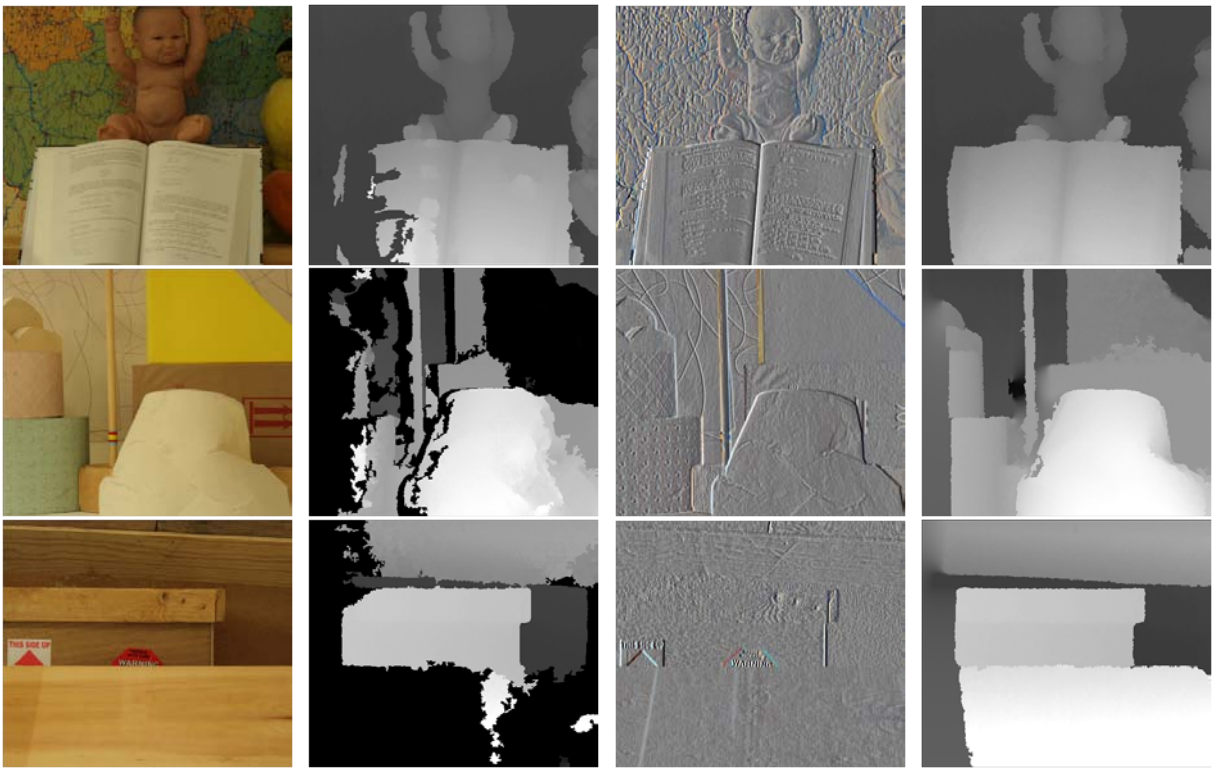
Figure 5.4: Results comparison on **Aloe2** stereo pair without radiometric changes: Absolute Difference (AD), Mutual Information (MI), Census (Cen) as well as their combinations (ADC and MIC) are applied both on intensity images and gradient images. All cost functions perform similarly. Matching using gradient images causes rugged object boundaries in discontinuity areas.

Matching cost	**Tsukuba**	**Venus**	**Teddy**	**Cones**
iC_{AD}	16.19 %	17.62%	24.76 %	14.06%
gC_{AD}	17.96 %	19.37%	26.48%	16.45%
iC_{MI}	13.41%	18.13%	24.08%	11.91%
gC_{MI}	17.01 %	18.9%	26.73 %	15.05%
iC_{Cen}	16.09 %	27.47%	24.42%	16.69%
gC_{Cen}	18.05%	27.66%	26.73%	19.59%

Table 5.4: Comparison of cost functions using intensity and gradient images applied to Middlebury online benchmark in the discontinuity areas (DISC). Consistently for each cost function, the results applied to intensity images outperform the results applied to gradient images clearly.

We select the stereo scene, **Aloe**, to show the results using different cost functions without radiometric changes in Figure 5.4. All cost functions perform well and similarly, both using intensity and gradient images. Qualitatively, gradient based absolute difference (AD_{grad}) reaches the best performance and outperforms the worst cost function in that case, gradient based census (Cen_{grad}), only by 2%. This slight performing difference of match costs appears almost by each data set from the Middlebury online benchmark. Thus, the importance of cost functions for developing stereo methods is often ignored when focusing only on data captured in a constantly radiometric environment.

Compared with the results on the Middlebury online benchmark, a higher difference in



(a) Reference images (b) Results of (a) (c) References images (d) Results of (c)

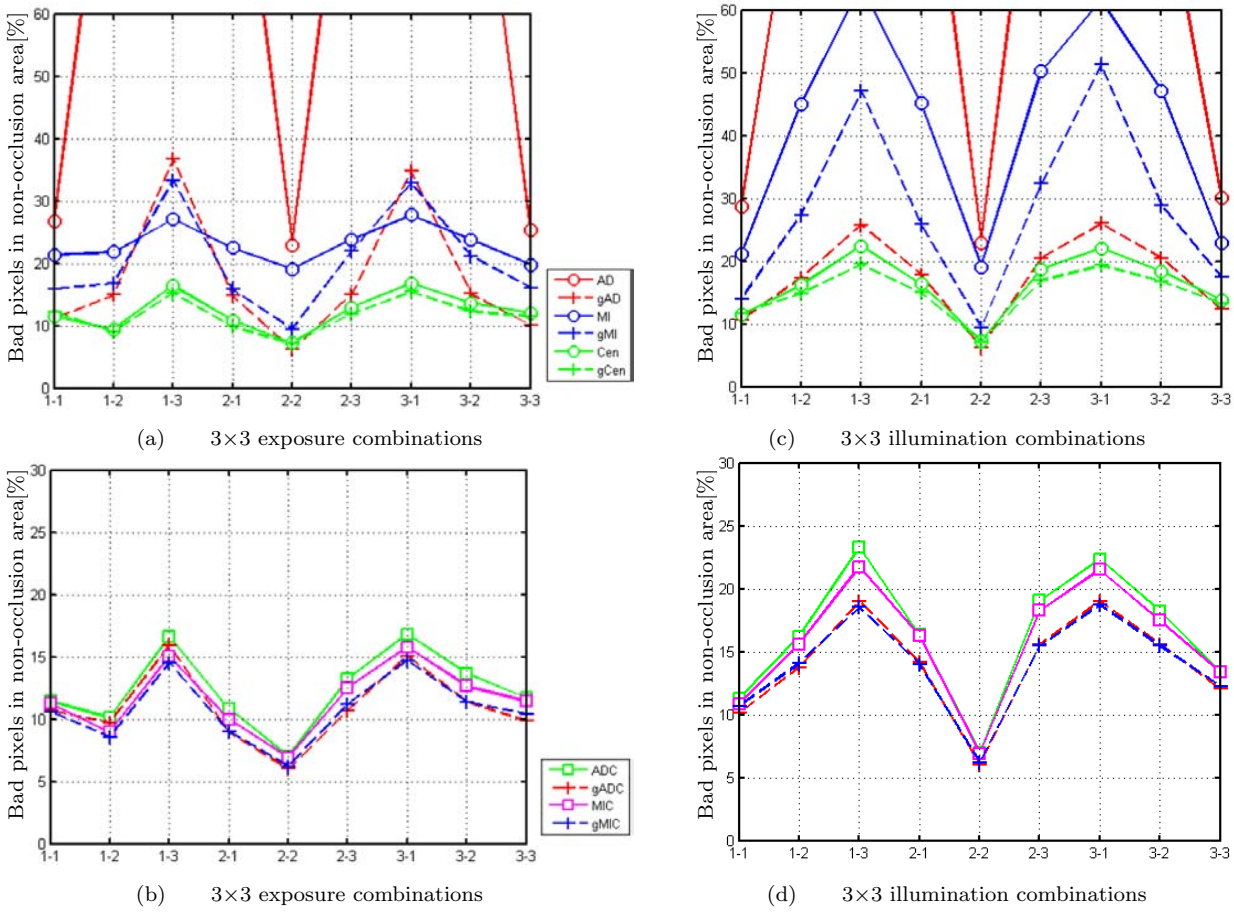
Figure 5.5: Results comparison of absolute difference using intensity and gradient images. **Baby2**, **Lampshade2** and **Wood2** are represented in color and as gradient images in the first and third column. In the areas with less texture, results using AD_{int} presented in the second column are less complete then AD_{grad} presented in the fourth column.

matching performance between cost functions is measured when using the Middlebury 2006 data sets. We selected three stereo scenes with less texture, **Baby2**, **Lampshade2** and **Wood2**, to show the impact of challenging data on the matching behavior. Figure 5.5 illustrates the results using absolute difference on intensity and gradient images (AD_{int} and AD_{grad}). The qualitative comparison is shown in Table 5.3. In the homogenous areas, matching using gradient images is more complete than using intensity images due to Non-Lambertian reflectance on surface. Gradient images reduces the influence of Non-Lambertian reflectance and improves the matching robustness, especially for parametric matching costs like absolute difference. In contrast, non-parametric matching costs like census perform constantly in the areas with less texture. It seems that the local features are less changed in intensity and gradient image.

We note that results using intensity images have sharper edges then using gradients, if they are successfully matched. Table 5.4 illustrates the comparison between matching costs using intensity and gradient images applied to the Middlebury online benchmark in the discontinuity areas. For each cost function, the results using intensity images outperform the results using gradients, because derivation gradients causes the location shifting of edges on the object boundaries. This appearance prevents matching using gradient images for remote sensing data, especially for the urban areas, where sharp building boundaries are expected.

5.2.2.2 Results on Data with Radiometric Changes

The Middlebury 2005 and 2006 data sets are captured with three different levels of exposure time under three different illuminations. Thus, each view of a scene has nine different images that exhibit significant radiometric differences. We apply the evaluation over all 9×9 combinations of either exposure-time and light source for each stereo scene. Figure 5.7 illustrates six captures of **Dolls** under three different illuminations with the same exposure and with three different exposures under the same illumination in color and derived in gradient.



(a) 3×3 exposure combinations

(c) 3×3 illumination combinations

(b) 3×3 exposure combinations

(d) 3×3 illumination combinations

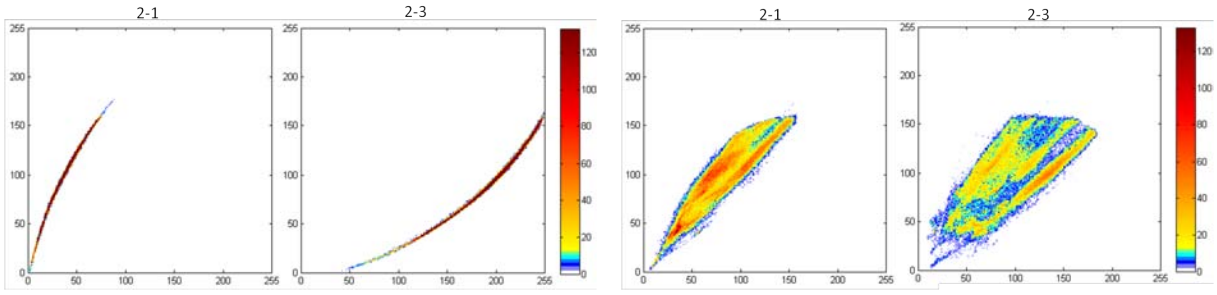
Figure 5.6: Result analysis on stereo pairs using different combinations of illuminations or exposures. The number notation $l - r$ on the x axis denotes the illumination and exposure combination between left and right view. The y axis denotes the percentages of bad-pixel. (a) and (b) summarize the case under the same illumination with different exposures using individual matching costs ($C_{AD}(\mathcal{I}^{3\to1})$, $C_{AD}(\bigtriangledown_e\mathcal{I})$, $C_{MI}(\mathcal{I}^{3\to1})$, $C_{MI}(\bigtriangledown_e\mathcal{I})$, $C_{Cen}(\mathcal{I}^{3\to1})$ and $C_{Cen}(\bigtriangledown_e\mathcal{I})$) and merged matching costs ($C_{AD+Cen}(\mathcal{I}^{3\to1})$, $C_{AD+Cen}(\bigtriangledown_e\mathcal{I})$, $C_{MI+Cen}(\mathcal{I}^{3\to1})$ and $C_{MI+Cen}(\bigtriangledown_e\mathcal{I})$) respectively. Results for different levels of illumination with the same exposure are shown in (c) and (d).

For a clear discussion, we classify our evaluations into two groups, one, the 3×3 exposure combinations under the same illumination and the other, the 3×3 light-source combinations with the same exposure time. The total matching error is calculated by averaging bad-pixel of all 27 sets from the Middlebury 2005 and 2006 data sets. Figure 5.6 shows the error analysis of different costs in this two groups: The individual match costs (AD_{int}, AD_{grad}, MI_{int}, MI_{grad},

Cen_{int} and Cen_{grad}) are illustrated in (a) and (b) as well as the merged match costs (ADC_{int}, ADC_{grad}, MIC_{int} and MIC_{grad}) in (c) and (d).

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

Figure 5.7: **Dolls** with illumination and exposure varyings in color and gradient variants. The images in the first three columns have been captured with the same exposure but under different lighting conditions. The right images in the last three columns are captured with three different exposures under the same illumination.



(a) Different exposure, intensity (b) Different illumination, intensity

(c)Different exposure, gradient (d) Different illumination, gradient

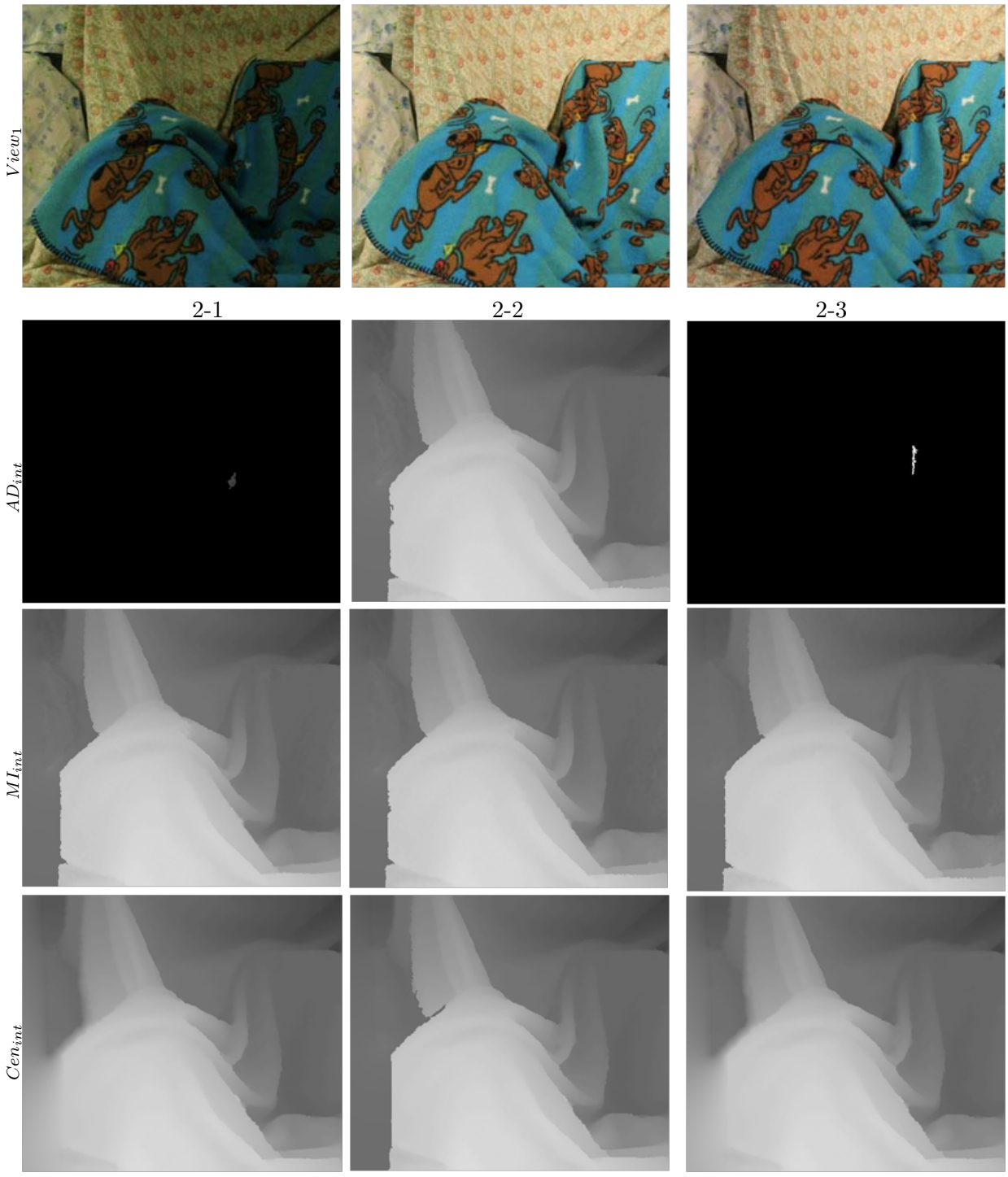
Figure 5.8: Joint histograms of **Cloth3**. **(a)** Joint histograms of images captured with different exposures under the same illumination. **(b)** Joint histograms of images captured with the same exposure under different illuminations. **Top:** Joint histograms generated by intensity images. **Bottom:** Joint histograms generated by gradient images. The notation $2-x$ denotes an exposure combination of two images.

Generally, census shows the highest robustness against variations of radiometric configuration. For stereo pairs whose left and right images are captured with the same exposure and illumination, census performs consistently well. Parametric match costs like absolute difference fail completely when using intensity images with radiometric changes. However, gradient images reduce the radiometric variations within a stereo pair so that the matching performance using

parametric costs can be obviously improved. In contrast, a similar improvement for census using gradient images is not measure due to robust local structures, which does not disappeare by generation of gradient image. Mutual information outperforms absolute difference clearly, but it fails mostly during changing light-sources. As shown in Figure 5.6 (a) and (c), the performance variation using mutual information under exposure changes is much less as under illumination changes, because changing light source causes local reflectance changes for different surfaces under different illuminations. A tangible improvement via applying gradient images is also observed for mutual information. Note that all merged matching costs perform similarly because all of them combine census in their formulations.

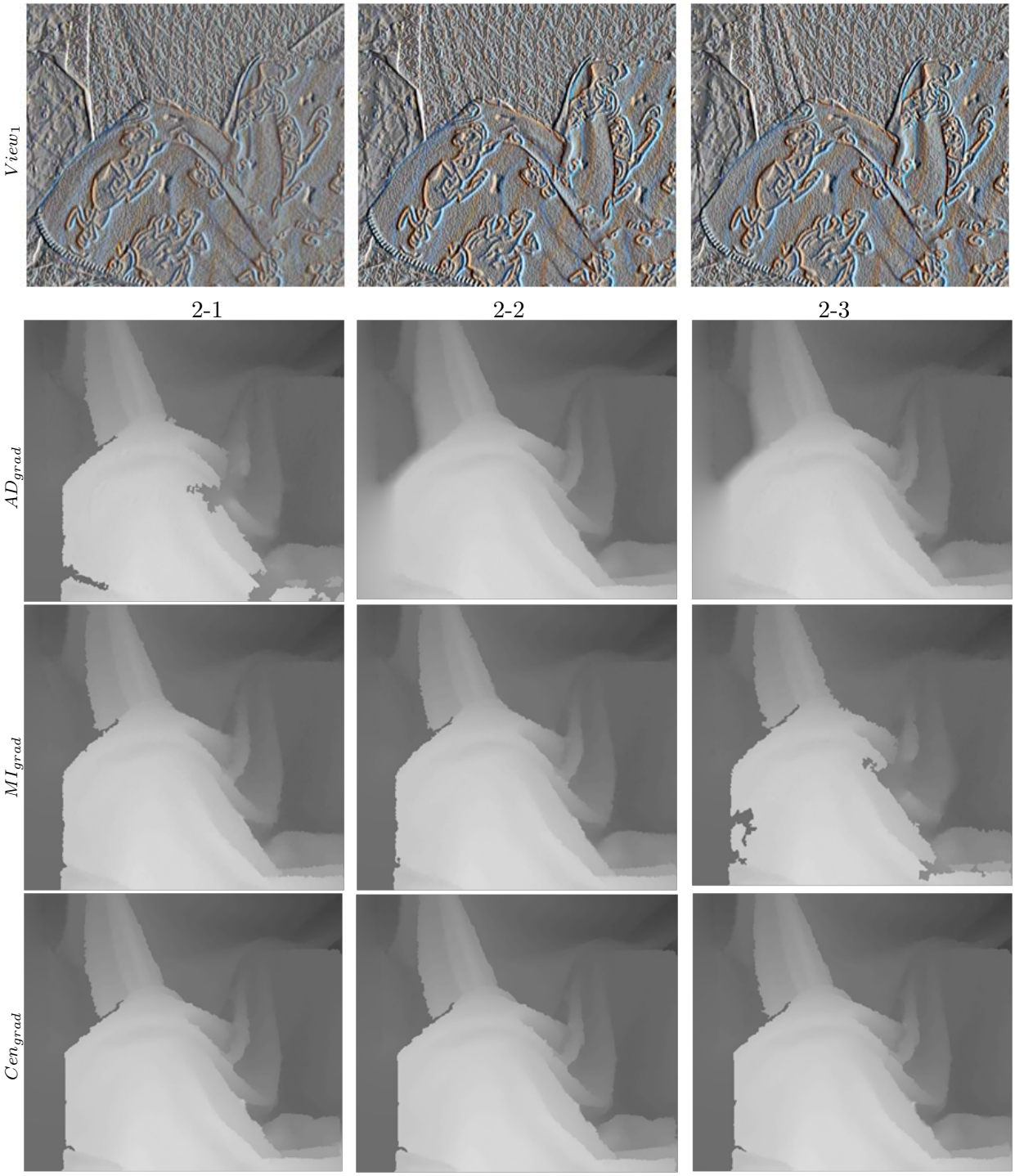
Radiometric changes of the Middlebury data are caused by either different exposures or varying illuminations. Which one of them is more challenging for stereo matching? Using **Cloth3**, we illustrate the joint histograms of the reference image captured with different exposure settings (Figure 5.8 (a)) or varying illuminations (Figure 5.8 (b)). The gradient images show consistent behaviors of the joint histograms that the variances of corresponding gradients diagonally distribute. The joint histogram of intensity image by changing exposure levels shows a little variance, however the diagonal line is bend to a curve. Match costs based on statistical analysis like mutual information should handle this situation well. The most challenging data are intensity images captured with different illumination levels. For example, Figure 5.9 and Figure 5.10 compare absolute difference, mutual information, and census with different exposures and under varying illuminations using **Cloth3**. In the intensity space, changing exposures causes a whole shifting of a joint histograms as a curve crooked in luminance-changed direction. In contrast, a modification of a joint histogram due to illumination changes is more difficult to forecast. The radiometric changes due to changing exposures can be dealt with mutual information effectively, while the matching using mutual information failed mostly on illumination-changed stereo pairs. In the gradient space, the joint histograms distribute nearby the diagonal lines by changing both exposures and illumination. The matching using absolute difference and mutual information in gradient domain can be achieved successfully in the most regions of the stereo pair. In both intensity and gradient spaces, census shows a consistent robustness.

In this subsection match costs are evaluated and compared using the Middlebury stereo sets with radiometric changes. In contrast to the results shown in Subsection 5.2.2.1, this part of evaluation demonstrates more performing difference using different match costs.



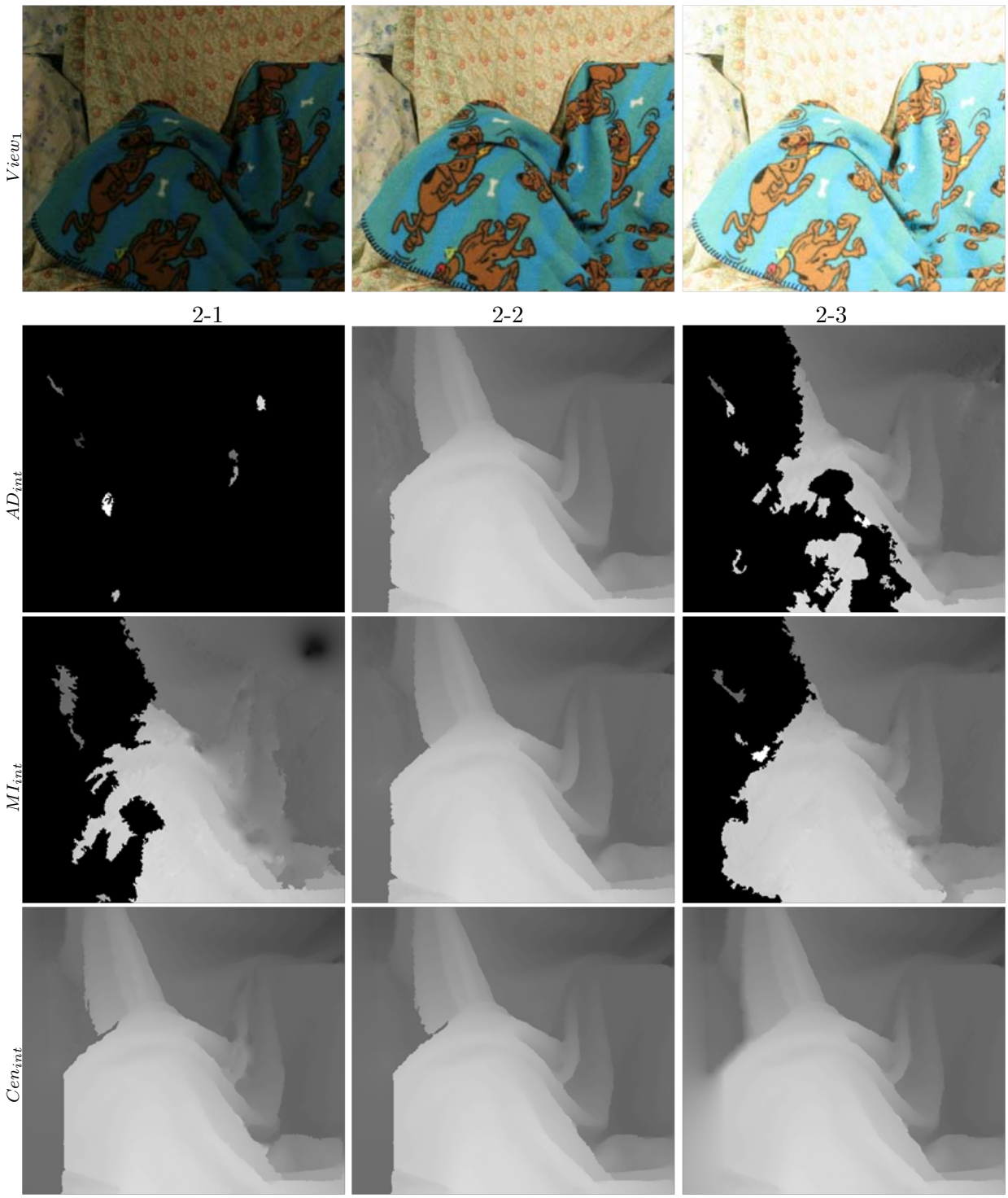
(a) Comparison using intensity images with different exposures.

Figure 5.9: (a) Results comparison of AD_{int}, MI_{int} and Cen_{int} using **Cloth3** under the same illumination with different exposures. The notation $l - r$ indicates different exposure combinations, i.e., the left view captured with the l-th exposure is matched with the right view captured with the r-th exposure. Thus, 2-2 means that the same exposure setting is used for both views in a stereo pair.



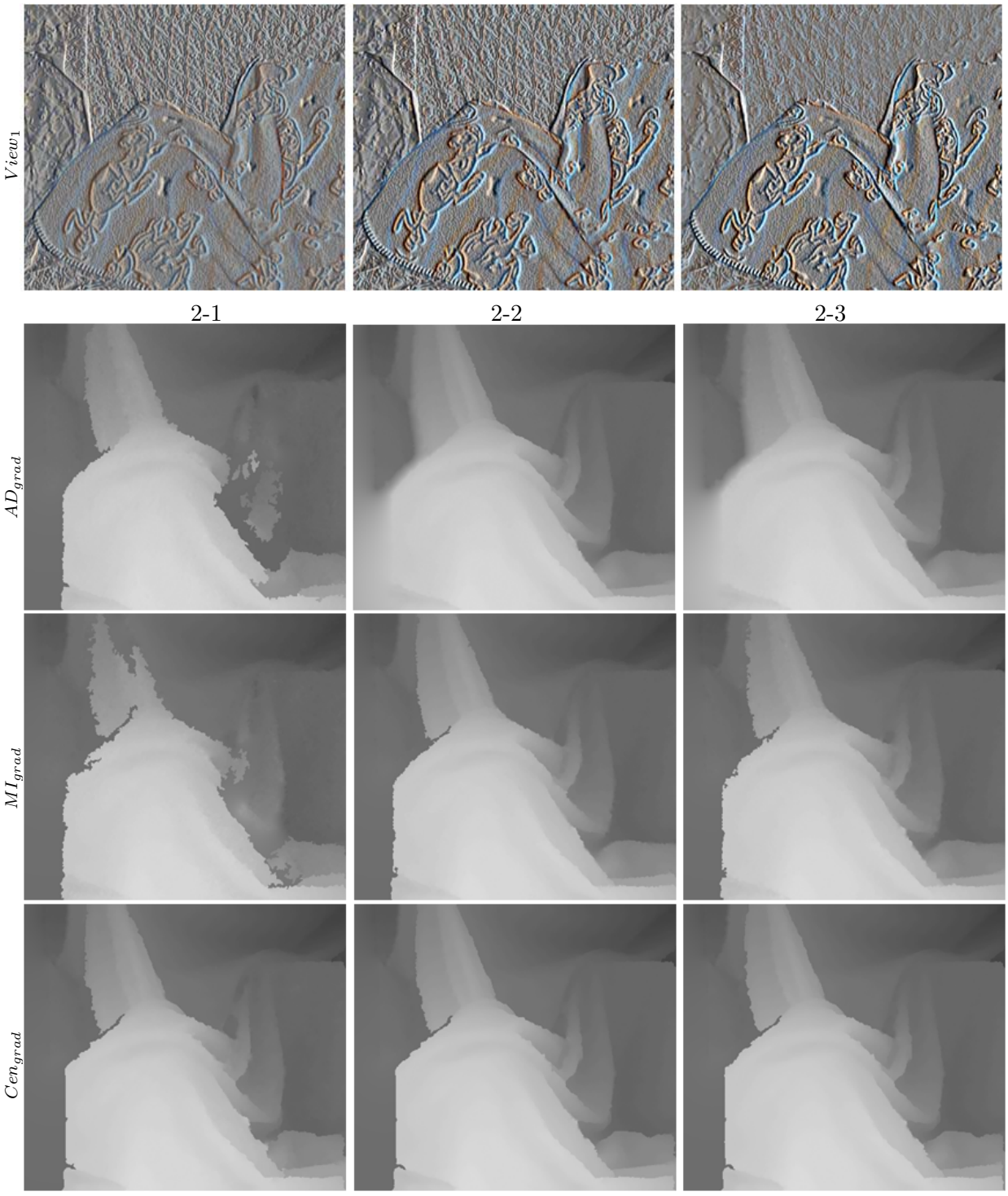
(b) Comparison using gradient images with different exposures.

Figure 5.9: (b) Results comparison of AD_{grad}, MI_{grad} and Cen_{grad} using **Cloth3** under the same illumination with different exposures.



(a) Comparison using intensity images under different illuminations.

Figure 5.10: (a) Results comparison of AD_{int}, MI_{int} and Cen_{int} using **Cloth3** with the same exposure under different illuminations. The notation $l - r$ indicates different illumination combinations, i.e., the left view captured under the l-th illumination is matched with the right view captured under the r-th illumination. Thus, 2-2 means that the same illumination setting is used for both views in a stereo pair.



(b) Comparison using gradient images under different illuminations.

Figure 5.10: (b) Results comparison of AD_{grad}, MI_{grad} and Cen_{grad} using **Cloth3** with the same exposure under different illuminations.

5.2.3 Results on Airborne Image Sequence

One limitation of the Middlebury data sets is that all stereo pairs have relatively small baselines. In this subsection, using the airborne image sequence, we study the interdependency among matching performance and match costs by progressively changing stereo perspective. Nine continually recorded images build eight stereo pairs with an increasing baseline length from 35 meters to 250 meters. The recorded urban region includes high buildings, wide and narrow streets as well as large homogenous roof areas. Both bad-pixel and normalized median absolute deviation (NMAD) are applied to describe the matching completeness and the deviation of correct matching. We evaluate our results in three specific areas shown in Figure 5.11: the city area, the church roof area, and the building boundary area. The city area covers the whole image and allows a general evaluation. The roof area consists of a large and high church in order to study the matching behavior in complicated situations like large homogenous surfaces and occlusions. The boundary area includes all building edges to analyze the discontinuities of a disparity map. The penalty factors P_1 and P_2 for each cost function are used based on parameter tuning of the Middlebury data and kept constant for all seven stereo pairs.

(a) (b) (c)

Figure 5.11: Evaluation masks on city and church area (a) and mask on boundary area (b). The LiDAR ground truth is shown in (c).

Master 1 8

Figure 5.12: Cutdown in the church area. M denotes the master image, which is respectively matched with 1 and 8 that the stereo pair M-1 has the shortest baseline and M-8 has the longest baseline.

To demonstrate the stereo perspective change during increasing the baseline length, a small

cutout of the church area is shown in Figure 5.12. The master image is matched with the sequentially recorded images respectively. All images are captured by the nadir camera of the 3K system. In total, eight stereo pairs covering the same area are evaluated.

Figure 5.13 illustrates the bad-pixel percentage and the NMAD of all individual cost functions (without merging) in the city area. The mis-match percentage (bad-pixel) decreases for all match costs from the first to the second stereo pair, because a too short baseline length causes a higher height error. After the second frame, the matching performance (NMAD) is slightly decreased, but holds relatively steady for the next three stereo pairs. The reason for this is that one the object similarities change during increasing the baseline length – the matching performance decreases, but statistically a matching result with relative large baseline length has less ground deviations. From the fifth stereo pair (M-5), the bad-pixel percentage rises due to notably changed observation views, particularly challenging for parametric costs like absolute difference. Matching costs using gradient images show a higher sensitivity to baseline increase due to uncertainties of edge locations. In general, census shows the highest robustness on the airborne images with increasing baseline length. Mutual information performs slightly better than census for stereo pairs with small baselines, but fails on data with large baselines, possibly due to the Non-Lambertian reflectance.

As shown in Subsection 5.2.2.2, match costs based on gradient images are almost invariant to the radiometric changes. However, in contrast to the results on the Middlebury stereo sets, gradient based match show less stability when stereo pairs have a large baseline. In the remote sensing data, the facades of high buildings are observed differently in two views with large baseline length, thus the gradients on objects boundaries are completely changed. Matching using gradient images is inferior to the intensity images. The similar behavior can not be observed on stereo pairs with small baselines, like the Middlebury benchmark. Because gradient location shifts little, if the parallax changes only slightly.

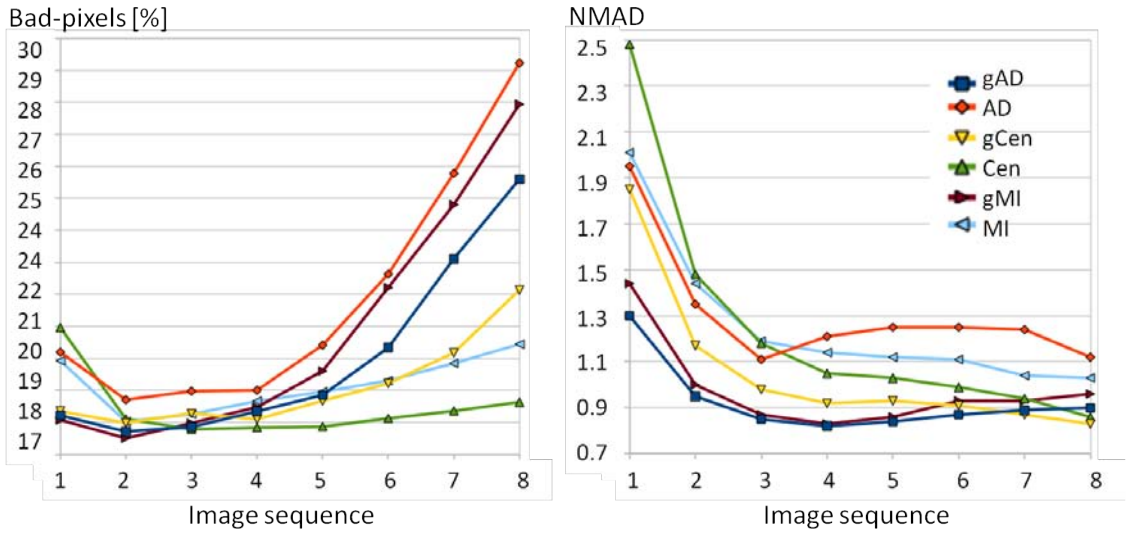


Figure 5.13: Evaluation on the city area: indexes 1 to 8 denote the different matching pairs with increasing baseline. The *Y-axis* in the left diagram denotes of bad-pixel (height error more than 5 meters). The *Y-axis* in the right diagram denotes the NMAD. Census shows the most robustness against the variation of parallax. Gradient based matching costs are more accurate than their gray value variants, if the matching is successfully executed.

Two stereo pairs ($M-2$ and $M-8$) are selected to observe the matching performance difference during increasing baseline length. Census (Cen_{int}), mutual information (MI_{int}), and gradient based absolute difference (AD_{grad}) are compared in Figure 5.14 (We have not shown the result of AD_{int}, because this match cost is too sensitive to large stereo baseline.). The results using different cost functions on the stereo pair $M-2$ with a small baseline length seem similarly, however, visible differences are measured on the results using the stereo pair $M-8$ with a large baseline length. While gradient based absolute difference fails completely on the stereo pair $M-8$, census and mutual information perform well in different regions: census generates more complete surfaces of homogenous regions; mutual information produces sharper edges on building boundaries. The robust performance of census in less texture areas is endowed from its window-based local support. But using neighbors for matching within a fixed window size causes blurring problem of census when apply it in discontinuity areas.

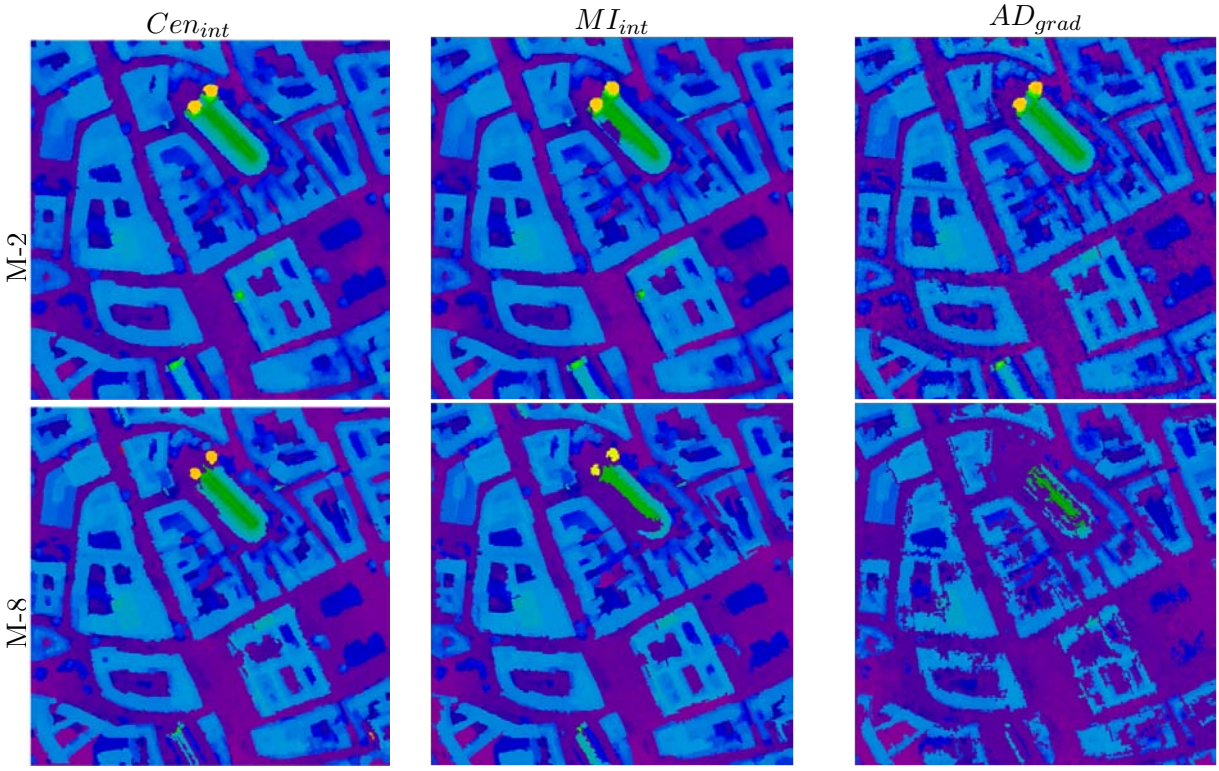


Figure 5.14: Results comparison between Cen_{int}, MI_{int} and AD_{grad}. The stereo pair, M-8, has a larger baseline than M-2. Census shows the most robustness during increasing baseline length. In contrast, AD_{grad} performs well on M-2, but failed mostly on M-8. The edges generated by MI_{int} are shaper then Cen_{int} and AD_{grad}.

In the church area with less texture, cost functions using gradient images show more robustness and higher accuracy than using intensity images. This character is consistent with and manifested on the Middlebury sets like **Wood** and **Plastic**. Absolute difference and mutual information are sensitive to radiometric changes and use only pixel-wise information for matching, hence, both of them fail mostly in the homogenous regions. Figure 5.13 shows the bad-pixel percentage and the NMAD of different matching costs in the church area. In the homogenous regions of stereo pairs with large baselines, census using intensity images performs more robust than gradient based absolute difference and mutual information. Note that, in Figure 5.13, the failed matching with a large deviation is dealt as an outlier and is not counted in the evaluation,

neither the bad-pixel percentage nor NMAD.

Figure 5.16 illustrates the comparison using census, gradient based absolute difference and mutual information on the church area. The Digital Elevation Models (DEMs) are generated by the $M-5$ stereo pair. Compared with gradient based absolute difference and mutual information, census generates less noisy but loses the roof apex. On the building borders, census shows more complexness than the gradient based matching costs. To evaluate the matching performance, we apply the third mask on the boundary area.

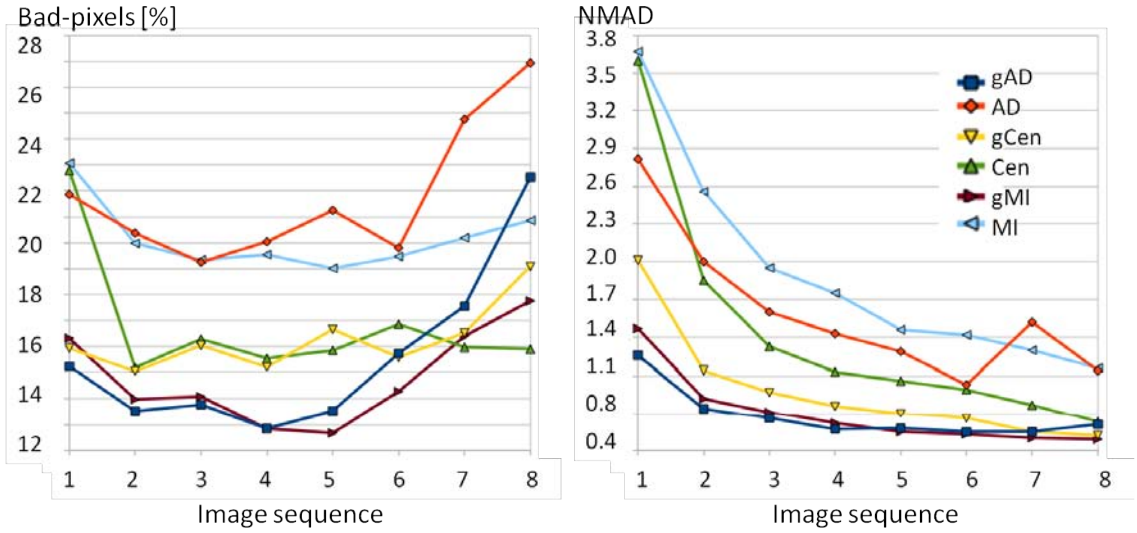


Figure 5.15: Evaluation on the church area. Gradient based matching costs show the most robustness and the highest accuracy in this area, but the failed matching increases during increasing baseline length. AD and MI failed mostly in this area.

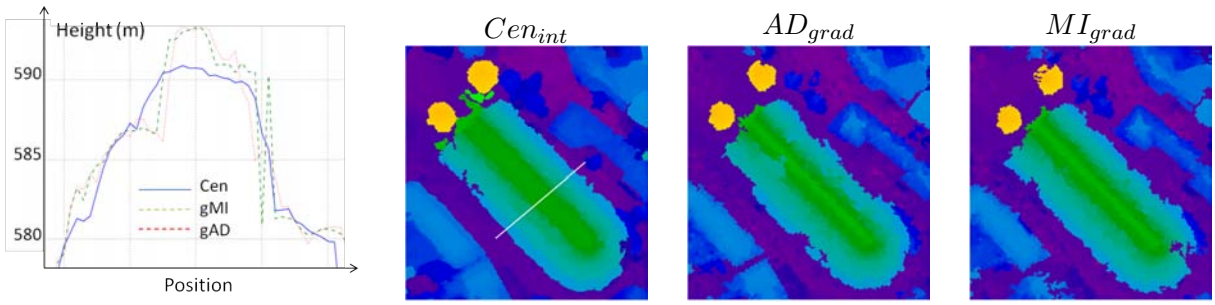


Figure 5.16: Smoothness and completeness comparison on the church area. The left diagram shows the height values for Cen_{int}, AD_{grad} and MI_{grad} along the profile indicated by the white line in the fist DEM. Compared with the gradient matching costs, Census generates less noisy but loses the roof apex.

Results on intensity images show mutual information produces sharper edges at discontinuity boundaries. In contrast, the size-fixed local support of census causes slightly blurred edges. Thus, the third mask is applied in order to detect matching performances of different on object boundaries. Figure 5.17 presents the bad-pixel percentages of census, mutual information and their combination (MIC_{int}) respectively. Generally, mutual information performs better than census, but the errors of mutual information rise observably during increasing baseline length.

MIC_{int} performs similar as MI_{int} on data sets with short baselines and show also robustness like census when mutual information failed. The weight w_{MI} of Mutual Information for this evaluation is kept constant at 0.6[1] for all stereo pairs.

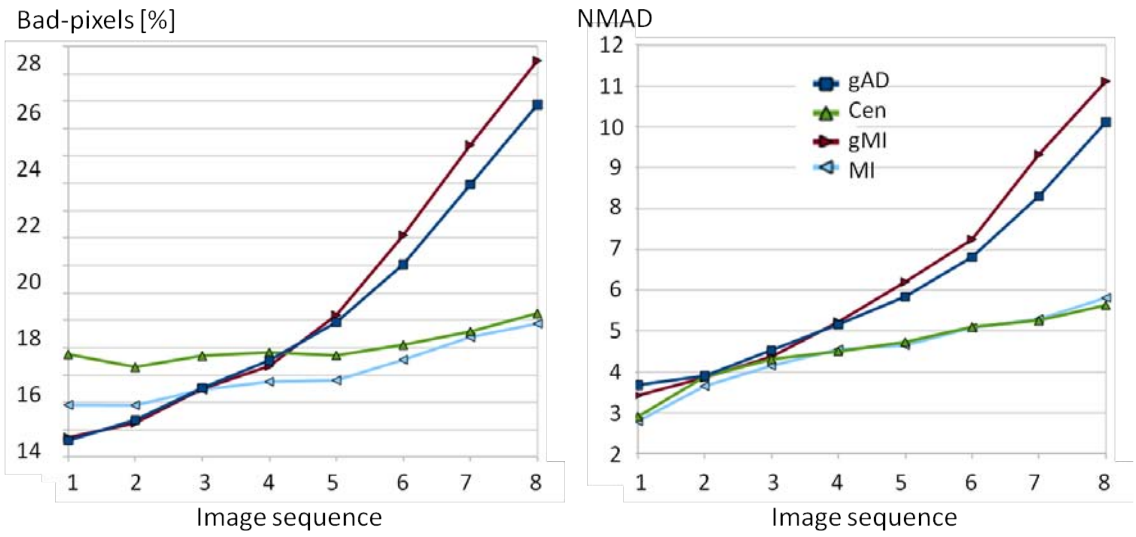


Figure 5.17: Evaluation on the boundary area. MI performs better than Census for the stereo pairs with small baseline lengths. The lead of MI compared to Census disappeared by increased baseline length due to the higher robustness of Census. Failed matching of AD_{grad} and MI_{grad} increases observably, if the angle of view changes strongly, while Census and MI stay relative stable.

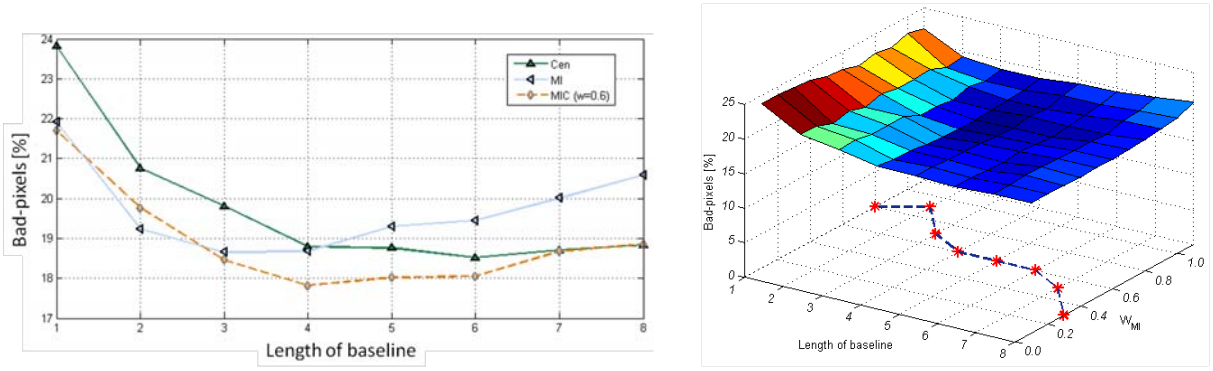
Since the different behaviors of mutual information and census depending on the baselines, we tuned the weight w_{MI} from 0 to 1 with step 0.1 for each stereo pair during changing baseline length, that the relationship between matching performance and cost functions depending on baseline can be measured and manifested by the cost weight. Figure 5.18 demonstrates the error analysis of different weight parameters to combine mutual information and census (MIC_{int}). The result shows the larger the baseline of the stereo pair has, the smaller w_{MI} performs better. In other words, census performs better than mutual information on data with larger baseline. The weight should be adapted to the observation constrains, like stereo baseline (adaptive MIC_{int}). The results of the statistical evaluation are shown in Table 5.5.

Figure 5.19 demonstrates the improvement through merging mutual information and census (MIC_{int}). In the boundary areas, MIC_{int} performs similarly as MI_{int} generating sharp edges and remains the completeness of census in the homogenous roof areas.

Finally, we illustrate results using mutual information, census and their combination with a fixed or adaptive weight parameter (fixed MIC_{int} and adaptive MIC_{int}) on the stereo pairs when stereo baseline length increases as shown in Figure 5.21. Both mutual information and census perform well on stereo pairs with a small baseline length. However, Mutual information fails gradually in the roof area during changing stereo perspective. In contrast, census can generate less mis-matches then mutual information, even on stereo pairs with very large baselines. The merged matching cost, MIC_{int} with a fixed weight parameter, improves the matching performance, but still fails in the roof areas. Adaptive MIC_{int}, whose weight parameter is tuned according to the baseline length, shows the most robustness and the highest accuracy presented

[1] We use the weight parameter tuned by Middlebury data sets.

in Table 5.5.

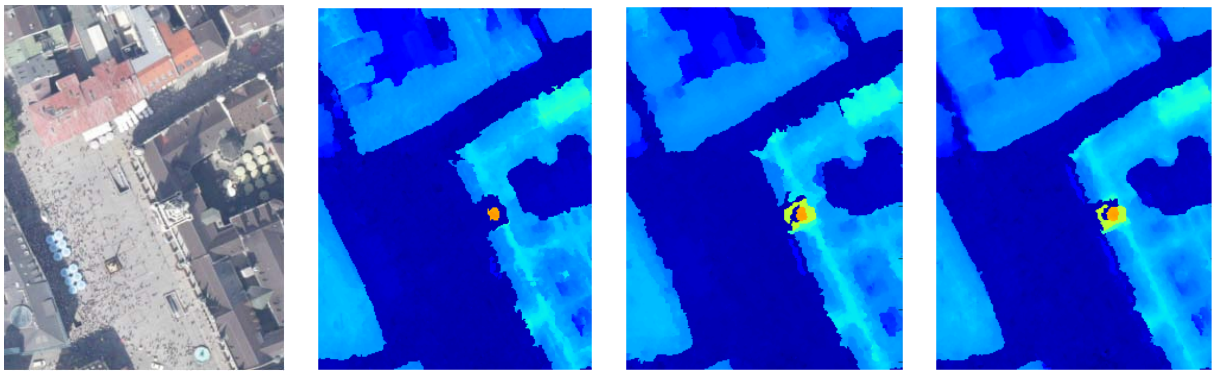


(a) (b)

Figure 5.18: MIC merged with fixed and adaptive weight. The left diagram illustrates the comparison among MIC with fixed weight parameter w_{MI}=0.6, MI and Cen. MIC performs similar to MI on stereo pairs with small baselines, but keeps the robustness of Census for larger baselines. The right diagram shows the weight tuning for w_{MI} during increasing of baseline length. The positions getting minimal error are projected at the Baseline-Weight surface and show the weight trend.

	M-1	M-2	M-3	M-4	M-5	M-6	M-7	M-8
w	(35)	(70)	(105)	(140)	(175)	(210)	(245)	(280)
0.0	23.8	20.7	19.8	18.7	18.7	18.5	18.7	18.8
0.1	23.9	20.5	20.0	18.4	18.7	18.4	18.7	18.7
0.2	23.6	19.7	19.4	18.4	18.5	18.2	18.9	18.7
0.3	23.4	20.3	19.5	18.1	18.4	18.2	18.7	18.5
0.4	22.6	19.5	19.5	17.9	18.2	18.5	18.7	18.7
0.5	22.4	20.1	18.4	17.9	18.1	18.2	18.6	18.6
0.6	21.7	19.7	18.4	17.8	18.0	18.0	18.7	18.8
0.7	22.2	19.7	18.1	17.9	18.3	18.2	18.5	18.9
0.8	21.6	19.9	18.8	18.1	18.2	18.5	19.1	19.2
0.9	22.5	18.9	18.2	18.2	18.8	18.8	19.2	19.5
1.0	21.9	19.2	18.6	18.6	19.3	19.4	20.0	20.5

Table 5.5: Evaluation of matching results in the boundary areas using MIC_{int} with different w_{MI}. Bad pixels are the percentage of mismatching compared with LiDAR points with an error > 5 meters. The red number in each column is the minimum error percentage and denotes the best weight used.



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

Figure 5.19: Improvement by merging matching costs: MIC_{int} (d) generates sharper edges than Cen_{int} alone (c) and constructs the high building completely which is lost using only MI_{int} (b). One reference image is shown in (a).

5.2.4 Results on Satellite Data

A small cutout of the stereo data and the reconstruction results for an urban area (Terrassa in Spain, Barcelona) are shown in Fig. 5.20. The full data set covers mountainous, agricultural, forest, industrial and residential areas. The figure indicates that these images cannot be matched successfully using Mutual Information, while Census and the MIC_{int} perform reasonably well on this challenging data set. The large black background in the Mutual Information image was incorrectly filled using this data. Table 5.6 shows the results of evaluating the city area shown in Fig. 5.20 and two other test areas (hilly forest and industrial area) against the LiDAR reference data. It is clearly visible that MIC_{int} performs slightly better than Census and that Mutual Information does produce the largest errors. Experiments with various values for P_1, P_2 and W_{P_2} indicated that performance depends mainly on the cost function and not on the exact parametrization of the stereo algorithm.

Cost	P_1	P_2	W_{P_2}	w_{MI}	NMAD	Bad pixels
MIC_{int}	700	1400	200	0.3	0.72	15.8 %
Cen_{int}	600	1300	200	-	0.74	16.8 %
MI_{int}	700	1400	200	-	1.10	25.8 %

Table 5.6: Evaluation of Matching results in three test areas using ground truth LiDAR Data. NMAD is the normalized median deviation and Bad pixels is the percentage of pixels with an absolute height error > 2 m.

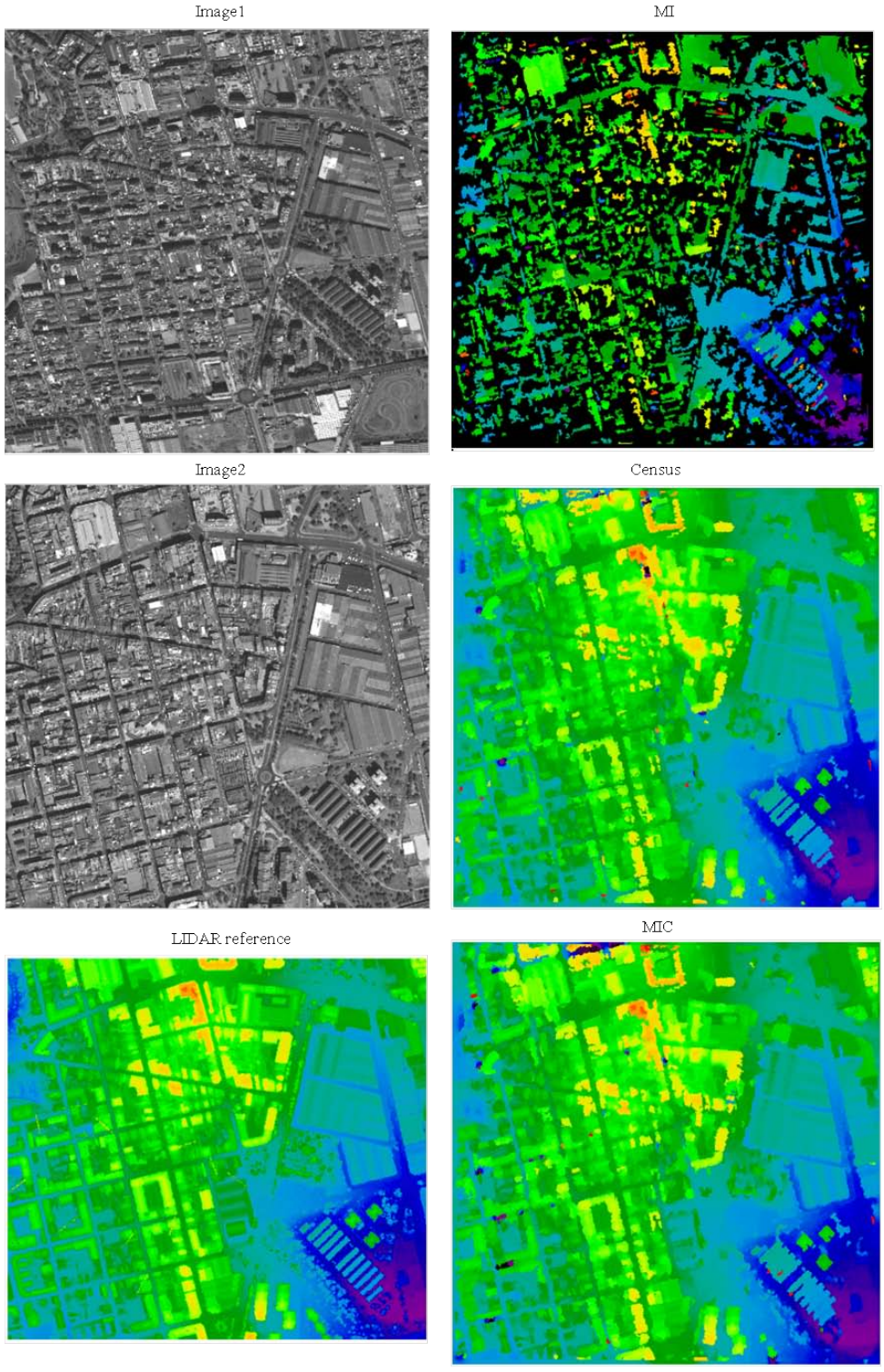


Figure 5.20: Small cutout of the Worldview-1 Stereo pair. First column: stereo pair and LiDAR reference data. Second column: Results after stereo matching with different cost functions, orthographic reprojection and discontinuity preserving interpolation.

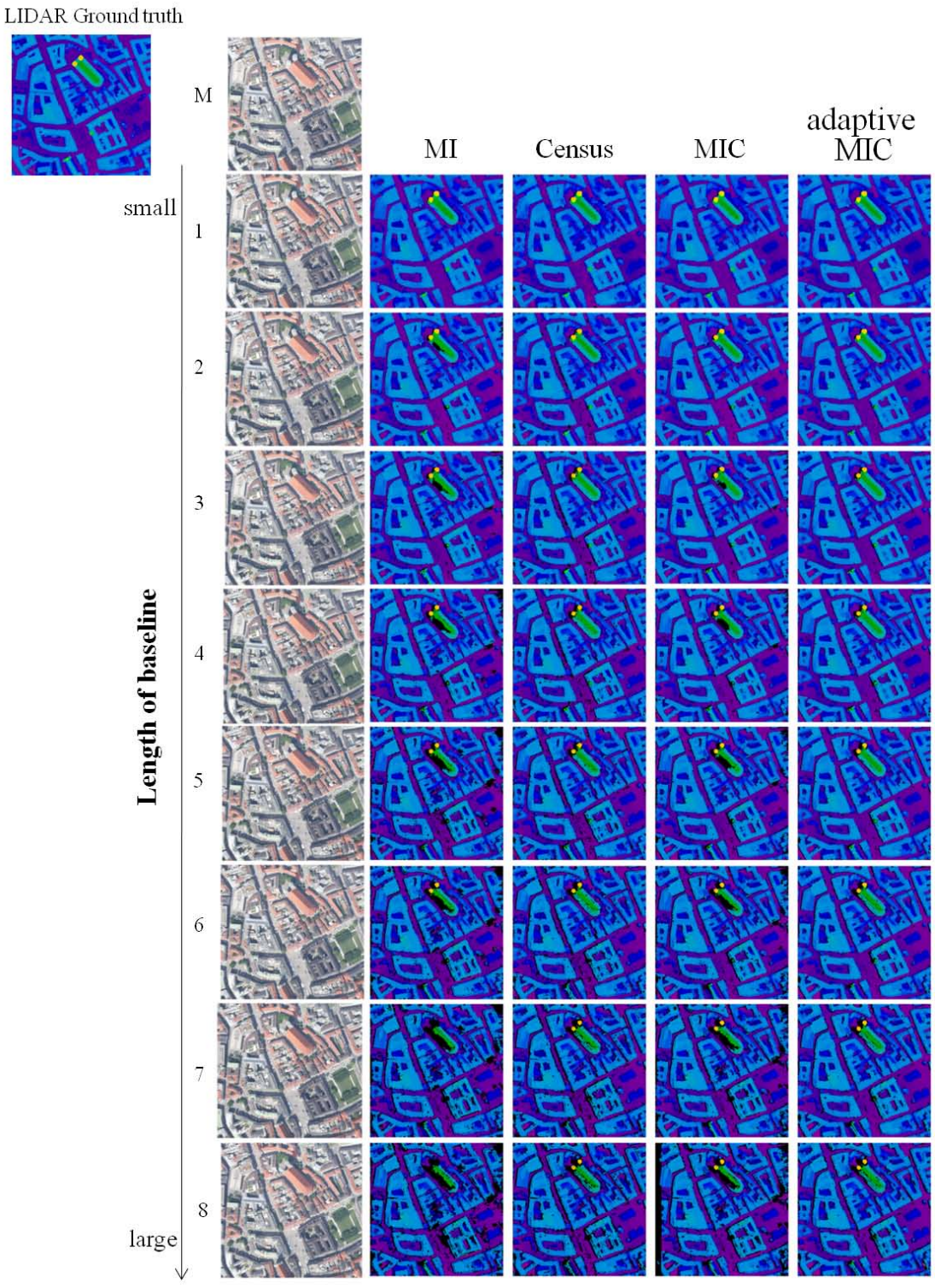


Figure 5.21: Disparity maps for stereo pairs with increasing baseline. The images 1 to 8 are matched with the center image C respectively [b]. The results for MI, Census, MIC and adaptive weighted MIC are shown in columns 2-5. The black areas indicate failures of the left-right check.

[b] Note that the shown reference images are not rectified.

5.2.5 Discussion

In this section we investigate the interdependencies among matching performance, match cost functions, and observation constraints using both the standard benchmark and remote sensing data sets. We summarize our study under four criteria as follows:

- **Insensitivity to radiometric changes:** Although absolute difference and Mutual Information perform well on Middlebury data under good radiometric conditions and remote sensing data with small baseline lengths, they fail mostly on Middlebury data with radiometric changes and remote sensing data with long baseline due to non-Lambertian reflectance. In contrast, census shows the highest robustness on both close-range data with radiometric changes and almost all used remote sensing data. Cost functions applied to gradient images are generally insensitive to radiometric changes and perform well on Middlebury data and remote sensing data with short baseline length.

- **Robustness during increasing baseline length and stereo angles:** Except non-Lambertian reflectance, one most challenge for stereo pairs with long baseline and large stereo angle is notable object shape change between left and right view, an extreme example is the facade area of a high building. This challenge causes location shift of gradients that matching costs using gradient images mostly fail. Mutual information applied to intensity images performs better as absolute difference on airborne stereo pairs with long baselines, but fails completely on the satellite data with very large stereo angles. Census achieves the highest robustness both on airborne and satellite images.

- **Completeness in homogenous areas:** Our evaluation shows that the matching performance of cost functions depends not only on the observation conditions, but also on the imaged areas. Regions with less texture impede matching by high ambiguities. Pixel-wise costs applied on gradient images show more completeness than the directly using intensity images, probably due to non-labertian reflectance. The bit-wise recording of census detects the relative relationship between the local ambiance and restricts outliers in small value that Census outperforms Absolute Difference and Mutual Information in these areas.

- **Blurring in discontinuity regions:** Cost functions applied to gradient images generate rugged edges due to location shifts of gradients in two stereo-views. In addition, the window-based local support of census causes blurring in discontinuity regions both on Middlebury data and airborne images. However, pixel-wise costs using intensity images show more accuracy in the object boundaries and sharper edges are reconstructed.

- Contribution of merging matching costs: The influence of data on matching performance shows, there is no almighty matching cost for all data in all regions. Merging different costs profits their individual advantages to improve the matching performance. In our evaluation, the cost combination using mutual information and census performs similarly as absolute difference and mutual information in large discontinuity area and as robust as census both in homogenous area and on data with large stereo angles.

5.3 Evaluation of the Confidence-based Surface Prior

In this section we evaluate the confidence-based surface prior and compare it with existing segment-based methods. Both the Middlebury benchmark and the 3K+ airborne stereo pairs are used. In our evaluation, the advantages of adding the confidence-based surface prior into a global framework are highlighted: We analyze the discontinuities along object boundaries to demonstrate sharper edges through the surface prior; Texture-rich stereo pairs are selected to show the decreasing sensitivity of our prior to a given segmentation; The profile through a building roof shows the improved matching completeness using the surface prior on large homogenous regions.

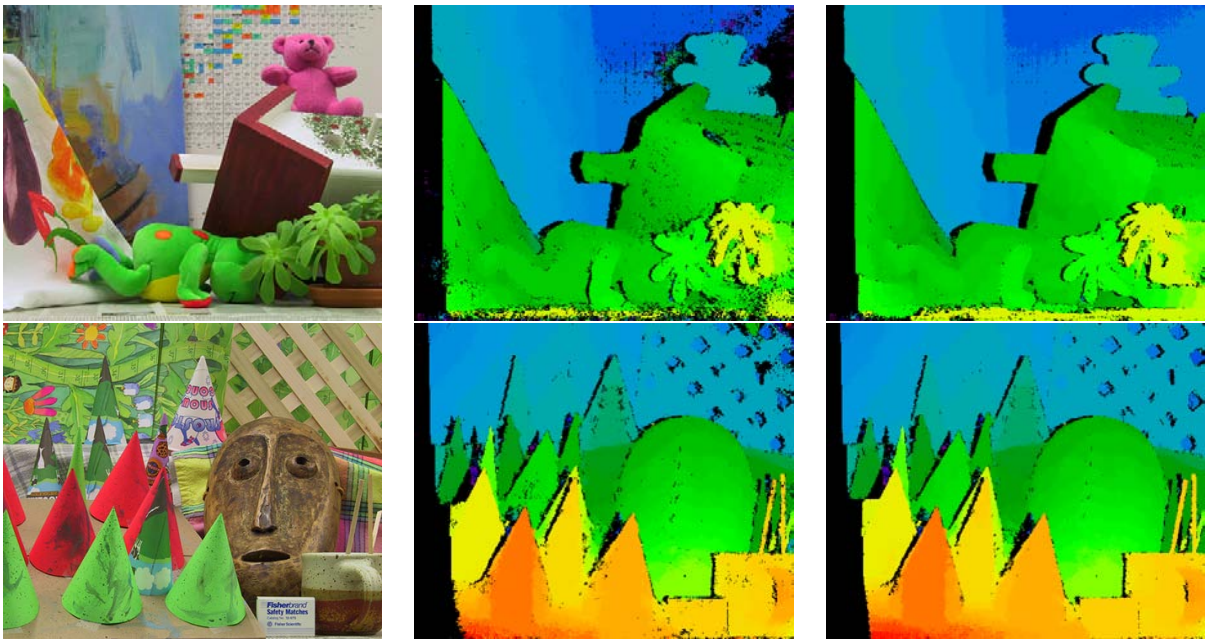
5.3.1 Results on Middlebury Benchmark

Table 5.7 demonstrates the quantitative improvement when applying our surface prior on the Middlebury online benchmark. Compared with our SGM implementation, improvement is obtained under all criteria for all four datasets. Figure 5.22 demonstrates the result comparison between SGM and our iSGM3 on **Teddy** and **Cones** from the Middlebury online benchmark. The dilation of the chimney and Teddy bear are reduced in (c) in contrast to (b). The mesh background is better defined and edges are sharper in (c). Our results perform similar as the work using soft surface prior [Woodford et al., 2009] except the result of **Venus**, because our segmentation is not tuned for the online benchmark. It seems, color segmentation works well on these four data sets, thus the methods using segmentation-based hard constraint perform very well [Sun et al., 2005; Wang and Zheng, 2008]. However, as shown in Figure 4.1, methods using hard plane constraint are extremely sensitive to the given segmentation. Oft over-segmentation is required. In texture-rich regions, artifacts from segmentation can appear.

Algorithm	Tsukuba			Venus			Teddy			Cones		
	nocc	*all*	*disc*	*nocc*	*all*	*disc*	*nocc*	*all*	*disc*	*nocc*	*all*	*disc*
iSGM3	2.40	3.16	10.1	1.47	2.49	14.2	10.2	16.3	21.4	4.2	11.4	12.7
SGM	2.73	3.60	11.4	2.0	3.32	15.9	12.1	18.0	23.2	5.41	13.5	13.8
2OP+occ	2.91	3.56	7.33	0.24	0.49	2.76	10.9	15.4	20.6	5.42	10.8	12.5
SymBP	0.97	1.75	5.09	0.16	0.33	2.19	6.47	10.7	17.0	4.79	10.7	10.9
DoubleBP	0.88	1.29	4.76	0.13	0.45	1.87	3.53	8.3	9.63	2.9	8.78	7.79

Table 5.7: Evaluation of bad-pixel percentage on Middlebury online benchmark with Error Threshold > 1. Our SGM results differ from Hirschmüller [2008] due to different match cost function and post processing for filling occlusion holes used in our implementation. The iSGM3 performs similar as 2OP+occ [Woodford et al., 2009] except the result of **Venus**, probably due to the color segmentation used. Methods using hard segment-based constraint like SymBP [Sun et al., 2005] and DoubleBP [Wang and Zheng, 2008] show the best performance on these data.

Furthermore, figure 5.23 demonstrates improved accuracy when utilizing our surface prior on more images from the Middlebury datasets other than the online benchmark. The proposed confidence-based surface prior converts the hard segment-based constraint from the plane-fit disparity map into a soft prior that is not as reliant on the plane-fit disparity. Sharp edges and sloped surfaces can be generated naturally via a global energy minimization without the hard constraint imposed by plane-fit disparity. For instance, the open book in **Baby2** is texture-less and has a smoothly bowed surfaces in the disparity space. The result without our prior contains



(a) Reference images (b) SGM (c) iSGM3

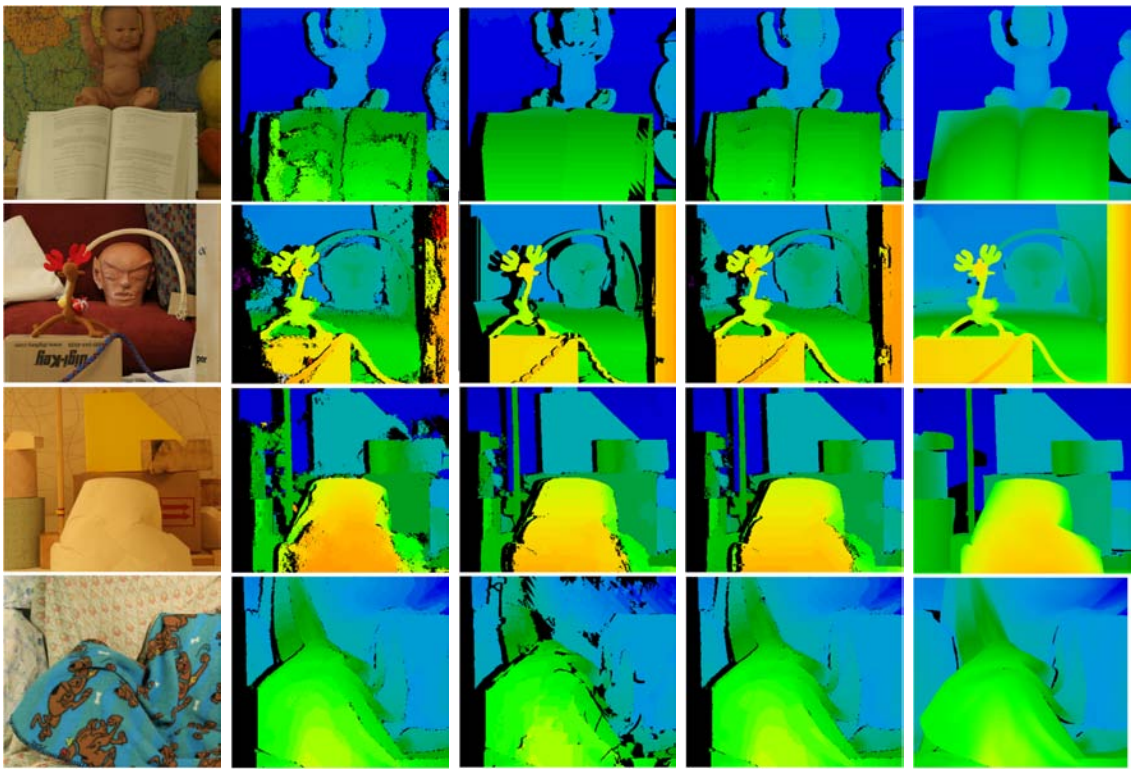
Figure 5.22: Comparison on Middlebury online benchmark. We compare the disparity results, after left-right consistency checking, with and without the proposed soft prior. **Top**: The dilation of the chimney and Teddy bear are reduced in (c) in contrast to (b). **Bottom**: The mesh background is better defined and edges are sharper in (c).

large miss-matched areas due to texture homogeneity. In contrast, the hard segmentation-based constraint enforces a single slanted plane for the book, which is not in consistent with the ground truth. In contrast, our surface prior allows the faithful reconstruction of the books surface. As edges of objects boundaries and slanted planes in the disparity maps of **Reindeer** and **Lampshade1** benefit from the hard plane constraints, **Cloth3** with rich textures and smoothly varying depth is cut apart by the segments artifacts. The hard constraint enforces each segment as a disparity plane instead of a segmentation-independent smooth surface. Our confidence-based prior improves shape edges and slanted surfaces, but still remains good results from initial matching.

Algorithm	Baby2			Reindeer			Lampshade1			Cloth3		
	nocc	*all*	*disc*	*nocc*	*all*	*disc*	*nocc*	*all*	*disc*	*nocc*	*all*	*disc*
iSGM3	10.4	12.3	14.1	14.0	17.6	22.5	13.6	14.5	18.8	3.6	4.2	5.5
SGM	17.3	19.3	21.6	19.3	25.3	31.2	16.3	18.1	21.2	3.6	4.2	5.5
Plane-fit	15.9	17.5	24.3	12.2	18.8	21.8	13.2	14.4	18.4	17.5	18.7	29.7

Table 5.8: Evaluation on the Middlebury data sets with bad-pixel error threshold > 1. The results are shown in Figure 5.23.

Our confidence-based surface prior differs from existing surface constraints in that it varies the per-pixel strength of the constraint to be proportional to the confidence in our given disparity estimation. In contrast to many methods using soft surface priors [Bleyer et al., 2010; Woodford et al., 2009], the proposed confidence measure for our surface prior is calculated from



(a) Reference images. (b) Results without our soft surface prior. (c) Plane-fit disparity from (b). (d) Results with our soft surface prior. (e) Ground truth.

Figure 5.23: Comparison on Middlebury 2005 and 2006 data sets. **Top (Baby2)**: Discrete disparity levels in the low-texture curved book surface in (c) is partially smoothed out in (d). The baby doll is also better reconstructed in (d). **Top-mid (Reindeer)**: The blue foreground rope is reconstructed better in (d) than in (b) or (c), and the brown head/mask has better defined edges in (d) than in (b). **Bottom-mid (Lampshade1)**: Both the thin stick and large surfaces are better reconstructed in (d). **Bottom (Cloth3)**: Our surface prior performs robust on texture-rich data. (b) and (d) are almost identical, as segmentation artifacts are observed in (c).

an energy minimization without reference to planar surfaces or regions. Thus it is independent of the plane fitting process. Using this confidence, the Gaussian interpretation of the plane-fit result in energy space makes our surface prior more robust to segmentation errors. To demonstrate the effectiveness of our confidence measure we compare disparity maps produced by our proposed confidence-based Gaussian surface prior and the soft surface prior defined using disparity distance between δ_p and d_p in Figure 5.24. Absent scaling by our confidence measure, the soft surface prior causes the reconstruction of the bowling ball to be relatively planar in vertical stripes. The round shape of the bowling ball is correctly reconstructed when using the confidence-based surface prior.

In contrast to real-world data, color segmentation works well on the Middlebury data sets. This is the main reason, why many segment-based methods reach high ranks of the online benchmark. However, effects like object shadowing, which often appear in real-world applications, are not included. Such failed color segments can lead to critical results. For instance, a building

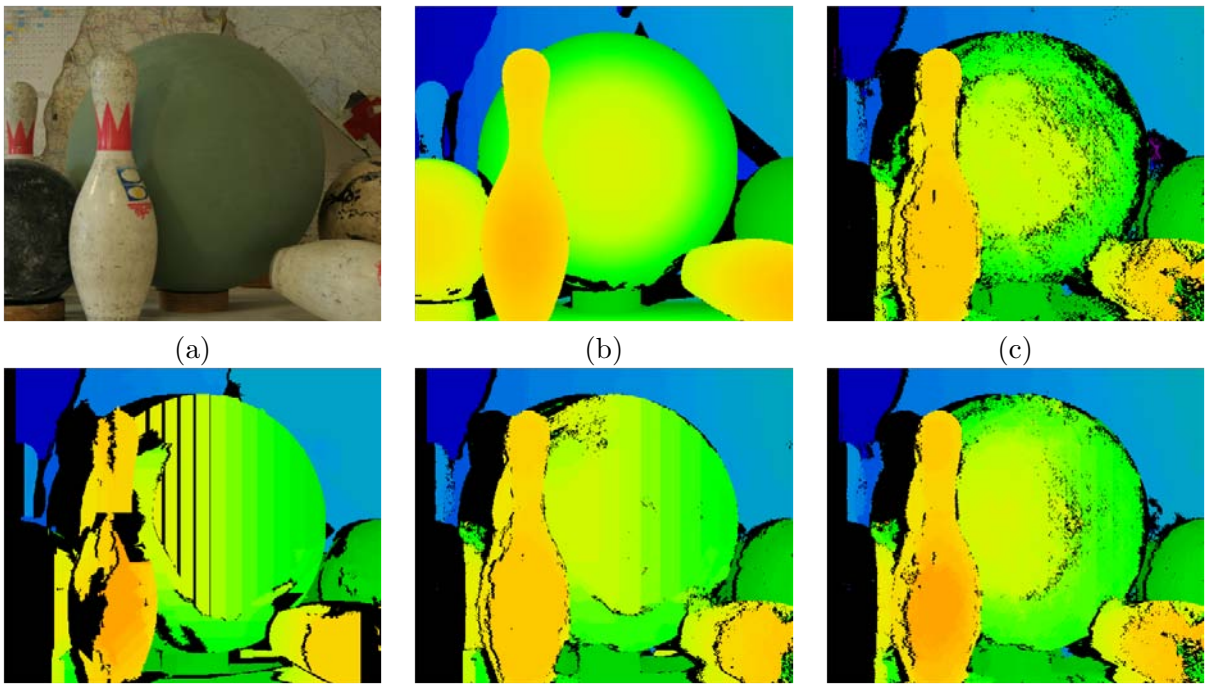


Figure 5.24: Comparison of simple disparity distance prior and the confidence-based prior on **Bowling2**: **(a)** Reference image. **(b)** Ground truth. **(c)** Result without soft prior: matching on the boundary areas of objects failed a bit. But the shape of the ball is well reproduced. **(d)** Plane-fit disparity from (c): The ball is reconstructed as a set of vertical stripes, and the pin has become poorly defined. **(e)** Result using soft prior without confidence: Edges are well defined, but some vertical striping of the bowling ball remains. **(f)** Result using our surface prior: The rounded shape of the ball is better reconstructed than in (e) and the edges are shaper than in (c).

facade is assigned in the same segment with its shadow on street, hard plane constraint enforces a slanted plane connecting the roof and the street, instead a perpendicular wall. In the next subsection, we show more contrast between our approach and segment-based method .

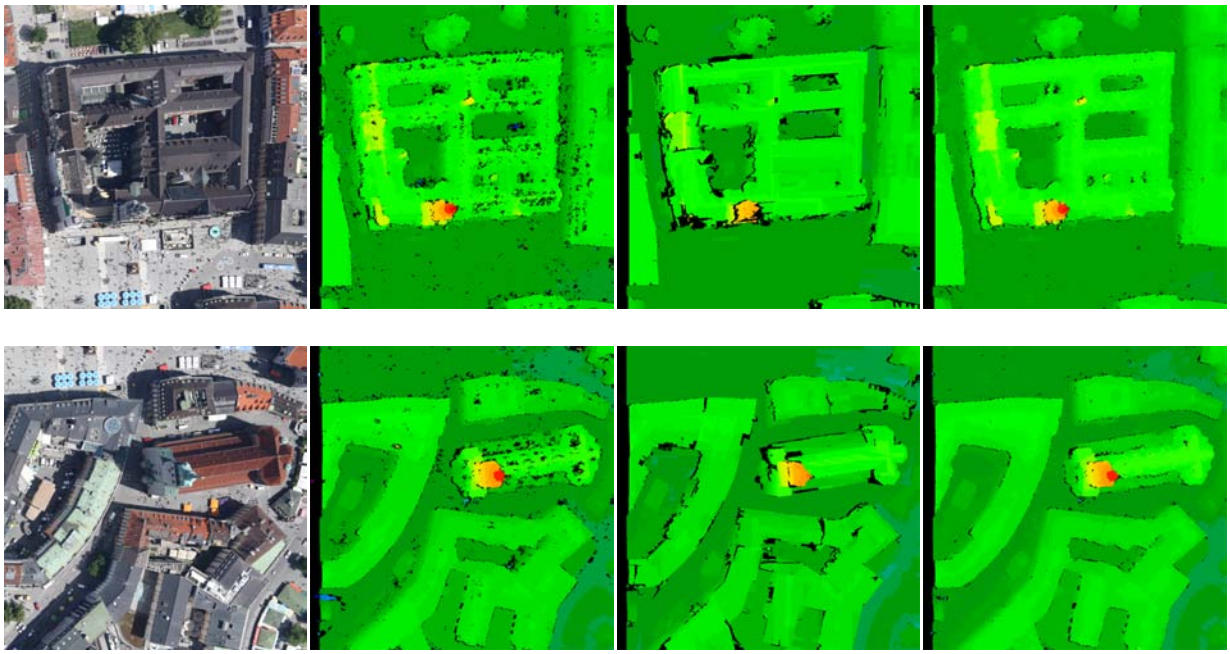
5.3.2 Results on Airborne Stereo Pairs

Non-lambertian reflectance and scene complexity all combine to make dense stereo matching on real-word data challenging in contrast to the Middlebury benchmark. The airborne 3K+ data include large homogenous roofs, high buildings, small streets, regions under shadows as well as moving pedestrians and vehicles. The lighting environment of typical airborne imagery is dynamic – the sun can be hidden by clouds of short duration. In this subsection we demonstrate results of our algorithm on airborne stereo pairs in an urban area. Obvious improvements in accuracy are obtained by our proposed algorithm: sharp building edges, complete roof regions, and less noise on the streets. Building boundaries benefit most from our method. They are important features, but cover only a small percentage of the whole image. Thus, we have manually generated building-edge segments to evaluate the disparity changes within a small location. This evaluation highlights the robustness of our surface prior when applied for a real-world scene. In addition, we use also stereo pairs with large baselines. For a fair comparison, we use the truncated absolute difference, both for both SGM and iSGM3.

We focus now on the stereo pairs with short baseline. The bad-pixel percentages of SGM and iSGM3 are very similar, 18.76% for SGM and 18.74% for iSGM3. In contrast to the comparison using the Middlebury data, the difference between both methods is not clear. There are several reasons for this: (1) The resolution of our groundtruth is much lower than the 3K+ images; (2) The 3K+ images were captured seven years after the groundtruth laser scanning. During these years, many buildings are were torn down and rebuild; (3) We observe that object boundaries benefit mostly from iSGM3 on this data set, but these cover only a small percentage of the image. However, Figure 5.25 shows two cutouts to compare results with or without the surface prior. Many mis-matches are observed on roofs, facades and streets where pedestrians are moving. Plane-fit method revamps hardly the building boundaries at cost of segment artifacts (shown as regions of black holes). In contrast, results using iSGM3 contain less mis-matches and no fitting artifacts.

Figure 5.26 shows the superior performance of our method on large homogenous regions. Along the indicated white arrow through the large homogeneous church roof area, we demonstrate the improvement obtained by our confidence-based surface prior in (c). In contrast to a lot of unreliable disparities after consistency checking on the disparity map without applying our soft surface prior 6, a smooth and gradual changing roof is reconstructed when incorporating our confidence-based surface prior. Additionally, the church edges are perfectly generated without any dilation into the shadowed streets below.

In order to evaluate the disparity changes on object boundaries, we manually generated edge segments shown as green masks in Figure 5.27 (b). Edges are selected, if the height difference beside them are very high. Figure 5.27 (a) explains the evaluation procedure using these edge masks. Given a small location threshold, disparities of pixels along the normals of the masks are compared. Only if the disparity change are higher than another disparity threshold, we count this location as a disparity discontinuity. The amount of discontinuities is divided by the pixel number of edges. A high percentage indicates that the disparity map has sharp edges. Five location thresholds from 1 pixel to 5 pixels are used. The disparity threshold is selected as 15 pixels difference at least. Table 5.9 demonstrates the comparison between SGM, iSGM3 and plane-fit generated disparity maps. Our surface prior perform superiorly in the large discontinuity regions. The hard constraint of plane-fit generates many artifacts, because segments near building boundaries often include some shadow regions on streets. In contrast to the bad-pixel evaluation of the whole image, the performance difference between results with and without our surface prior is very clear.



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

Figure 5.25: **Results on airborne images**: **(a)** Reference images. **(b)** Results without surface-prior (SGM). **(c)** Plane-fit disparity maps from (b). **(d)** Our results (iSGM3)

Method	Threshold = 1	2	3	4	5
SGM	24.03%	36.89%	48.94%	61.63%	74.75%
Plane − fit	10.34%	12.50 %	13.39%	17.97%	20.44%
iSGM3	36.76%	48.03%	59.16%	70.20%	83.25%

Table 5.9: **Evaluation using edge masks.** The location thresholds are selected from 1 pixel to 5 pixels along the normals of edges. iSGM3 performs better than SGM in all thresholds. Especially, the difference of the small thresholds is clear that results using the surface prior contain shaper edges. The plane fit results suffer from bad segmentation at object boundaries with shadows.

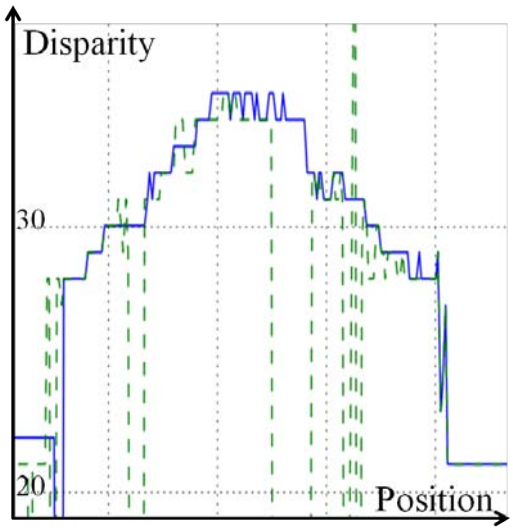
We evaluate the surface prior also on airborne images with large baselines. In Figure 5.28, the master image (M) is matched with the reference images 1 (short baseline) and 2 (long baseline). The rows, (a), (b), and (c) present the results without our soft region prior, of the plane-fitting of (a), and with our confidence-based soft surface prior, respectively. Matching on the M-2 stereo pair is much more difficult than the M-1 stereo pair due to the increased baseline separation. Even on the stereo pair with a large baseline, the contribution of our confidence-based soft surface prior is clear: more complete roofs, clear boundaries, and fewer outliers.

5.3.3 Discussion

In this section we compare the results with and without the confidence-based surface prior using both the Middlebury benchmark and the airborne stereo pairs. The iSGM3 uses a very similar energy minimization framework like SGM, the only difference between them is that we fuse costs from a plane-fit result into the aggregated match cost. However, results using the additional prior

(a) Reference image.

(b) Result without our surface prior (SGM).



(c) Profile of disparity change.

(d) Result with our surface prior (iSGM3).

Figure 5.26: **Comparison of results on the airborne images.** (b) Without our surface prior there is a lot of false matching on the street level and homogeneous roof regions. The reconstructed building edges are dilated into the shadow areas. (c) With our surface prior, accuracy is improved in all surface areas, building edges, and roof areas when compared with (b). (d) Profile on the homogeneous church roof area. The profile shows the disparity change along the indicated white arrow both with (blue line) and without (green line) our surface prior. The use of our surface prior reconstructs a closer approximation to the roof surface, without the extreme outliers observed in (b). Additionally, the building edges are sharper when our surface prior is utilized.

contain shaper edges and more complete surfaces than results using energy minimization only including data and smoothness terms. This fact – improvement by adding high-level priors like surface prior even using the same optimization framework – inspires us to make joint inferences between stereopsis and object recognition (see Future Work).

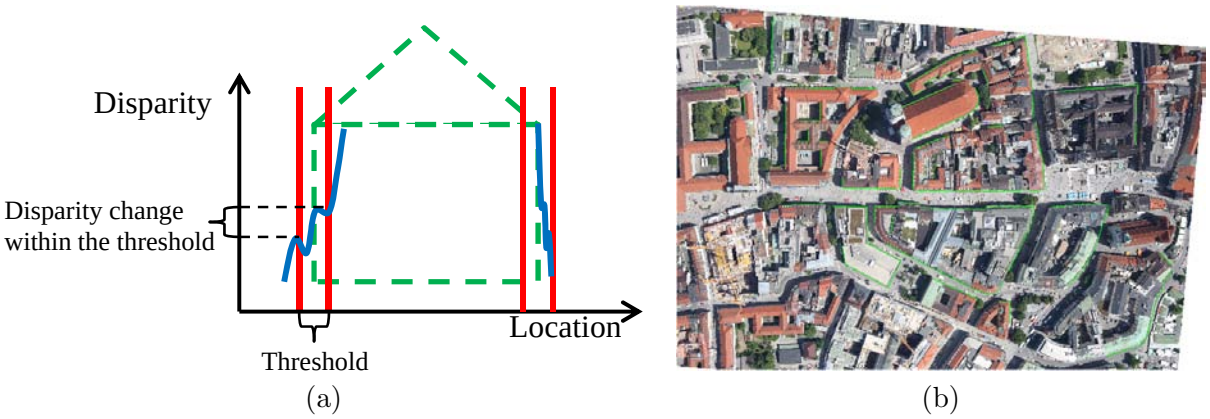


Figure 5.27: **Evaluation using edge masks** (a) Evaluation procedure: The blue shapes denote the reconstructed building facades. Disparity changes are calculated within a small location threshold (shown as red lines). (b) Edges masks of a rectified airborne image: The edges segments shown as green lines are selected on the building sides in order to check the large discontinuities besides the lines in a disparity map.

In addition, this evaluation indicates that region-based methods are preferable around object boundaries and in large homogeneous areas; the use of object regions can propagate strong matches into sub-regions where poor matches appear. However, defining an energy minimization over regions, rather than pixels, imposes a hard constraint that forces depths to lie on the smooth surface associated with a region; removing fine-level details from the depth map in the process. The results are very sensitive to the given segmentations. In contrast, our prior learns the strength of a pixel lying on a surface from the previous result and fuses this cost with match cost and smoothness penalties together. The semi-global framework aggregates the fused costs path-wise. Thus, the artifacts of a plane-fit disparity map are avoided.

5.4 Summary

In this chapter the interdependencies among matching performance, cost functions, and observation conditions are investigated using both close-range and remote sensing data: Mutual Information performs well on stereo pairs with global radiometric changes. Non-parametric match costs like Census works robust in texture-less regions and on stereo pairs with large baseline. Based on this study, the cost-merging strategy is developed in order to combine the advantages of different match cost functions and gives consideration to imagery configurations.

In the second part of this chapter, a confidence-based surface prior is introduced for global energy minimization. This surface prior varies in the existing hard segment-based priors, which force each pixel belonging to a spatial surface. In this work the hard plane-fitting result is modeled as a distribution using the previous (aggregated) match costs. Thus, segmentation artifacts can be effectively avoided. Moreover, the addition of this prior has shown more benefits: sharp object-boundary edges in areas of depth discontinuity; and accurate disparity in surface regions.

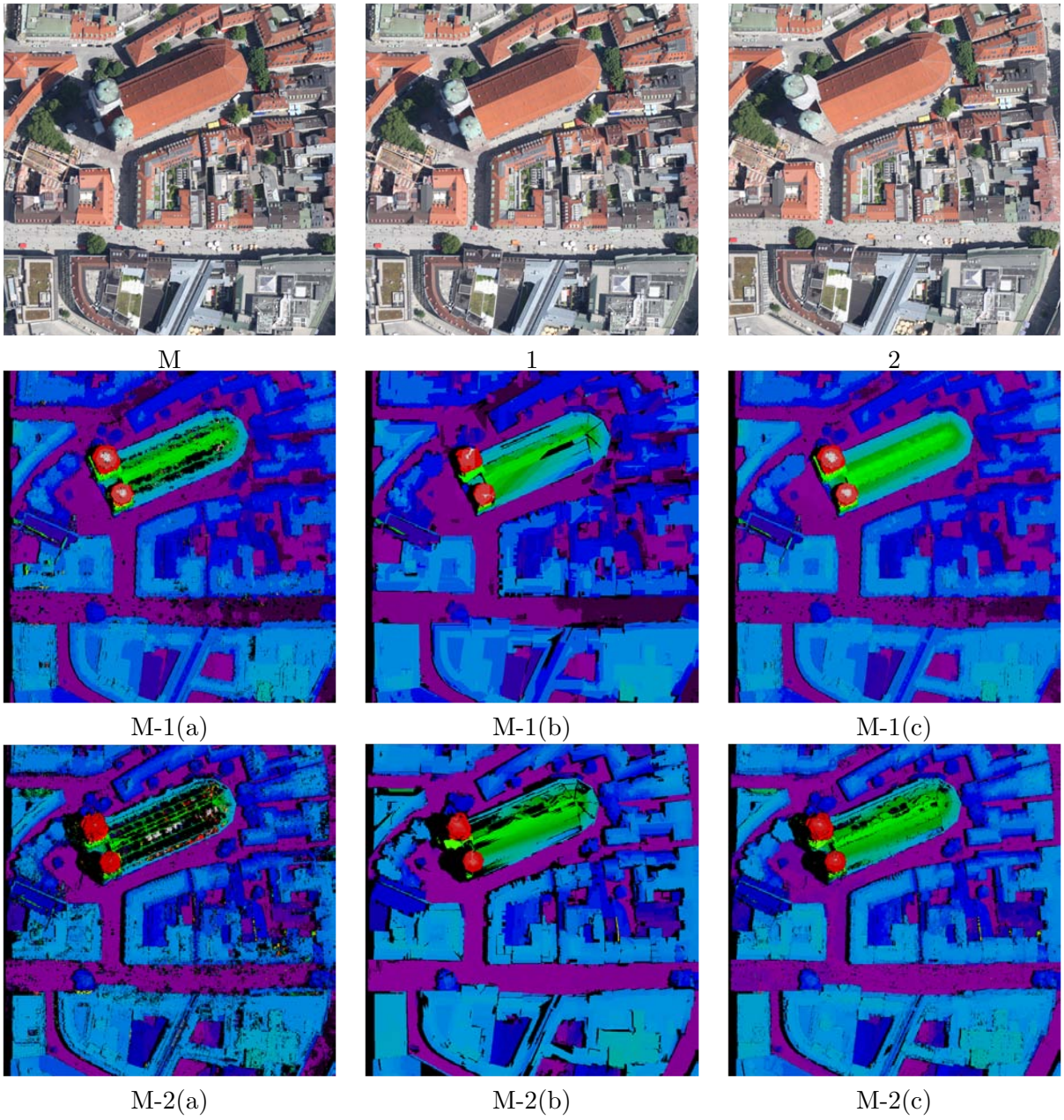


Figure 5.28: Comparison on stereo pairs with increasing baselines. M denotes the master image, which is rectified and matched with a short baseline (1) and long baseline (2) image. M-x(a) presents the results without our soft surface prior. M-x(b) are the plane-fit disparity maps to M-x(a). M-x(c) are the results with our soft surface prior.

Chapter 6

Conclusion and Outlook

The goal of this dissertation is to develop robust stereo matching method for challenging real-world data. To this end, we conduct two components of the global stereopsis frameworks: the robust match cost functions and the addition of the confidence-based surface prior in global energy formulations.

Match cost is the most fundamental component for correspondence computations. All stereo methods use a cost function to compute the similarity of possible correspondences. To develop reliable match costs, we investigate into the interdependencies among matching performance, cost functions, and observation conditions. Parametric, non-parametric match costs, and the match cost based on mutual information are evaluated and compared in context of data (radiometric changes, object features) and observation conditions (stereo baselines). Both close-range and remote-sensing data are used. The analysis in this dissertation indicates that there is no single cost function that performs best for all circumstances. Instead, the match cost function has to be chosen based on the properties of imagery conditions and data applied. We introduce some general guidelines for choosing suitable match cost functions:

- Parametric match costs and mutual information allow accurate reconstruction of fine details, such as building edges, but lead to increased sensitivity to radiometric changes.
- Non-parametric match costs encode image local structures, which improves robustness, but leads to reduced accuracy on fine details and edges.
- The matching performance of a cost function depends on observation constraints such as baseline length. For stereo pairs with large baselines, non-parametric costs perform more robust than parametric match costs and mutual information.
- Features (homogeneity, discontinuity, visibility, etc.) of imagery objects confront match costs with a variety of challenges. For example, in texture-less regions, non-parametric and window-based cost functions are suggested.

For these studies, we apply both standard benchmarks and challenging remote sensing data sets. We demonstrate the challenges such as non-Lambertian reflectance, complex scene, and radiometric changes in real-world applications. Our investigation on match costs indicates that different match cost functions are complementary in different situations. Thus, we contribute a cost-merging strategy with respect to the stereo baseline length, in order to combine advantages of different cost functions and improve matching performance in robustness and accuracy.

Currently almost all of the best stereo-matching algorithms are framed as global energy minimizations, which aim to solve the stereopsis using some regularizer – typically the smoothness of resulting disparity maps. However, such energy optimization is computed pixel-wisely

– both the cost calculation and the energy propagation. High-level priors are oft completely missing. The experience of a lot of researchers indicate that segment-based methods are preferable around object boundaries and in large homogeneous areas; the use of color segments helps propagate strong matches into sub-regions with poor matches. However, defining an energy minimization over segments, rather than pixels, imposes a hard constraint that forces depths to lie on the smooth surface associated with a region; removing fine-level details from the depth map in the process. The results are very sensitive to the given segmentations. Thus, in this dissertation we introduce a novel confidence-based surface prior for energy minimization formulations of dense stereo matching. Given a dense disparity estimation we fit planes, in disparity space, to regions of the image. For each pixel, the probability of its depth lying on an object plane is modeled as a Gaussian distribution, whose variance is determined using the confidence from a previous matching. We then recalculate a new disparity estimation with the addition of our novel confidence-based surface prior. The process is then repeated. Unlike many region-based methods, our method defines an energy formulation over pixels, instead of regions in a segmentation; this results in a decreased sensitivity to the quality of the initial segmentation. The introduced confidence-based surface prior differs from existing surface constraints in that it varies the per-pixel strength of the constraint to be proportional to the confidence in the given disparity estimation. Based on the experiment of this dissertation, useful and general conceptions for global stereopsis methods are summarized:

- Energy function formulation with additional object-level prior. Standard energy formulations using data and smoothness terms have reached their limitation of matching performance. High-level priors can inject semantical knowledge of the physical world into the goal function. However, addition of high-level priors leads oft to a hard constraints and generates artifacts.

- Pixel-wise energy optimization. To solve the global energy function, optimization should be operated over pixels. Optimization methods, which are not defined in pixel space, often impose hard constraints. Once a mis-match (in fact, a large region of mis-matches) appears, it is very difficult to eliminate it.

- Probabilistic costs fusion. An efficient way to combine additional costs into aggregated match costs is to build a probabilistic framework such that the hard constraint can be relaxed in the energy optimization.

- Iterative processing. The knowledge about the unknown result can only be groped in a gradual manner. Iterative methods allow an update or/and addition of high-level priors into the processing step-by-step.

As a concrete realization iSGM3 is introduced in this work. This iterative pixel-wise energy minimization method solves energy function defined with data, smoothness and surface priors. The cost aggregation of iSGM3 is similar to SGM and the penalty for surface constraint is fused with the path-wise aggregated costs. The addition of the confidence-based surface prior has three main benefits: sharp object-boundary edges in areas of depth discontinuity; accurate disparity in surface regions; and low sensitivity to segmentation.

Thus, our goal – developing stereo methods for challenging data – is achieved by the *cost merging* strategy depending on the baseline length and the *confidence-based surface prior* incorporating a novel energy optimization method. Both novelties are generally applicable for almost all extended stereo methods. The ideas behind our contributions may guide researchers to develop their own algorithms.

6.1 Outlook

Dense stereopsis formulated as a global energy minimization problem has been intensely investigated since last decade. Regarding to different components in global frameworks, there are still match costs functions remain to be investigated. There are also many optimization algorithms that might be considered by designing stereopsis method. This dissertation aims at developing new stereo methods that solve two problems: less robustness of match costs and absence of semantic information about the scene to be reconstructed. However, three main perspectives are of special interest for future research:

- **Stereo matching based on superpixels.** Over-segmentation is required for many existing segment-based stereo matching methods. Although the proposed iSGM3 using a confidence-based surface prior is less sensitive to a given color segmentation, artifacts do appear in segments of large areas. Superpixels can be used to replace the rigid structure of the pixel grid [Radhakrishna Achanta and Suesstrunk, 2012]. The sizes of superpixels should be smaller as color segments and more similar with each other. Moreover superpixels should adhere well to image boundaries and reduce computational complexity.

- **Jointly solves for depth and image segments.** In this dissertation color segmentation is once generated and used for the iterative processing. In fact a depth map can also be used in order to improve a segmentation in image space. Grouping or splitting segments with respect to the spatial depth-boundaries may modify incorrect segments, especially in texture-rich regions. The segmentation and depth estimation might be formulated as a joint energy minimization problem that energy of each pixel is the sum of energies encoding shape appearance and depth, taking special care into modeling occlusions [Yamaguchi et al., 2013].

- **Object detection and tracking using disparity maps.** Depth maps generate a new view of the scene. In contrast to object tracking in image space which requires detecting features and determining their correspondences, object boundaries can be better detected and tracked using disparity. Moreover, extracted disparity forms can be used to match geometric model of objects to identify interesting regions before tracking.

What makes the stereopsis interesting are its applications. For instance, stereo camera systems are used for recognizing road surface undulations in advance that the suspension of ego-vehicle can be adjusted to suit the situation [Kruse, 2014]. However, very challenging real-world data as shown in the work of Meister et al. [2012] are only restricted solvable due to the physical limitations of optical systems. In the future, to develop highly robust and accurate perception systems for understanding dynamical environments,´we will be applying additional sensors like radar/laser-scanner and fusing their data with stereo systems.

Appendix A: Data Set


Aloe


Art


Baby1


Baby2


Baby3


Books


Bowling1


Bowling2


Cloth1


Cloth2


Cloth3


Cloth4


Dolls


Flowerpots


Lampshade1


Lampshade2


Laundry

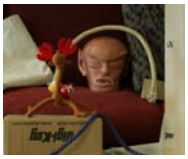

Midd1


Midd2


Moebius


Monopoly


Plastic


Reindeer


Rocks1


Rocks2


Wood1


Wood2

Figure 6.1: Left reference images of the used Middlebury data.

Figure 6.2: Rectified airborne stereo pairs with increasing baseline length.

References

Tarkan Aydin and Yusuf Sinan Akgul. Stereo depth estimation using synchronous optimization with segment based regularization. *Pattern Recognition Letters*, 31(15):2389–2396, November 2010. 35

H. H. Baker and T. O. Binford. Depth from edge and intensity based stereo. In *International Joint Conference on Artificial Intelligence*, pages 631–636, 1981. 19

Christian Banz, Sebastian Hesselbarth, Holger Flatt, Holger Blume, and P. Pirsch. Real-time stereo vision system using semi-global matching disparity estimation: Architecture and fpga-implementation. In *2010 International Conference on Embedded Computer Systems (SAMOS)*, pages 93–101, 2010. 19

Christian Banz, Peter Pirsch, and Holger Blume. Evaluation of penalty functions for semi-global matching cost aggregation. In *XXII ISPRS Congress*, 2012. 18, 36

Adrian Barbu and Song-Chun Zhu. Generalizing swendsen-wang to sampling arbitrary posterior probabilities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1239–1253, 2005. 16

Stephan Barnard. Stochastic stereo matching over scale. *International Journal of Computer Vision*, 3(1):17–32, 1989. 16

Benedichte Bascle and Rachid Deriche. Stereo matching, reconstruction and refinement of 3d curves using deformable contours. In *International Conference on Computer Vision*, pages 421–430, 1993. 19

A. Bensrhair, P. Miche, and R. Debrie. Fast and automatic stereo vision matching algorithm based on dynamic programming method. *Pattern Recognition Letters*, 17:457–466, 1996. 17

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, 36(2):192–236, 1974. 14

S. Birchfield and C. Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *PAMI*, 20(4):401 – 406, 1998. 22

S. Birchfield and C. Tomasi. Depth discontinuities by pixel-to-pixel stereo. *International Journal of Computer Vision*, pages 269–293, 1999. 17

M. Bleyer and M. Gelautz. Graph-based surface reconstruction from stereo pairs using image segmentation. In *SPIE*, volume 5665, pages 288–299, 2005. 1, 15

Michael Bleyer, Carsten Rother, and Pushmeet Kohli. Surface stereo with soft segmentation. In *CVPR*, pages 1570 – 1577, 2010. 33, 34, 35, 71

Michael Bleyer, Carsten Rother, Pushmeet Kohli, Daniel Scharstein, and Sudipta Sinha. Object stereo — joint stereo matching and object segmentation. In *CVPR*, pages 3081 – 3088, 2011. 33, 35

A. F. Bobick and S. S. Intille. Large occlusion stereo. *International Journal of Computer Vision*, 33(3):181–200, 1999. 11

Y. Boykov, O. Veksler, and R. Zabih. A variable window approach to early vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1283–1294, 1998. 26

Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001. 1

Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 26(9):1124–1137, 2004. doi: 10.1109/TPAMI.2004.60. 16, 36

A.T. Brint and M. Brady. Stereo matching of curves. *Image and Vision Computing*, 8(1):50–56, 1990. 19

Alan Brunton, Chang Shu, and Gerhard Roth. Belief propagation on the gpu for stereo vision. In *Canadian Conference on Computer and Robot Vision*, pages 26–26, 2006. 16

Xuanping Cai, Dongxiang Zhou, Ganhua Li, and Zhaowen Zhuang. A stereo matching algorithm based on color segments. In *International Conference on Intelligent Robots and Systems*, volume 3372 -3377, 2005. 35

D. Chandler and J. K. Percus. Introduction to modern statistical mechanics. *Physics Today*, 41, 1988. 16

Huahua Chen. Stereo matching using dynamic programming based on occlusion detection. In *International Conference on Mechatronics and Automation*, 2007. 17

R. Chrastek and J. Jan. Mutual information as a matching criterion for stereo pairs of images. *Analysis of Biomedical Signals and Images*, 14:101–103, 1998. 22

Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 24(5):603 – 619, 2002. doi: 10.1109/34.1000236. 36

Juan Jose de Dios and Narciso Garcia. Face detection based on a new color space ycgcr. In *International Conference on Image Processing*, volume 3, pages 909–12, 2003. 24

R. Deriche and O. Faugeras. Tracking line segments. In *European Conference on Computer Vision*, 1990. 19

Geoffrey Egnal. Mutual information as a stereo corresponce measure. In *Computer and Information Science*, number Technical Report MS-CIS-00-20. University of Pennsylvania, Philadelphia, USA, 2000. 27

F. Forbes and G. Fort. Combining monte carlo and mean-field-like methods for inference in hidden markov random fields. *IEEE Transactions on Image Processing*, 16:824–837, 2007. 16

Sven Forstmann, Yutaka Kanou, Jun Ohya, Sven Thuering, and Alfred Schmitt. Realtime stereo by using dynamic programming. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 29–29, 2004. 16, 17

Andrea Fusiello, Emanuele Trucco, and Alessandro Verri. A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, 12:16–22, 2000. 8, 9

Stefan Gehrig and Clemens Rabe. Real-time semi-global matching on the cpu. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 85–92, 2010. 19

Stefan Gehrig, Felix Eberli, and Thomas Meyer. A real-time low-power stereo vision engine using semi-global matching. In *Computer Vision Systems*, volume 5815, pages 134–143. Lecture Notes in Computer Science, 2009. 19

Stefan K. Gehrig and Uwe Franke. Improving stereo sub-pixel accuracy for long range stereo. In *International Conference on Computer Vision*, 2007. 17

A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *Computer Vision - ACCV2010*, volume 6492 of *Lecture Notes in Computer Science*, pages 25–38, 2010. 19, 20

S. German and D. German. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 6:721–741, 1984. 13, 16

Minglun Gong and Ruigang Yang. Image-gradient-guided real-time stereo on graphics hardware. pages 548–555, 2005a. doi: 10.1109/3DIM.2005.55. 36

Minglun Gong and Yee-Hong Yang. Fast stereo matching using reliability-based dynamic programming and consistency constraints. In *ICCV*, volume 1, pages 610 – 617, 2003. doi: 10.1109/ICCV.2003.1238404. 36, 38

Minglun Gong and Yee-Hong Yang. Near real-time reliable stereo matching using programmable graphics hardware. In *CVPR*, volume 1, pages 924–931, 2005b. doi: 10.1109/CVPR.2005.246. 11, 17, 26

Minglun Gong, Ruigang Yang, Liang Wang, and Mingwei Gong. A performance study on different cost aggregation approaches used in real-time stereo matching. *IJCV*, 75(2):283–296, 2007. 11, 23

J.Y. Goulermas and P. Liatsis. A new parallel feature-based stereo-matching algorithm with figural continuity preservation, based on hybrid symbiotic generic algorithms. *Pattern Recognition*, 33:529–531, 2000. 19

W. E. L. Grimson. Computational experiments with a feature based stereo algorithm. *IEEE Trans. On Pattern Analysis Machine Intelligenz*, 7(1):17–34, 1985. 19

I. Haller, C. Pantilie, F. Oniga, and S. Nedevschi. Real-time semi-global dense stereo solution with improved sub-pixel accuracy. In *IEEE Intelligent Vehicles Symposium*, pages 369–376, 2010. 18

R. M. Haralick and L. G. Shapiro. *Computer and robot vision.* Addison-Wesley Longman Publishing Co., Inc., Boston, 1992. 8

Yong Seok Heo, Kyoung Mu Lee, and Sang Uk Lee. Illumination and camera invariant stereo matching. In *CVPR*, 2008. 23

Yong Seok Heo, Kyong Mu Lee, and Sang Uk Lee. Robust stereo matching using adaptive normalized cross-correlation. *PAMI*, 33-4:807–822, 2011. 22

Simon Hermann and Reinhard Klette. Iterative semi-global matching for robust driver assistence systems. In *ACCV*, 2012. 18

Simon Hermann, Sandino Morales, Tobi Vaudrey, and Reinhard Klette. Illumination invariant cost functions in semi-global matching. volume 6469 of *Lecture Notes in Computer Science*, chapter 25, pages 245–254. ACCV 2010 Workshops, Berlin, Heidelberg, 2011. ISBN 978-3-642-22818-6. doi: 10.1007/978-3-642-22819-3_25. 18, 23

J. Höhle and M. Höhle. Accuracy assessment of digital elevation models by means of robust statistical methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(4):398–406, 2009. 49

H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *PAMI*, 30 (2):328–341, 2008. doi: 10.1109/TPAMI.2007.1166. 18, 27, 36, 38, 40, 70

H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *PAMI*, 31(9):1582–1599, 2009. doi: 10.1109/TPAMI.2008.221. 1, 4, 18, 23, 27, 34, 48

Heiko Hirschmüller and Stefan Gehrig. Stereo matching in the presence of sub-pixel calibration errors. 2009. 18

Heiko Hirschmüller, Maximilian Buder, and Ines Ernst. Memory efficient semi-global matching. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXII ISPRS Congress*, volume 1-3, 2012. 19

R. Horaud and T. Skordas. Stereo correspondence through feature grouping and maximal cliques. *PAMI*, pages 1168–1180, 1989. 19

M. Humenberger, T. Engelke, and W. Kubinger. A census-based stereo vision algorithm using modified semi-global matching and plane fitting to improve matching quality. In *CVPR Workshops*, pages 77–84, 2010. doi: 10.1109/CVPRW.2010.5543769. 18, 22

Seunghun Jin, Junguk Cho, Xuan Dai Pham, Kyoung Mu Lee, Sung-Kee Park, Munsang Kim, and Jae Wook Jeon. FPGA design and implementation of a real-time stereo vision system. *IEEE Transactions on Circuits and Systems for Video Technology*, 20-1:15–26, 2010. 2

P.-M. Jodoin and M. Mignotte. An energy-based framework using global spatial constraints for the stereo correspondence problem. In *ICIP*, volume 5, pages 3001 – 3004, 2004. doi: 10.1109/ICIP.2004.1421744. 36

John R. Jordan, Wilson S. Geisler, and Alan C. Bovik. Color as a source of information in the stereo correspondence process. *Vision Research*, 30(12), 1990. 1955-1970. 24

X.Y. Ju, J.P. Siebert, B.S. Khambay, and A.F. Ayoub. Self-correction of 3d reconstruction from multi-view stereo images, 2009. ICCV Workshop. 2

T. Kanade. Development of a video-rate stereo machine. In *Image Understanding Workshop*, page 549–557, Monterey, CA, 1994. Morgan Kaufmann Publishers. 25, 48

T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 16(9):920–932, 1994a. 11

T. Kanade and M. Okutomi. Stereo matching algorithm with an adaptive window: theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9):920–932, 1994b. 26

Jörg H. Kappes, Bjoern Andres, Fred A. Hamprecht, Christoph Schnörr, Sebastian Nowozin, Dhruv Batra, Sungwoong Kim, Bernhard X. Kausler, Jan Lellmann, Nikos Komodakis, and Carsten Rother. A comparative study of modern inference techniques for discrete energy minimization problems. In *CVPR*, 2013. 16

J. Kim, V. Kolmogorov, and R. Zabih. Visual correspondence using energy minimization and mutual information. In *International Conference Computer Vision*, 2003. 27

J.-C. Kim, K.M. Lee, B. T. Choi, and S. U. Lee. A dense stereo matching using two-pass dynamic programming with generalized ground control points. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1075-1082, 2005. 11

Andreas Klaus, Mario Sormann, and Konrad Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *ICPR*, volume 3, pages 15–18, 2006. 15, 16, 33, 34, 35, 36

Vladimir Kolmogorov and Ramin Zabih. Computing visual correspondence with occlusions using graph cuts. In *International Conference on Computer Vision*, volume 2, pages 508–515, 2001. 1, 16, 21

A. Koschan, V. Rodehorst, and K. Spiller. Color stereo vision using hierarchical block matching and active color illumination. In *International Conference on Pattern Recognition*, volume 1, pages 835–839, 1996. 35

Jochen Kruse. More comfort when cornering, June 2014. URL `http://technicity.daimler.com/en/mbc-eng/`. 81

Ritwik Kumar, Angelos Barmpoutis, Arunava Banerjee, and Baba C. Vemuri. Non-Lambertian reflectance modeling and shape recovery of faces using tensor splines. *PAMI*, 33-3, 2011. 48

F. Kurz, R. Müller, M. Stephani, P. Reinartz, and M Schroeder. Calibration of a wide-angle digital camera system for near real time scenarios. In *ISPRSWorkshop High Resolution Earth Imaging for Geospatial Information*, 2007. 46

Franz Kurz, Sebastian Türmer, Oliver Meynberg, Dominik Rosenbaum, Hartmut Runge, Peter Reinartz, and Jens Leitloff. Low-cost optical camera systems for real-time mapping applications. *PFG Photogrammetrie, Fernerkundung, Geoinformation*, 2:159–176, 2012. 2, 33

Franz Leberl, Horst Bischof, Thomas Pock, Arnold Irschara, and Stefan Kluckner. Aerial computer vision for a 3d virtual habitat. *Computer*, 43(6):24–31, 2010. doi: 10.1109/MC.2010.156. 2

Sang Hwa Lee, Jong Il Park, and ChoongWoong Lee. A new stereo matching algorithm based on bayesian model. In *International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 2769–2772, 1998. 16

Victor Lempitsky, Carsten Rother, and Andrew Blake. Logcut - efficient graph cut optimization for markov random fields. In *ICCV*, 2007. 9, 36

Maxime Lhuillier and Long Quan. Match propagation for image-based modeling and rendering. *IEEE Trans. On Pattern Analysis Machine Intelligenz*, 24:1140–1146, 2002. 20

Gang Li and Steven W. Zucker. Surface geometric constraints for stereo in belief propagation. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2355–2362, 2006. 16

S.Z. Li. Inexact matching of 3d surfaces. In *Technical Report VSSP-TR-3/90*. Vision Speech & Signal Processing, Dept. Electronic and Electrical Engineering, University of Surrey, UK, February 1990. 12

S.Z. Li. Markov random field models in computer vision. In Lecture Notes in Computer Science, editor, *Computer Vision - ECCV'94*, volume 801, pages 361–370, 1994. 13

S.Z. Li, H. Wang, and K.L. Chan. Energy minimization and relaxation labeling. *Journal of Mathematical Imaging and Vision*, 7:1–162, 1997. 13

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 8, 19

L. Matthies, T. Kanade, and R. Szeliski. Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision,*, 3(3):209–238, 1989. 25

Gerad Medioni and Ramakant Nevatia. Segment-based stereo matching. *Computer Vision, Graphics, And Image Processing*, 31:2–18, 1985. 19

C. Medrano, J. E. Herrero, J. Martínez, and C. Orrite. Mean field approach for tracking similar objects. *Computer Vision Image Understanding*, 113(8):907–920, 2009. 16

S. Meister, B. Jähne, and D. Kondermann. Outdoor stereo camera system for the generation of real-world benchmark data sets. *Optical Engineering*, 51(02):021107, 2012. 81

Dongbo Min, Jiangbo Lu, and M.N. Do. A revisit to cost aggregation in stereo matching: How far can we reduce its computational redundancy? In *ICCV*, 2011. 23

D. Neilson and Yee-Hong Yang. Evaluation of constructable match cost measures for stereo correspondence using cluster ranking. In *CVPR*, pages 1–8, 2008. doi: 10.1109/CVPR.2008.4587692. 1

D. Neilson and Yee-Hong Yang. A component-wise analysis of constructible match cost functions for global stereopsis. *PAMI*, 33(11):2147–2159, 2011. doi: 10.1109/TPAMI.2011.67. 23, 48

Daniel David Neilson. *A Study of Match Cost Functions and Colour Use In Global Stereopsis*. PhD thesis, University of Alberta, 2009. 17

Y. Ohta and T. Kanade. Stereo by intra- and interscanline search using dynamic programming. *IEEE Trans. PAMI*, 7:139–154, 1985. 17

Giorgio Parisi. *Statistical Filed Theory*. Addison-Wesley, 1988. 16

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962. 16

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Intelligent Systems: Networks of Plausible Inference, San Mateo, CA, 1988. 36

C. Peterson and J.R. Anderson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987. 16

Thomas Pock, Thomas Schoenemann, Gottfried Graber, Horst Bischof, and Daniel Cremers. A convex formulation of continuous multi-label problems. In *ECCV*, 2008. 23

S. B. Pollard, J. E. W. Mayhew, and J. P. Frisby. Pmf: A stereo correspondence algorithm using a disparity gradient limit. *Perception*, 14:449–470, 1985. 12, 24

Kevin Smith Aurelien Lucchi Pascal Fua Radhakrishna Achanta, Appu Shaji and Sabine Suesstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 2012. 81

P. Reinartz, P. d'Angelo, T. Krauß, D. Poli, K. Jacobsen, and G. Buyuksalih. Benchmarking and quality analysis of dem generated from high and very high resolution optical stereo satellite data, 2010. ISPRS Symposium Commission I. 2, 47

Erwin Riegler, Gunvor Elisabeth Kirkelund, Carles Navarro Manchón, Mihai-Alin Badiu, and Bernard Henry Fleury. Merging belief propagation and the mean field approximation: A free energy approach. *IEEE Transactions on Information Theory*, 2012. 17

L. Robert and Olivier D. Faugeras. Curve-based stereo. In *CVPR*, 1991. 19

A. Rosenfeld, R. A. Hummel, and S. W. Zucker. Scene labeling by relaxation operations. *IEEE Trans. SMC*, 6:420–433, 1976. 12

T. W. Ryan, R. T. Gray, and B. R. Hunt. Prediction of correlation errors in stereo-pair images. *Optical Engineering*, 19(3):312–322, 1980. 22, 25

Hajar Sadeghi, Payman Moallem, and S. Amirhassn Monadjemi. Feature-based dense stereo matching using dynamic programming and color. *International Journal of Information and Mathematical Sciences*, 4(3):179–186, 2008. 8, 19

Gorkem Saygili. Improving segment based stereo matching using surf key points. In *IEEE International Conference on Image Processing*, 2012. 35, 36

D. Scharstein. View synthesis using stereo vision. In *Lecture Notes in Computer Science (LNCS)*, volume 1583. Springer-Verlag, 1999. 8

D. Scharstein and R. Szeliski. Stereo matching with nonlinear diffusion. *IJCV*, 28(2):155–174, 1998. 8

D. Scharstein and R Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002. http://vision.middlebury.edu/stereo/. 1, 4, 8, 21, 22, 23, 33, 44, 45, 48

Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003. 45

C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27: 379–423, 1948. 27

Jianbo Shi and Carlo Tomasi. Good features to track. In *CVPR*, pages 593 – 600, 1994. 39

C. Strecha, R. Fransens, , and L. Van Gool. Combined depth and outlier estimation in multi-view stereo. In *Computer Vision and Pattern Recognition*, 2394-2401, 2006. 16

Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):787–800, 2003. 16, 36

Jian Sun, Yin Li, Sing Bing Kang, and Heung-Yeung Shum. Symmetric stereo matching for occlusion handling. In *CVPR*, volume 2, pages 399–406, 2005. 35, 70

R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *PAMI*, 30(6):1068–1080, 2008. doi: 10.1109/TPAMI.2007.70844. 1, 9, 16

R. Szeliski, R Zabih, D. Scharstein, and O. Veksler. Middlebury mrf minimization website, Online. URL `http://vision.middlebury.edu/MRF/`. 16

Y. Taguchi, B. Wilburn, and L. Zitnick. Stereo reconstruction with mixed pixels using adaptive over-segmentation. In *CVPR*, 2008. 33, 34, 35, 36

H. Tao and H. Sawhney. Global matching criterion and color segmentation based stereo. In *Workshop on the Application of Computer Vision*, pages 74–81, 2000. 26, 35

Marshall F. Tappen and William T. Freeman. Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *International Conference on Computer Vision*, volume 2, pages 900–906, 2003. 16

Camillo J. Taylor. Surface reconstruction from feature based stereo. In *International Conference on Computer Vision*, 2003. 8, 19

Engin Tola, Vincent Lepetit, and Pascal Fua. A fast local descriptor for dense matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 19

F. Tombari, S. Mattoccia, L. Di Stefano, and E. Addimanda. Classification and evaluation of cost aggregation methods for stereo correspondence. In *CVPR*, 2008. 23

M. Tomono. Robust 3d slam with a stereo camera based on an edge-point icp algorithm. In *ICRA*, pages 4306–4311, 2009. 19

O Veksler. Fast variable windowfor stereo correspondence using integral images. In *IEEE Conference on Computer Vision and Pattern Recognition*, page 556–561, 2003. 26

Olga Veksler. Semi-dense stereo correspondence with dense features. In *Computer Vision and Pattern Recognition*, 2001. 19

P. Viola and W. M. Wells. Alignment by maximization of mutual information. In *IJCV*, volume 24, page 137–154, 1997. doi: 10.1109/ICCV.1995.466930. 22, 27, 34

Radim Šára. Finding the largest unambiguous component of stereo matching. In *European Conference on Computer Vision*, volume 2, pages 900–914, 2002. 10

L. Wang, S.B. Kang, Shum H.-Y., and G. Xu. Cooperative segmentation and stereo using perspective space search. In *Asian Conference on Computer Vision*, page 366–371, 2004. 26

Zeng-Fu Wang and Zhi-Gang Zheng. A region based stereo matching algorithm using cooperative optimization. In *CVPR*, 2008. 33, 34, 35, 36, 38, 39, 70

Wei Wei and King Ngi Ngan. Disparity estimation with edge-based matching and interpolation. In *In IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, pages 153–156, 2005. 19, 20

Y. Wei and L. Quan. Region-based progressive stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 106–113, 2004. 35

O. J. Woodford, P. H. S. Torr, I. D. Reid, and A. W. Fitzgibbon. Global stereo reconstruction under second-order smoothness priors. *PAMI*, 31-12:2115–2128, 2009. 35, 70, 71

K. Yamaguchi, D. McAllester, and R. Urtasun. Robust monocular epipolar flow estimation. In *In Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 81

Michael Ying Yang and Wolfgang Förstner. Plane detection in point cloud data. Technical Report 1, Department of Photogrammetry Institute of Geodesy and Geoinformation, January 2010. 39

Qingxiong Yang, Liang Wang, Ruigang Yang, Henrik Stewenius, and David Nister. Stereo matching with color-weighted correlation, hierachical belief propagation and occlusion handling. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2006. 16

Qingxiong Yang, Chris Engels, and Amir Akbarzadeh. Near real-time stereo forweakly-textured scenes. In *BMVC*, 2008. 38

J. S. Yedidia, W. T. Freeman, and Y. Weiss. *Exploring artificial intelligence in the new millennium*. Morgan Kaufmann Publishers Inc, 2003. 17

K.-J. Yoon and I.-S. Kweon. Locally adaptive support-weight approach for visual correspondence search. In *IEEE Conference on Computer Vision and Pattern Recognition*, 924-931, 2005. 11

A. L. Yuille, D. Geiger, and H. Bülthoff. Stereo matching with the distinctive similarity measure. In *European Conference on Computer Vision*, 73-82, 1990. 16

Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *ECCV*, pages 151–158, 1994. 22, 28, 34

Ka Zhang, Yehua Sheng, and Chun Ye. Stereo image matching for vehicle-borne mobile mapping system based on digital parallax model. *International Journal of Vehicular Technology*, 2011. 2

Z. Zhang, D. Deriche, O. Faugeras, and Q. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence Journal*, 78:89–117, 1995. 13

K. Zhu, P. d'Angelo, and M. Butenuth. Comparison of dense stereo using cuda, 2010. ECCV, Workshop 'Computer Vision on GPUs'. 11, 19, 26

K. Zhu, P. d'Angelo, and M. Butenuth. A performance study on different stereo matching costs using airborne image sequences and satellite images. In *Photogrammetric Image Analysis, Lecture Notes in Computer Science, Springer*, pages 159–170, 2011. 19, 23, 34

S. W. Zucker, Y. G. Leclerc, and J. L. Mohammed. Continuous relaxation and local maxima selection. *IEEE Trans. PAMI*, 3:117–127, 1981. 12