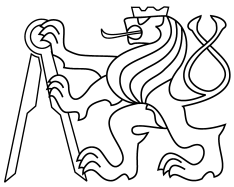
CENTER FOR
MACHINE PERCEPTION



CZECH TECHNICAL
UNIVERSITY IN PRAGUE

PhD THESIS

ISSN 1213-2365

Accurate and Robust Stereoscopic Matching in Efficient Algorithms

Jan Čech

cechj@cmp.felk.cvut.cz

CTU–CMP–2009–05

February 9, 2009

Available at
ftp://cmp.felk.cvut.cz/pub/cmp/articles/cech/Cech-thesis-2009.pdf

Thesis Advisor: Radim Šára

The research presented in this thesis was supported by the Czech Academy of Sciences under projects 1ET101210406 and 1ET101210407, by the European Union under Project FP6-IST-027113 eTRIMS.

Research Reports of CMP, Czech Technical University in Prague, No. 5, 2009

Published by

Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6, Czech Republic
fax +420 2 2435 7385, phone +420 2 2435 7637, www: http://cmp.felk.cvut.cz

Accurate and Robust Stereoscopic Matching in Efficient Algorithms

A dissertation

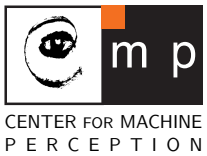
presented to the Faculty of Electrical Engineering of the Czech Technical University in Prague in partial fulfillment of the requirements for the Ph.D. degree in Study Programme No. P 2612—Electrical Engineering and Information Technology, branch No. 3902V035—Artificial Intelligence and Biocybernetics, by

Jan Čech

February 9, 2009

Thesis Advisor

Radim Šára



Center for Machine Perception
Department of Cybernetics
Faculty of Electrical Engineering
Czech Technical University in Prague

Karlovo náměstí 13, 121 35 Prague 2, Czech Republic
fax: +420 224 357 385, phone: +420 224 357 637
http://cmp.felk.cvut.cz

Research Reports of CMP, Czech Technical University in Prague, **No. 5, 2009**
ISSN 1213-2365
Published by
Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University in Prague
Technická 2, 166 27 Prague 6, Czech Republic
fax: +420 224 357 385, phone: +420 224 357 637
www: http://cmp.felk.cvut.cz

Typeset in LaTeX, February 9, 2009.

Abstract

The thesis studies dense stereoscopic techniques which are usable for accurate, robust and fast matching of high-resolution images of complex 3D scenes.

The main contributions are: (1) Image sampling invariant and affine insensitive complex correlation statistic (CCS) which is based on representing the image point neighbourhood as a response to complex Gabor filters. The CCS is a complex number with a magnitude of invariant similarity and a phase of estimated maximum position between pixels. (2) Methods for refining a disparity to sub-pixel precision - as an outcome of CCS phase, and alternatively also as a single continuous optimization problem based on a simple quadratic criterion. (3) A fast matching algorithm which avoids computing correlations for the entire disparity space by growing promising correspondence hypotheses from initial (even random) seeds. The growth is coupled with a confidently stable matching algorithm by Šára, ECCV 2002, which robustly selects the matching among competing hypotheses. (4) An algorithm for verification of given correspondences by uncalibrated dense matching. It is destined for selecting correspondences before RANSAC in challenging matching problems, with low ratio of inliers, in cluttered scenes where standard descriptor-based approach fails. An efficient procedure driven by Wald's sequential decision process grows a given correspondence while collecting statics until the decision based on learned models.

Some methods presented in the thesis go beyond the scope of 3D reconstruction, and they are applicable in many problems where the correspondences between images are sought.

Resumé in Czech

Disertační práce studuje techniky hustého stereoskopického párování, které jsou použitelné pro přesné, robustní a rychlé párování obrazů ve vysokém rozlišení zachycujících složitou 3D scénu.

Hlavní příspěvky jsou: (1) Na vzorkování obrazu invariantní a affinně necitlivá komplexní korelační statistika (CCS), která je založena na reprezentaci okolí obrazového bodu pomocí odezev Gaborových filtrů. Statistika CCS je komplexní číslo, jehož absolutní hodnota představuje invariantní podobnost obrazových okolí, fáze představuje odhad polohy maxima mezi pixely. (2) Metody pro zpřesnění disparity na pod-pixelovou úroveň - použitím výsledeku z fáze CCS a alternativně jako jediný spojitý optimalizační problém, který je založený na jednoduchém kvadratickém kritériu. (3) Rychlý párovací algoritmus, který se vyhýbá vypočítávání korelací pro celý disparitní prostor tak, že narůstá slibné korespondenční hypotézy z počátečních (dokonce i náhodných) semínek. Narůstání je spojeno se stabilním párovacím algoritmem Šára, ECCV 2002, který nakonec robustně vybere párování mezi soutěžícími hypotézami. (4) Algoritmus pro ověření správnosti dané korespondence použitím nekalibrovaného hustého párování. Algoritmus je navržen pro výběr korespondencí před procedurou RANSAC v obtížných párovacích problémech, s nízkým poměrem korespondencí vyhovujících modelu, ve scénách se složitým rušivým pozadím, kde standardní přístup založený na desktriptorech selhává. Efektivní procedura řízená Waldovým rozhodovacím procesem narůstá danou korespondenci a sbírá statistiky až do rozhodnutí, které je založeno na naučených modelech.

Některé metody představené v disertační práci sahají nad rámec 3D rekonstrukce a jsou aplikovatelné v mnoha problémech počítačového vidění, kde jsou hledány korespondence.

Acknowledgements

I am greatly indebted to my thesis advisor Radim Šára who guided me throughout my research. His friendly support, patience, and valuable feedbacks were important to the successful completion of my PhD thesis. My grateful thanks belong to Jiři Matas for his inspiring ideas and for his enthusiastic approach to the research. My thanks go to all my colleagues at Center for Machine Perception, mainly to Jana Kostlivá, Martin Matoušek, Daniel Martinec, Michal Perďoch, and Štěpán Obdržálek for many discussions and cooperation. I would like to express my heartfelt gratitude to my family, especially to my beloved wife Hana.

The research presented in the thesis was supported by Czech Academy of Sciences under projects 1ET101210406 and 1ET101210407, and by EC project FP6-IST-027113 eTRIMS. Any opinions expressed in this paper do not necessarily reflect the views of the European Community. The Community is not liable for any use that may be made of the information contained herein.

Contents

1 Introduction

This thesis deals with various correspondence problems. Points in the images are said to be *corresponding* if they are projections of an identical point in three-dimensional (3D) scene. In reality, we always search for a correspondence among pixels or larger image regions, although the correspondence is elementarily defined among image points.

The correspondence problem is one of the fundamental problems in computer vision. The tasks like 3D scene reconstruction from images, image registration, stitching, object recognition, image retrieval rely on correspondences among images.

Our primary motivation is the automatic 3D reconstruction from images. First, we will introduce the reader to basic notions and concepts of stereoscopic vision. Then we will give a brief sketch of our approach to the entire task, i.e. getting a complete 3D model of a scene from unorganized set of its images. Finally, we show where the thesis contributes.

1.1 Stereoscopic vision

The 3D scene reconstruction is generally[1] feasible using at least two images captured with a camera at displaced viewpoints. This is the case of binocular stereo.

Consider Fig. 1.1: $\mathbf{C_l}$, $\mathbf{C_r}$ are centers of projection of perspective cameras, $\pi_\mathbf{l}$, $\pi_\mathbf{r}$ are the camera image planes, where the images $\mathbf{I_l}$, $\mathbf{I_r}$ are their finite subsets with own local (pixel) coordinate system. $\mathbf{X}$ is a 3D point of the observed scene. The $\mathbf{x_l} \in \pi_\mathbf{l}$, $\mathbf{x_r} \in \pi_\mathbf{r}$ are projections of the scene point $\mathbf{X}$ onto respective camera image planes. These points are said to be *corresponding*.

The camera projection is algebraically realized as a linear mapping from scene coordinate system into local image coordinate system of vectors in homogenous representation

$$\lambda\tilde{\mathbf{x}} = \mathbf{P}\tilde{\mathbf{X}}, \tag{1.1}$$

where λ is a non-zero real scalar, the symbol $\sim$ denotes homogenous coordinates, $\mathbf{P}$ is the *camera projection matrix* of (3×4) elements, which encapsulates the camera position and orientation and the camera intrinsic transformation realizing the mapping from the

[1] Under very strong assumptions about the scene, there exist works producing a simplified and approximate 3D model from a single image, e.g. [133, 6, 65]. However, this is a completely different approach to ours.

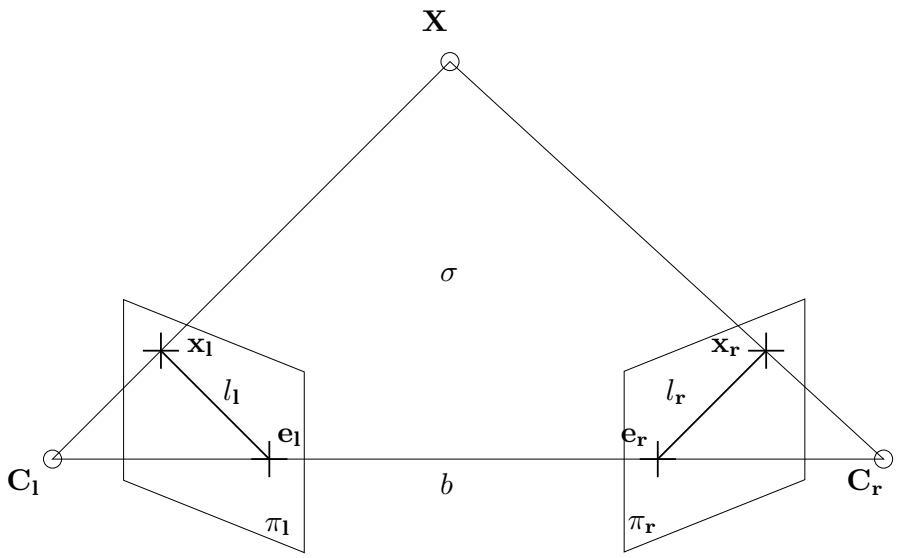


Figure 1.1: Binocular stereo setup.

camera coordinates into local image coordinate system; the parameters are focal length and pixel scales.

First, let us assume the cameras are fully calibrated. We know the projection matrices $\mathbf{P_l}, \mathbf{P_r}$, i.e. we know where $\mathbf{C_l}$, $\mathbf{C_r}$, $\pi_\mathbf{l}$, $\pi_\mathbf{r}$ lie in the coordinates of the scene. Then by intersecting the rays $(\mathbf{C_l}, \mathbf{x_l})$ and $(\mathbf{C_r}, \mathbf{x_r})$, we obtain the scene point X. The essential for the reconstruction is to establish the *matching* of the corresponding points between the images.

The line segment connecting camera centers $\mathbf{C_l}$ and $\mathbf{C_r}$ is called the *baseline*. Usually, the wider the baseline is, the more exact triangulation of $\mathbf{X}$ can be achieved, but wide baseline complicates the matching, since image neighbourhoods of corresponding points are potentially more different for more distant viewpoints.

There exists a geometric relation between the images which significantly simplifies the matching. This is the *epipolar geometry*. The triangle $\mathbf{C_l}$, $\mathbf{C_r}$ and $\mathbf{X}$ defines a plane σ, called the *epipolar plane*. The intersection of the baseline with image planes $\pi_\mathbf{l}$, $\pi_\mathbf{r}$ are the *epipoles* $\mathbf{e_l}$, $\mathbf{e_r}$. The epipoles are the images of the other camera centers. The intersection of the epipolar plane σ with image planes $\pi_\mathbf{l}$, $\pi_\mathbf{r}$ are the epipolar lines $l_\mathbf{l}$, $l_\mathbf{r}$ respectively.

An image point lying on the epipolar line in one camera lies on the corresponding epipolar line in the other camera. This constraint is called the *epipolar constraint*. Algebraically it relates the image points as,

$$\tilde{\mathbf{x}}_\mathbf{r}^T \mathbf{F} \tilde{\mathbf{x}}_\mathbf{l} = 0, \tag{1.2}$$

where $\mathbf{F}$ is the *fundamental matrix* of (3×3) elements. Thanks to the epipolar constraint,

the correspondence problem becomes a one-dimensional search along the epipolar lines only.

Moreover, having the fundamental matrix $\mathbf{F}$, the images may be geometrically transformed such that the epipolar lines coincide with the same row in both images. This process is called *epipolar rectification*. After the rectification it holds:

$$\bar{\mathbf{I}}_{\mathbf{r}}(x, y) \leftrightarrow \bar{\mathbf{I}}_{\mathbf{l}}(x + \mathbf{d}(x, y), y), \tag{1.3}$$

where $\bar{\mathbf{I}}_{\mathbf{l}}$, $\bar{\mathbf{I}}_{\mathbf{r}}$ are rectified images, y is image row coordinate, x is image column coordinate and $\mathbf{d}(x, y)$ is called the *disparity*. The disparity in the current point (x, y) is inversely proportional to the depth of the scene. The "$\leftrightarrow$" is a symbol of the correspondence. The map of the same size as image $\bar{\mathbf{I}}_{\mathbf{l}}$, where a pixel (x, y) has a disparity $\mathbf{d}(x, y)$ we call a *disparity map* $\mathbf{d}$. Note that disparity may not be defined for certain pixels, e.g. occluded.

A different situation occurs when neither the calibration nor the epipolar geometry are known. Then one needs to use the corresponding points as well to estimate the calibration. This is a task of so called Wide Baseline Stereo (WBS). Using (1.2), the fundamental matrix $\mathbf{F}$ can be directly estimated from several corresponding points. Also, camera matrices $\mathbf{P}_{\mathbf{l}}, \mathbf{P}_{\mathbf{r}}$ can be estimated from corresponding points. These methods are described in e.g. [60], where minimal number of correspondences and degenerate configurations are analyzed.

Since the epipolar geometry is unknown, the correspondence problem remains a two-dimensional search, which is susceptible to mismatches. Therefore a robust estimation of the model is necessary. A usual approach is to generate a set of promising tentative correspondences and then fit the model in a robust manner using methods like RANSAC [40] to detect the mismatches, which are outliers from the model. This procedure is exponentially dependent on the complexity of the model, therefore minimal number of correspondences to determine the model is desired. And it is polynomial in the ratio of outliers, therefore obtaining high quality tentative correspondences is important.

1.2 Difficulty of the correspondence problem

As we have seen, a correspondence problem is crucial for the 3D reconstruction and appears in both cases with and without the epipolar geometry. However, matching the correspondences in the images is generally a tough ill-conditioned problem. The main sources of difficulty are the following:

- **Non-Lambertian scene**

  Lambertian bidirectional reflectance distribution function (BRDF) [110] is required. It means the surface point radiance is uniform with respect to viewing direction. This is an abstraction which can never be fulfilled. The model is a

good approximation for opaque dielectric surfaces with subsurface scattering and no boundary reflection.

The only known method for visual stereopsis with arbitrary BRDFs of the scene is using a special configuration of illumination and camera, exploiting the Helmholtz reciprocity principle. Loosely, it says that swapping the position of the light source and the camera, the outgoing radiance is the same for both original and swapped camera positions regardless the surface BRDF [184].

- **Constant intensity regions in the images**

  There is an inherent ambiguity in stereo. Baker et al. [5] demonstrated this fact constructing the same lightfield (set of all measurements that could be used by a stereo algorithm) via albedo (the ratio of reflected to absorbed power) alteration for different depth object surfaces. Weakly-textured regions, regions with low SNR are ambiguous, since their image points are not distinguishable for matching.

- **Repetitive pattern in the images**

  When there is a repetitive texture in the images, the matching may be ambiguous, due to pattern self-similarity. The probability of occurrence of such event can be reduced using more (calibrated) cameras for matching [115].

- **Occlusions**

  There may exist parts of the scene which are visible in one camera only and not in the other. Then the correspondence does not exist for the occluded regions. Matching algorithm must tackle this problem. This may be caused by an object in front of the background or an abrupt change in the scene depth.

  In a multi-view case, there are even more types of occlusions. The scene part may be visible by a subset of the cameras only.

There are several other causes of difficulty, e.g. noise in the images, discretization artifacts, perspective or non-linear distortion of the images, etc.

1.3 From unorganized images to a 3D model

In this section, we briefly introduce an automatic 3D reconstruction pipeline which was developed in Center for Machine Perception [31, 69]. The algorithm produces a 3D model from a set of unorganized images.

There exist other 3D reconstruction pipelines in literature, e.g. the method [1] produces a dense 3D model of urban environment from uncalibrated video sequences, or Microsoft's product Photosynth[2] started as Photo tourism project [144] which works with Internet photo-collections. The original idea was that the algorithm calculates the camera position and a sparse reconstruction and a user navigates inside the virtual

[2] `http://livelabs.com/photosynth/`

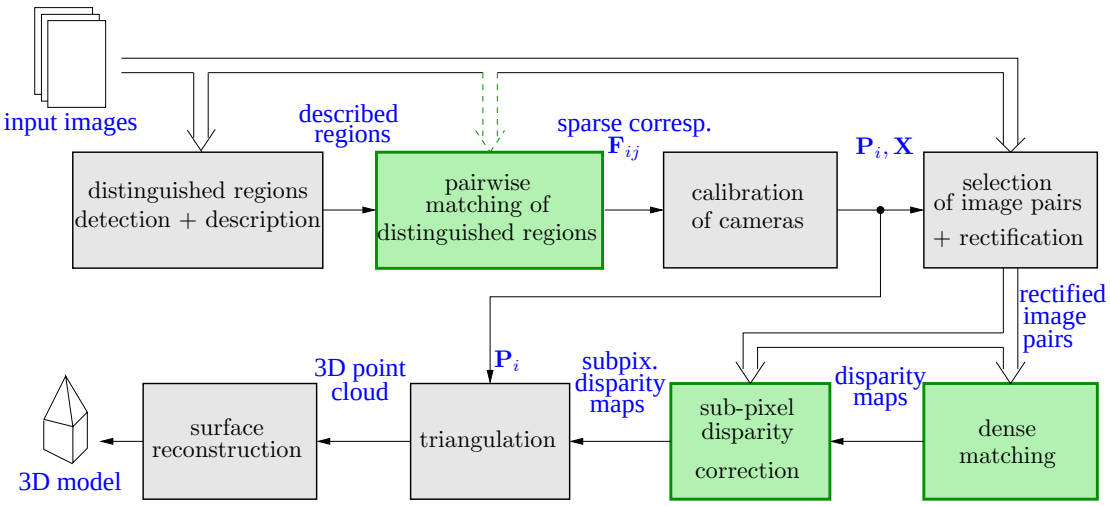


Figure 1.2: 3D reconstruction data processing pipeline. Blocks where the thesis contributes are highlighted by a green color.

Figure 1.3: Input images of Henri Miller's statue L'écoute in Paris. The entire set used for 3D reconstruction contains 26 images.

space and the system then offers the closest virtual view from a set of stored images. Recently, the authors added a dense multi-view stereo to produce a dense 3D model automatically [52].

A flowchart of our 3D reconstruction pipeline is sketched in Fig. 1.2. The algorithms works with unorganized images, which are captured by a consumer camera. An example is shown in Fig. 1.3. Notice, the object is occluded by people walking around and climbing the statue. This occurs in many images. However, the algorithm is robust to this type of occlusions too.

First, distinguished regions are detected in all n input images. These are the regions which are repeatably-detectable under image transformations induced by various viewpoints. We used Maximally Stable Extremal Regions (MSERs) [99] or Hessian Affine points [134]. An example of distinguished regions detected in a stereopair is in Fig. 1.4.

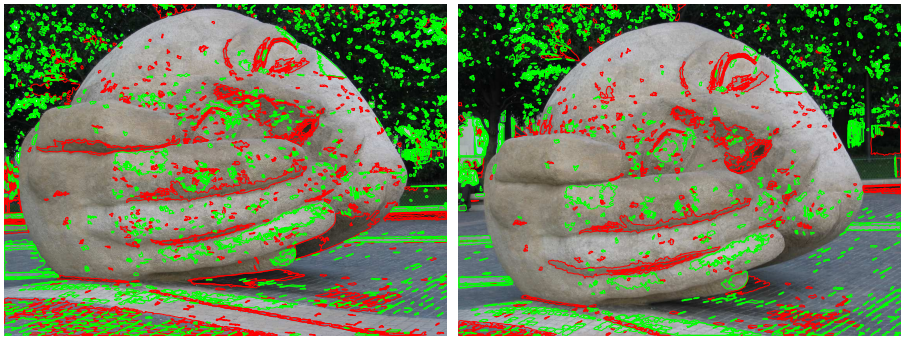


Figure 1.4: Distinguished regions detected in the images of a stereopair; MSER+ are in green and MSER- in red color [99].

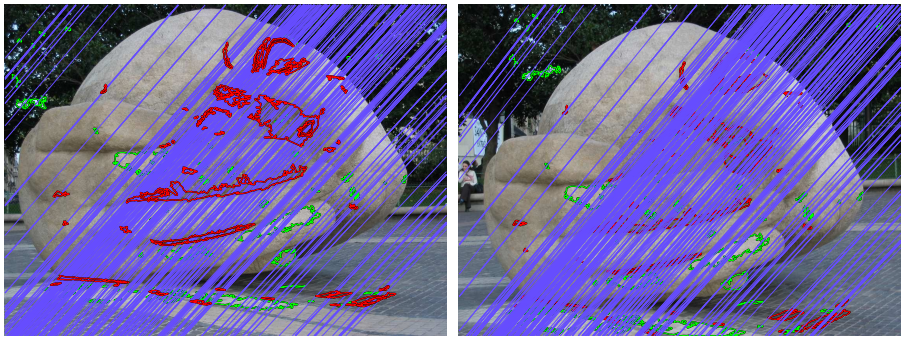


Figure 1.5: The stereopair with estimated epipolar geometry; distinguished regions which are inliers of the epipolar geometry, epipolar lines in blue color.

These regions are geometrically normalized according to regions's local affine coordinates constructed from local characteristics of the region. The normalized regions are then described with SIFT image descriptor [95]. The descriptor carries the viewpoint invariant (modelled as affine transformation invariant) information about the region's neighbourhood in the image.

Next, under no assumptions about the acquisition of images, each of $\binom{n}{2}$ image pairs are attempted to match in order to establish their epipolar geometry.[3] Loosely speaking, the pairs of regions with closest descriptors are inserted into a set of tentative correspondences for each image pair [99]. The epipolar geometry is found by fitting the model of fundamental matrix $\mathbf{F}$ in (1.2) into the set of tentative correspondences using RANSAC. The estimated epipolar geometry for the stereopair is shown in Fig. 1.5. However, before the RANSAC is applied, we propose to verify the tentative correspondences first using an efficient dense matching-based algorithm which tries to match image neighbourhoods of tentative correspondences and evaluates their quality. This is important in difficult

[3] We are aware that this is too expensive for a large set of images. This can be avoided by image retrieval techniques, clustering the space of descriptors, e.g. [143, 121].

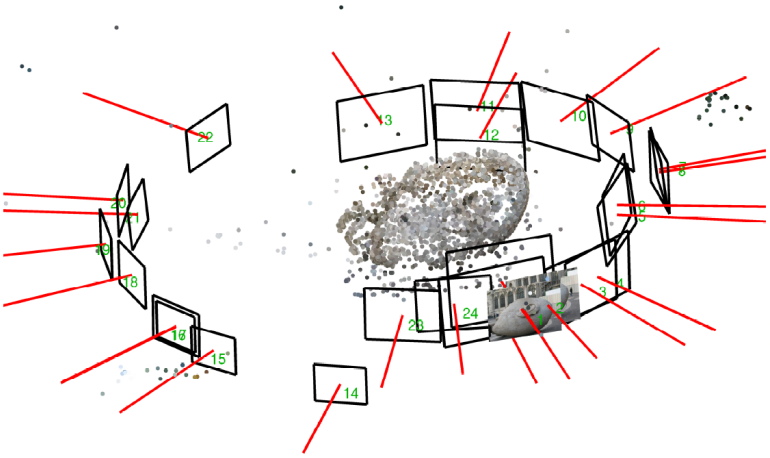


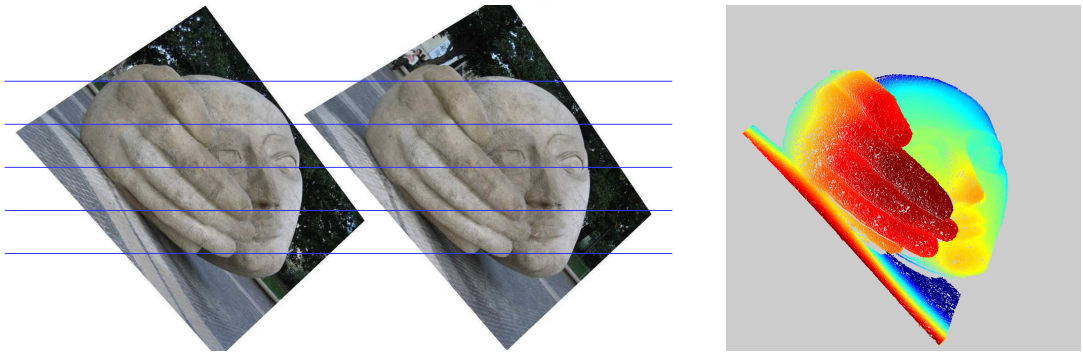
Figure 1.6: Estimated cameras $\mathbf{P}_i$ and sparsely reconstructed points $\mathbf{X}$. *Courtesy Daniel Martinec.*

matching problems, where compact descriptors do not perform well, problems with high ratio (above 90 percent) of outliers among tentative correspondences, typically complex 3D structure with many occlusions. This method is described in Chapter 6, originally published in [23].

The correspondences between pairs of images which are inliers of the epipolar geometries are used for calibration of cameras. This is done by an algorithm described in [98, 97]. The procedure robustly estimates the camera matrices $\mathbf{P}_i$ and sparse 3D points $\mathbf{X}$. The estimate of camera parameters and 3D points is refined by bundle-adjustment [157], which minimizes the reprojection error, i.e. the least squares error between the points detected in the images and reprojection of reconstructed points by estimated cameras. Reconstructed cameras together with sparse 3D points for the exemplar scene are shown in Fig. 1.6.

Next, suitable image pairs for subsequent rectification and dense matching are selected. These are the image pairs with large number of sparse correspondences. The selected pairs are epipolarly rectified, i.e. the images are warped by homographies such that the epipolar lines are aligned with common image rows. A review of the rectification methods can be found in [101]. The rectification is important since it makes the distortion of local image neighbourhoods due to slanted surfaces well defined and allows the subsequent dense matching algorithm to be computationally efficient. The rectified stereopair from Fig. 1.5 is shown in Fig. 1.7(a).

The dense matching algorithm finds correspondences between pixels. For each rectified image pair, it produces a disparity map, see (1.3). We use the algorithm described in Chapter 5 which was originally published in [26]. The disparity map for the above stereopair is shown in 1.7(b). Our approach is that the dense matching algorithm must not make too many errors, since the matching errors make the 3D model reconstruction difficult. However, the effort of making fewer matching errors is at the cost of a lower



(a) (b)

Figure 1.7: The rectified stereo-pair with epipolar lines shown in a blue color (a). The color-coded disparity map; warmer colors correspond to higher disparities, colder color to lower disparities, gray color means unassigned disparity (b).

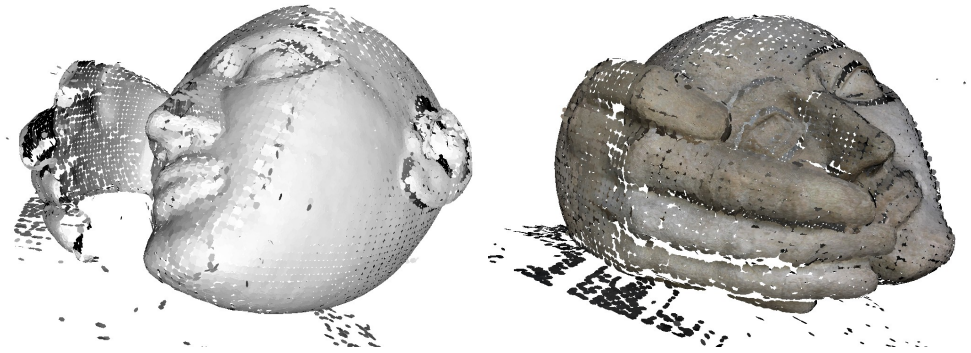


Figure 1.8: Resulting 3D model as fish-scales [132]. The first view is untextured while the other is textured.

density of the disparity map. The matching algorithm has the 'reject option', i.e. it does not decide a disparity in case of ambiguous data.

Integer disparity maps are refined to sub-pixel precision. Although we are using relatively high-resolution images (above 1 MPx), the sub-pixel disparity correction visually improves the 3D model. This is possible by several approaches, our methods are described in Chapter 4, and partially published in [25].

Resulting sub-pixel disparity maps are 'unrectified', i.e. the correspondences are recalculated with respect to images before rectification, and triangulated using the estimated cameras. This results in a dense 3D point cloud.

The simplest version of the surface reconstruction is based on 'fish-scales' [132]. Loosely, they are small planar discs tangent to the surface, spanned by two largest eigenvectors of covariance matrices found by the K-means algorithm. This simple method reduces the amount of data without significant loss of detail and it is able to eliminate certain mismatches. The resulting 3D fish-scale model of the exemplar scene is shown in Fig. 1.8 in two different views both untextured and textured. If needed, the triangulated mesh is obtained by standard methods [4, 73].

The cardinal techniques which will be described in the thesis are related to dense stereoscopic matching algorithms. Therefore, we first review the state-of-the-art methods, identify the problems and formulate the program of the thesis.

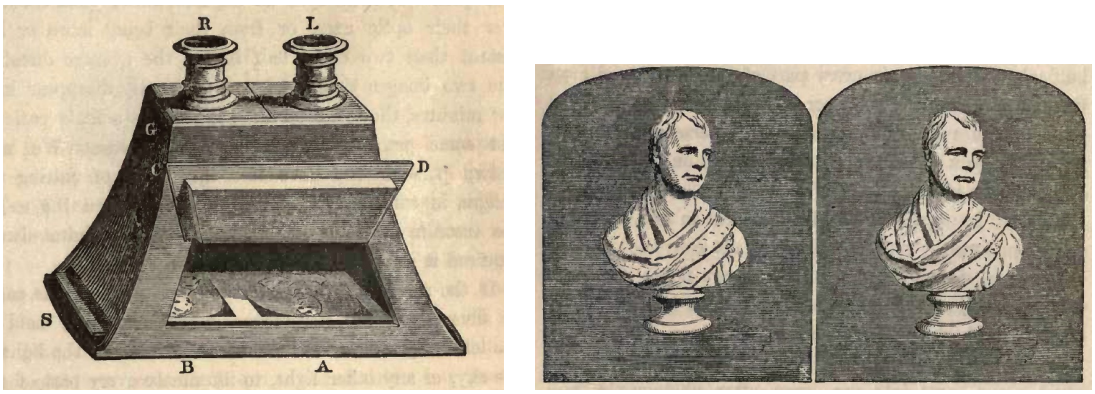
2 State of the art

This chapter overviews state of the art in the dense stereoscopic matching. The following description is not intended to be complete due to enormous number of research papers in the field, however we will try to characterize the fundamental methods which represent different paradigms. There exists several papers reviewing the stereo algorithms, e.g. [36, 80, 19].

The adjective *stereoscopic* comes from a *stereoscope* which is derived from two Greek words *στερεός* meaning *solid* and *σκοπεῖν* which means *to observe*. The stereoscope is an instrument which allows a human to perceive a depth of the virtual scene from two pictures with a disparity. The stereoscope was invented and named by Wheatstone [172] in 1838 and further improved by Brewster [18], see Fig. 2.1. Nevertheless, the binocular projection was studied even earlier, which is referred in [18], by Euclid in ancient Greece and later for instance by Leonardo da Vinci in renaissance. A large development of stereoscopy followed after the invention of photography by Niépce and Daguerre around 1830. The photography initiated a discipline closely related to stereoscopy called a *photogrammetry* around 1850 [21]. The photogrammetry employs methods for estimating 3D coordinates of the scene from 2D measurements in two or more of its photographs, i.e. basically, identification of correspondences and triangulation.

The research of stereoscopic vision in the 20th century was firstly driven by neuro-physiologists and psychologists studying the human stereo-vision, Julesz [68], and an effort to construct stereo algorithms which mimic it, Marr [96].

The early beginnings of stereoscopic matching were based on matching distinctive features (regions) of images, e.g. edges, corners, zero-crossings of local image Laplacian, etc. The reason was the low computational power of machines in the seventies of the last century. The results were sparse, a disparity was assigned to few pixels only. The review of these *feature-based* methods can be found in [36]. Nowadays, feature-based methods are used in wide baseline stereo (WBS) as discussed in the previous chapter. We will not discuss feature-based methods in detail. We are interested in *dense-matching*, i.e. in algorithms which find correspondences between pixels as densely as possible up to occlusions or ambiguity.



(a) (b)

Figure 2.1: The stereoscope (a) and a pair of stereo-images (b). A trained person is able to perceive the depth of the stereo-pair directly by crossing the eyes. *Reproduction from [18].*

2.1 Taxonomy of dense matching methods

In the following section we will try outlining a taxonomy of the dense matching methods. The taxonomy is not strict and captures our view of the field.

There exist two basic approaches in dense matching methods: First, methods formulating the task as an explicit global optimization problem incorporating the prior model (A). Second, the methods where no explicit prior is employed and the correspondences are selected according to a certain principle (B). In the middle, there are heuristic methods, which try to improve the matching iteratively.

2.1.1 Energy minimization methods (A)

These methods are formulated as an inverse problem with a shape prior. They are also sometimes called *global* since the matching task is formulated as an optimization of a single criterion. The inherent ambiguity of the stereo matching is solved by regularization via incorporating a prior model of the scene.

The common justification of these methods is a probabilistic formulation in a Bayesian framework. The optimal disparity map $\mathbf{d}^*$ has the maximum aposteriori probability (MAP), i.e. the most probable solution given the input images $\mathbf{I_l}$ and $\mathbf{I_r}$

$$\mathbf{d}^* = \arg\max_{\mathbf{d}\in\mathcal{D}} p(\mathbf{d} \mid \mathbf{I_l}, \mathbf{I_r}). \tag{2.1}$$

We denote $\mathcal{D}$ a domain of disparity map $\mathbf{d}$, see (1.3). Using the Bayes law, the problem becomes

$$\mathbf{d}^* = \arg\max_{\mathbf{d}\in\mathcal{D}} p(\mathbf{I_l}, \mathbf{I_r} \mid \mathbf{d}) p(\mathbf{d}). \tag{2.2}$$

The data term, which reflects an agreement of the solution with data, is of the form

$$p(\mathbf{I_l}, \mathbf{I_r} \mid \mathbf{d}) = \frac{1}{Z_D} e^{-\frac{D(\mathbf{I_l}, \mathbf{I_r}, \mathbf{d})}{\lambda_D}}, \tag{2.3}$$

where $D(\mathbf{I_l}, \mathbf{I_r}, \mathbf{d})$ is a non-negative function which measures how well the current disparity map $\mathbf{d}$ maps the images $\mathbf{I_l}$, $\mathbf{I_r}$ onto each other, λ_D is the scale (dispersion of the distribution) and Z_D is a normalization constant.

The prior term is of the form

$$p(\mathbf{d}) = \frac{1}{Z_R} e^{-\frac{R(\mathbf{d})}{\lambda_R}}, \tag{2.4}$$

where $R(\mathbf{d})$ is a non-negative regularizing function which penalizes the current disparity map $\mathbf{d}$, λ_R is the scale, and Z_R is a normalization constant.

Substituting terms (2.3) and (2.4) into (2.2), applying logarithm, setting $\lambda_D = 1$ and renaming $\lambda = \frac{1}{\lambda_R}$, we get the original problem in the energy minimization form

$$\mathbf{d}^* = \arg\min_{\mathbf{d} \in \mathcal{D}} \Big[D(\mathbf{I_l}, \mathbf{I_r}, \mathbf{d}) + \lambda R(\mathbf{d}) \Big] = \arg\min_{\mathbf{d} \in \mathcal{D}} E(\mathbf{d}), \tag{2.5}$$

where a non-negative scalar λ is the parameter which controls a relative strength of the prior term.

Now, according to the domain $\mathcal{D}$ of the disparity map which is considered, we further distinguish the energy minimization methods. If domain $\mathcal{D}$ is a space of *discrete* functions of a discrete variable, we are speaking about discrete optimization methods (A.1). If domain $\mathcal{D}$ is a space of *continuous* (differentiable) functions, we are speaking about variational methods (A.2).

Discrete optimization methods (A.1) The domain $\mathcal{D}$ is considered a space of discrete functions of discrete variables. Then, the disparity map is a finite set of pixels with assigned discrete labels. We denote L a finite set of possible labels, which consists of a range of integer disparity values, and possibly some extra labels for occlusions and missing correspondences. We denote $d_t \in L$ a label which is assigned at pixel $t \in T$ located at position (x, y), where T is a finite set of all disparity map pixels. Linearized notation d_t is a shortcut for $d_{x,y}$. The problem is often modeled using a Markov Random Field (MRF). The data term (2.3) factorizes over pixels

$$p(\mathbf{I_l}, \mathbf{I_r} \mid \mathbf{d}) = \frac{1}{Z_D} \prod_t e^{-\frac{\delta(\mathbf{I_l}, \mathbf{I_r}, d_t)}{\lambda_D}}, \tag{2.6}$$

where $\delta(\mathbf{I_l}, \mathbf{I_r}, d_t)$ is unary potential, a function which measures similarity of pixel intensities for disparity at pixel t. For instance using Gaussian error distribution, its simple

form is $\delta(\mathbf{I_l}, \mathbf{I_r}, d_t) = \delta(\mathbf{I_l}, \mathbf{I_r}, d_{x,y}) = \big(\mathbf{I_l}(x+d_{x,y}, y) - \mathbf{I_r}(x, y)\big)^2$. The prior term factorizes over cliques (pairs of neighbouring pixels t,t')

$$p(\mathbf{d}) = \frac{1}{Z_R} \prod_{t,t'} e^{-\frac{\rho(d_t, d_{t'})}{\lambda_R}}, \tag{2.7}$$

where $\rho(d_t, d_{t'})$ is binary potential, a function reflecting compatibilities of labels ar neighbouring pixels. For instance, the simple example is $\rho(d_t, d_{t'}) = 0$ if $d_t = d_{t'}$, otherwise $\rho(d_t, d_{t'}) = 1$. The prior then becomes Potts model [7].

Performing the same manipulation as above, the original problem written in the energy minimization form is

$$\mathbf{d}^* = \arg \min_{\mathbf{d} \in L^{|T|}} \sum_t \delta(\mathbf{I_l}, \mathbf{I_r}, d_t) + \lambda \sum_{t,t'} \rho(d_t, d_{t'}). \tag{2.8}$$

All these methods typically use very simple data models, such as the mentioned squared difference of pixel intensities, or some robust alternative of it. Various methods differ in the way they model occlusions (by introducing extra labels) or the methods differ in the algorithm how they solve the discrete optimization problem (2.8).

The problem (2.8) is NP-hard in general. The brute-force approach is of complexity $\mathcal{O}(|L|^{|T|})$. There exist solvable sub-classes of the problem which are applicable to stereo [42, 138]. Solvability depends on the problem topology, structure of label set and the form of pairwise potentials. For the remaining problems, one can use approximate algorithms which find a sub-optimal solution. This is e.g. (loopy) belief propagation [120, 38], Kolmogorov's TRW-S algorithm [75], max-sum diffusion by Kovalevsky and Flach [86, 41], Schlesinger's linear programming relaxation [137, 170, 171] or recent method via dual decomposition by Komodakis et al. [79].

Other approach to avoid NP-hard optimization is to decompose 2D problem into a collection of independent problems along scanlines, i.e. image rows which are epipolar lines in rectified images. It leads to a search of the shortest path through the correlation table[1] with an extra penalty for occlusions. This is found by dynamic programming. The early representatives are by Gimel'farb [51] and Cox et al. [32].

Belhumeur [8] proposed three different prior models: smooth surfaces, piecewise smooth surfaces, and piecewise smooth surfaces with steeply sloping surfaces. He also proposed an extension to this approach incorporating a inter-scanline (vertical) smoothing via iterative optimization initialized from the independent-scanline solution.

Bobick [14] noticed a large sensitivity of the solution to the cost of occlusions and proposed using Ground Control Points, it means highly-reliable matches. Sensitivity to the occlusion penalty is reduced, as well as the computational complexity. Bobick also proposed exploiting the possible coincidence of occlusion boundaries with intensity discontinuities in the images.

[1] Correlation table is a set of all potential matches with assigned (dis)similarity.

Torr and Criminisi [156] pointed out that traditional approach using separate scanline dynamic programming suffers from streaky disparity map. To avoid these artifacts, the authors try to exploit the information from previously applied corner and edge detector to establish "pivots." The trajectory is not forced to go through the pivots, they have set a low cost in order to attract the path to them. They observed a speedup of the algorithm due to the pivoting.

Gong and Yang [54] proposed reliability based dynamic programming. The reliability of each potential match is computed, such that it is the cost difference of the best path that goes through the match and the best path that does not go through that match. Hereby, there is a reliability threshold which rejects ambiguous matches, beside the parameter for smoothness. This thresholding is in fact an incomplete version of stable matching [129]. Stability alone as a matching criterion has been used by Šára, as will be discussed in Sec. 2.1.2.

Several authors made an effort to approximate the isotropic 2D prior encoded in the four connected neighbourhood graph of MRF by its acyclic sub-graph, which is optimized using dynamic programming. The idea is that the dynamic programming is run along different directions, not only along epipolar lines and the cost of the paths is aggregated. This approach is applied for instance by Mozerov [106] or recently by Hirschmüller [63] in a successful algorithm applied to high-resolution aerial images of cities. Alternatively, similar mechanism with trees are presented for instance by Veksler [165] or by Bleyer and Gelautz [13]. This approach is promising, since it is fast and results are still good.

A relatively successful is the algorithm which transforms the optimization (2.8) into finding a minimum cut in a graph constructed above image pixels. Boykov et al. [17] demonstrate a stereo algorithm via energy minimization with graph cuts. It was one of the first proper formulations with fully isotropic prior model. The class of functions which can be optimized via graph cuts is specified by Kolmogorov and Zabih [78]. The authors designed stereo matching algorithm with explicitly modeling occlusion using this optimization technique [76, 77]. Graph cuts have become a popular technique which is applied in various stereo algorithms combined with other methods or monocular cues. For instance, high-quality results are reported by Bleyer and Gelautz [12], where the image segmentation is employed. It positively effects the accuracy of occlusion boundaries and the regions of a weak texture.

The drawback of graph cuts is a relatively narrow class of problems where this algorithm is applicable. In certain methods, the problem (2.8) is optimized by belief propagation, e.g. the algorithm by Sun et al. [149] or by Felzenszwalb and Huttenlocher [38]. The problem of belief propagation is that the convergence is not guaranteed in graphs containing cycles, which is the case of problem (2.8). Tappen and Freeman [153] compared the optimization of problem (2.8) with identical parameters using graph cuts with belief propagation in a controlled setup with ground truth. They concluded that results are comparable, although Graph Cuts finds a disparity map with lower energy, not necessarily better with respect to the ground truth. Another similar comparative study is given by Szeliski et al. [152]. The authors compare minimization with graph cuts, belief

propagation and TRW-S algorithm in various benchmark discrete minimization tasks, including stereo.

The main drawback of the methods using a strong prior model is their tendency to smooth out occlusion boundaries and interpolate through regions with significantly low data support. This regularization of the matching problem is often erroneous and misleading for a structure reconstruction problem. The prior used is usually too simple and does not capture the natural scene statistics well. This is due to a difficulty to design and to minimize more complex priors. Nevertheless, methods that learn the priors has appeared, e.g. Scharstein and Pal [135] learn parameters of the (conditional) MRF from ground truth dataset. Recently, Woodford et al. [173] reported an improvement when incorporating second-order smoothness, which requires higher-order cliques in MRF.

What is also missing is a reject option, since in many applications the algorithm should not assign a doubtful disparity. It should identify the ambiguous regions instead. However, this is not easy to introduce into MAP methods.

The consequent artifacts are demonstrated in Fig. 2.2 and Fig. 2.3. There are two representatives of class A methods (employing a strong prior model): independent scan-line search using a dynamic programming by Cox et al. [32] Fig. 2.2(a), Fig. 2.3(a) and a method with isotropic prior using graph cuts by Kolmogorov and Zabih [76] Fig. 2.2(b), Fig. 2.3(b). The parameters were tuned to give visually the best results. Notice the streaks in the dynamic programming results and assigning an incorrect disparity in the regions of the sky where there is no data. Graph cuts produce a disparity map with a few patches of a constant disparity, which is not accurate and mostly erroneous, e.g. the regions around the the apple tree trunk and poor results for the larch grove. On the other hand class B methods, described later, represented here by Confidently Stable Matching algorithm [129] Fig. 2.2(c), Fig. 2.3(c) and GCS matching [26] (described in Chapter 5 of the thesis) Fig. 2.2(d), Fig. 2.3(d) do not suffer from such artifacts.

Variational methods (A.2) The following methods consider the domain of disparity map a space of continuous differentiable functions and express the problem (2.5) as a variational task

$$\mathbf{d}(x,y)^* = \arg\min_{\mathbf{d}(x,y)} \iint D\big(\mathbf{I_l}, \mathbf{I_r}, \mathbf{d}(x,y)\big) + \lambda R\big(\nabla \mathbf{d}(x,y)\big) dxdy. \tag{2.9}$$

To solve the problem, Euler-Lagrange equation, a necessary condition, is written. This is a partial differential equation (PDE). There are two main approaches to solve the PDE. It can be solved such that $\mathbf{d}$ becomes also a function of time $\mathbf{d}(x,y,t)$. It results in a gradient descent diffusion process which starts from an initial estimate of disparity map $\mathbf{d}(x,y,0) = \mathbf{d}_0(x,y)$ and should converge to the optimum $\mathbf{d}^*$ when time $t \to \infty$. This is for instance the case of work of Robert and Deriche [126], Alvarez et al. [2] or Strecha et al. [147]. Alternatively, the solution of PDE is found by the level set method, see e.g. Deriche et al. [34], Faugeras and Keriven [37], or more recently Lhuiller and Quan [94].

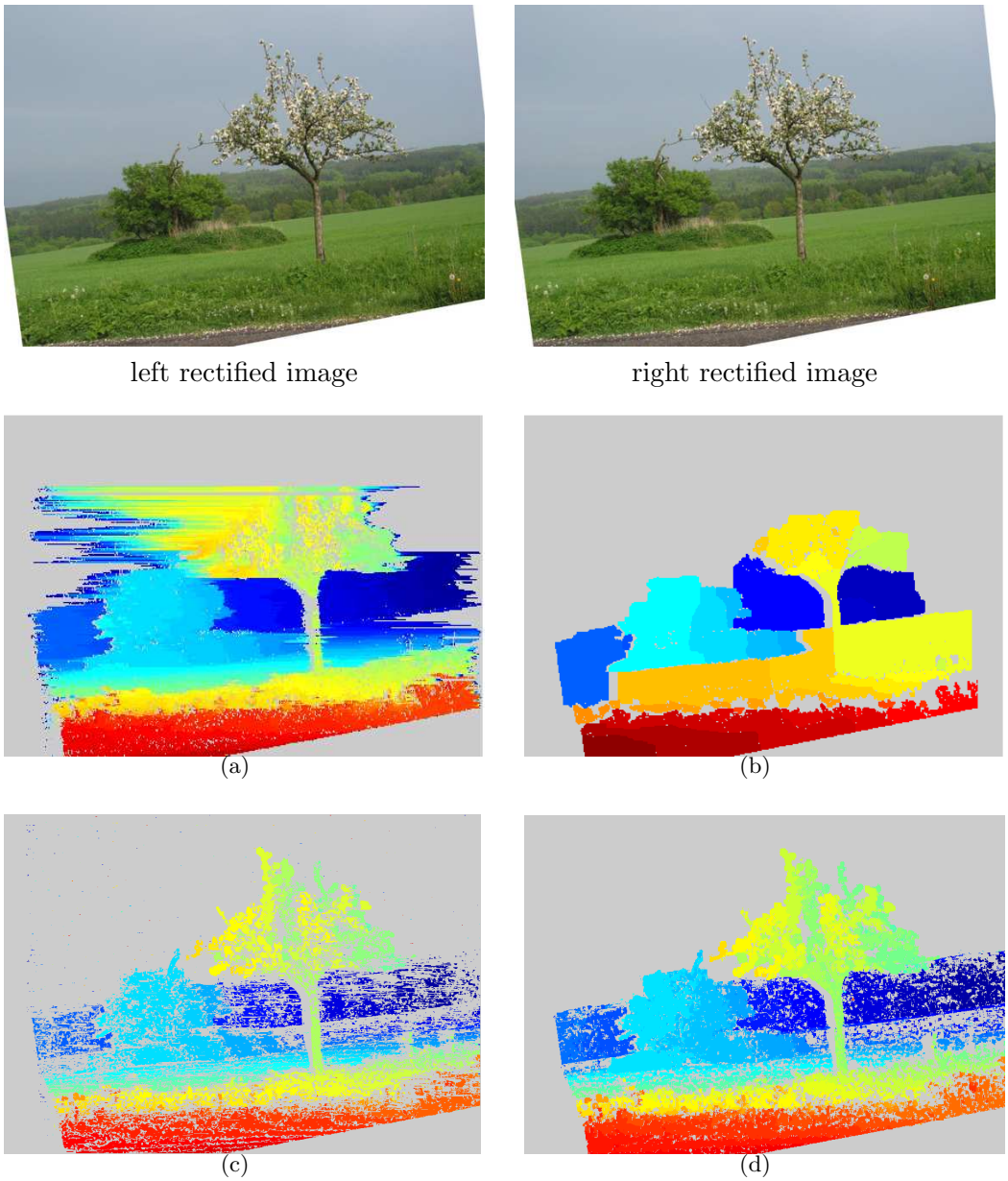


Figure 2.2: Apple tree data set. Disparity maps from (a) Dynamic Programming [32], (b) Graph Cuts [76], (c) Confidently Stable Matching [129], (d) Growing Correspondence Seeds [26] (the proposed method described in Chapter 5). *Disparity maps (a) and (b) by courtesy of Jana Kostlivá.*

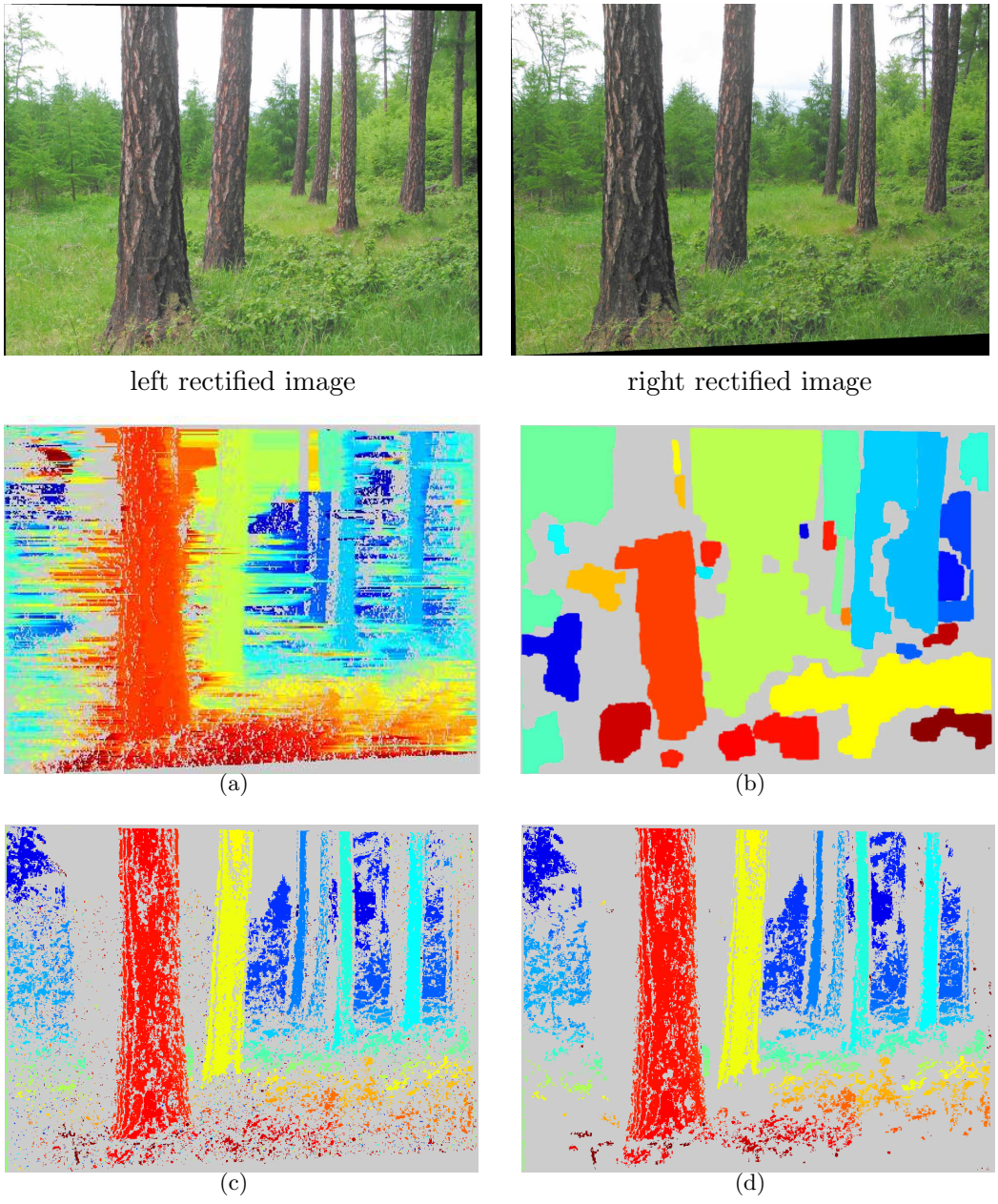


Figure 2.3: Larch grove data set. Disparity maps from (a) Dynamic Programming [32], (b) Graph Cuts [76], (c) Confidently Stable Matching [129], (d) Growing Correspondence Seeds [26] (the proposed method described in Chapter 5). *Disparity maps (a) and (b) by courtesy of Jana Kostlivá.*

Variational methods are suitable for reconstruction of a single surface, rather than for reconstruction of a scene which contains multiple objects. They are often used in multi-view surface reconstruction. A criticism is that they produce over-smoothed results, although modern regularizers preserve depth discontinuities correctly, e.g. Zimmer et al. [185]. Another criticism of these methods is their reliance on a proper initialization and potential stucking in a local extremum. However, impressive results of reconstructed surfaces from multiple-views with many fine details are shown by Strecha [147], where the initialization is from a sparse correspondences used for camera calibration. The algorithm proceeds in a multi-resolution fashion, where the surface estimated at the coarser level of the image initializes the process at the finer level.

On the other hand, an advantage of these methods compared to discrete optimization methods is, besides the natural sub-pixel accuracy, an algorithmic efficiency in the sense of memory resources and also computational time. Hereby, they allow to work with high-resolution images and thus achieve higher accuracy, as also pointed out by Strecha [146].

2.1.2 Discriminability based correspondence selection (B)

These methods are formulated as a constrained correspondence recognition problem. They proceed such that they recognize what belongs together based on the sufficient discriminability of the image point description without an explicit prior model of the scene. They use only simple geometrical constraints, like uniqueness and ordering[2].

Each image point is described by a feature vector. It is often a vector containing pixel intensities of a square neighbourhood of the image point (a window) or other more sophisticated descriptors. Then the algorithms construct a matching table, which is a set of all potential matches together with their similarity statistics computed over the feature vectors. The similarity statistic used is often SSD (Sum of Squared Differences), SAD (Sum of Absolute Differences), NCC (Normalized Cross Correlation), MNCC (Moravec's Normalized Cross Correlation) [105], etc.:

$$\mathrm{SSD}(\mathbf{W}_i, \mathbf{W}_j) = \sum_{k=1}^{n} (\mathbf{W}_i(k) - \mathbf{W}_j(k))^2, \tag{2.10}$$

$$\mathrm{SAD}(\mathbf{W}_i, \mathbf{W}_j) = \sum_{k=1}^{n} |\mathbf{W}_i(k) - \mathbf{W}_j(k)|, \tag{2.11}$$

[2] *Uniqueness* constraint says that the matching algorithm can establish one to one correspondences only, i.e. a correspondence where one image point is matched to more than one image point in the other image is not allowed. This assumes opaque surfaces.
Ordering constraint says that the order of image points on the epipolar line (row of the rectified image) is the same for both cameras, i.e. a set of correspondences violating this order is not allowed. Note that continuous surface that is visible in both cameras does not violate ordering.

$$\mathrm{NCC}(\mathbf{W}_i, \mathbf{W}_j) = \frac{\mathrm{cov}(\mathbf{W}_i, \mathbf{W}_j)}{\sqrt{\mathrm{var}\,\mathbf{W}_i \,\mathrm{var}\,\mathbf{W}_j}}, \tag{2.12}$$

$$\mathrm{MNCC}(\mathbf{W}_i, \mathbf{W}_j) = \frac{2\,\mathrm{cov}(\mathbf{W}_i, \mathbf{W}_j)}{\mathrm{var}\,\mathbf{W}_i + \mathrm{var}\,\mathbf{W}_j}, \tag{2.13}$$

where the $\mathbf{W}_i$ and $\mathbf{W}_j$ are n-dimensional feature vectors of image point i and j respectively. Zabih and Woodfill [178] proposed a similarity statistics which is based on rank transform and census transform of the intensities in the window. Such transforms make the resulting statistics insensitive to certain photometric perturbations and also more robust to occlusions, as reported. A review and benchmark of similarity statistics used for stereo can be found in a recent paper by Hirschmüller and Scharstein [64].

Finally, the matching algorithm establishes the matches based on the correlation table according to a certain *principle*. This classic approach is described in Sec. 3.1.

The simplest algorithm of this family is the Winner Takes All (WTA), see for example [82]. It selects the match with the highest similarity for each row (or column) of the matching table, without any other constraints. Since the uniqueness constraint is not generally preserved, the algorithm does not produce one-to-one matching. This algorithm may be used as a reference method only or in real-time applications, where the speed is highly important. Mulligan et al. [108] presented their real-time stereo algorithm using trinocular reconstruction from five cameras. Jia et al. [66] constructed special hardware using FPGA, which can produce 640×480 disparity map in 30 fps.

People noticed soon, the result of WTA is not perfect and posed other constraints on the disparity map. The famous one was the algorithm of Pollard, Mayhew and Frisby using the disparity gradient limit [123]. Disparity gradient limit is a constraint that generalizes the ordering constraint. The authors attempted to improve results by not allowing steep disparity changes. Although they were considered successful at their time, today they are overtaken by MAP methods with isotropic prior.

Šára [130] revised the optimality condition and proposed a *stability* as a criterion for finding solution. The stability principle says that the solution should not change much with a small perturbation of data. Šára designed a Confidently Stable Matching (CSM) algorithm [129] which finds the largest unambiguous matching, according to preselected confidence level. The algorithm constructs an oriented graph on the matching table cells in such a way that edges are oriented from the higher to lower similarity values within the forbidden zone generated by uniqueness and ordering constraint. Intervals around the similarity values derived from the confidence level are compared instead of the similarity values themselves. The task leads to finding the kernel of such graph. The principle of the algorithm is also described in Chapter 5.

Methods using a rectangular window to compute the similarity statistic suffer from artifacts when the disparity is not constant within the window, especially near occlusion boundaries. This effect causes a "fattening" of thin objects and generally decreases the discriminability of the correlation statistic within the matching table. An analysis of the cases of this problem was studied by Šára and Bajcsy [131]. There are several ideas to

cope with this problem. Bhat and Nayar [9] proposed a robust rank correlation. Noguchi and Ohta [112] propose to limit the value when computing the sum of absolute differences, but the remaining problem is a dynamic estimation of this limit. Hirschmüller [62] proposes to divide the correlation window into sub-windows and compute the correlations separately. The final correlation is computed from the central sub-window plus N best correlations. Okutomi and Katayma [116] use offset-windows, where the pixel of interest is not in the window center.

Kostková and Šára [83] proposed a method, where the matching window is adapted according to disparity-components in the disparity space re-projected into images. The discriminability enhancement is significant.

There are algorithms which do not model the neighbourhood of a pixel by taking the surrounding pixel intensities, but describe it with other feature vectors, based on coefficients of various filter responses. This is the case of Jones and Malik [67] and all phase-based techniques, which are described in Sec. 2.2. Clerc [30] uses Gabor wavelets for a local image description and MNCC of these features. She models the first order disparity, i.e. estimating both constant disparity and its slope. The experiments demonstrate a high discriminability even for repetitive-patterned texture, where standard correlation fails.

2.1.3 Heuristic methods

There are several algorithms which try to find the matching with an iterative process, usually based on some heuristics. These algorithms usually attempt to incorporate a prior model into the class B methods.

Zhang and Kambhamettu [182] assume the disparity varies smoothly within a homogeneous intensity image segment. Initially, the algorithm establishes certain matches and classifies segments as valid, semivalid or invalid (occluded) according to the ratio of the certain matches within. Then the algorithm iteratively propagates the reliability information. Szeliski and Scharstein [150] perform the matching on per-scanline interpolated images. Initially, a few matches with high confidence values are established. Then the algorithm propagates confidence labels into so far ambiguous regions aggregating the matching table with successively larger windows. Similar progressive matching strategy where the more confident correspondences are matched before the less confident, in order to reduce the ambiguous space, can be found for instance in papers by Zhang and Shan [183] or Wei and Quan [168].

2.2 Phase-based techniques

By the phase, it is meant the Fourier phase of a signal or the angle of a response to a complex-valued (quadrature) filter in general, see e.g. Granlund and Knutsson [55]. The phase is supposed to be a discriminable and stable feature under various global illumination changes and geometric deformation. Phase-based approach allows a simultaneous

sub-pixel estimation of the disparity, exploiting the Fourier shifting theorem: a shift in the spatial domain corresponds to a shift of the phase in the frequency domain.

Weng [169] introduced the windowed Fourier phase as a matching primitive. The matching algorithm successively refines the disparity in the coarse-to-fine framework, using a hierarchical image pyramid. The approach is fully 2D, epipolar geometry needs not be known.

Sanger [127] used Gabor filter responses to compute the disparity. The final disparity is calculated as a weighted average of the disparities obtained from various filter bands. The weights are derived from the similarity of absolute value of the responses.

Fleet and Jepson [44, 43] also used Gabor filters. They introduced a stability criterion of the phase, which is based on the discrepancy of the tuning frequency of the Gabor filter and the local frequency of the response. Unstable estimates are not used in the subsequent iterative algorithm.

Xiong and Shaffer [175, 174, 176] derived a stability criterion similar to [44] defining more conditions for the phase stability. They proposed to use the moment and hypergeometric filters beside the Gabors. These filters are orthogonal, complete and have special recursive properties in both spatial and frequency domain. They use high-order Taylor expansion of the response which allows affine invariant estimation, but it requires solving a set of equations by an iterative procedure for each match.

There are other papers, which present matching algorithms combining phase-based techniques with other well-established methods. Fröhlinghaus and Buhmann [46] proposed a regularization with smoothness in a Bayesian framework. Similarly, Pan and Magarey [119] build a hierarchical image pyramid and run a stochastic relaxation at each level optimizing the function which includes occlusion and smoothness term.

2.3 Other methods

There exist several other techniques in the literature which are based on a completely different approaches than methods mentioned above. We name a few of them.

Tomasi and Manduchi [155] designed an algorithm which does not search in the space of disparities. They construct an intrinsic curve for each scanline using features like local intensities and their derivatives. This is a translation (disparity) invariant representation. The matching is then transformed into nearest neighbour problem in the space of intrinsic curve features.

The *space carving* methods solve the correspondence problem in a reversed way. The main principle is the following: A large set of fully calibrated cameras is assumed. The observed scene is divided into voxels. The scene is reconstructed such that rays from each voxel to all cameras are projected to the images. Then, the voxel is set to be either empty or filled according to the variance of its projection into images. More details can be found for example in papers by Kutulakos and Seitz [88, 89].

The *plane sweeping* methods are applicable under strong assumptions on the planarity or piecewise planarity of the scene. These algorithms sweep a family of planes through

the volume generating correspondence hypotheses evaluated by (dis)similarity statistics between images transformed by a homography induced by the sweeping plane. This approach is employed recently in a real-time multi-view stereo system by Gallup et al. [49] for urban environment.

Another approach is the model based stereo. The strong prior model is fitted into stereo measurements. For instance, Amberg et al. [3] presented recently a successful algorithm of this kind for human face reconstruction, which is otherwise extremely difficult by other methods due to lack of texture and non-Lambertianity. They fit a 3D morphable model which is a space of 3D faces spanned by a linear combination of basis-faces.

2.4 Polynocular stereo

Although this thesis deals with a binocular (two-view) stereo, it is important to mention methods which use more views for stereo matching. These methods are called polynocular (multi-view) stereo.

It has been shown by many authors that incorporating more views can significantly improve the matching. The ambiguity due to the repetitive pattern in images is reduced and the noise in the image is filtered out. However, there are several other problems which complicate the task. First of all, a highly precise calibration of cameras is required, otherwise the effect is spoiled. This has been recently noticed by Furukawa and Ponce [48], where they suggest to refine the calibration using bundle adjustment from dense point cloud. Other source of difficulties are for example: It is impossible to directly rectify images for more than three views, there are more occlusion types since only a subset of cameras may see a scene point, etc. The important question for the polynocular stereo is the internal representation of the scene, due to large amount of data. The representation can be volumetric (voxels), using triangular mesh, or image based (depth maps).

Okutomi and Kanade [115] assume that the images are aligned in a sequence such that epipolar lines coincides with the same rows for all images. They use one reference view in a way that SSDs between this reference image and other images are summed, they call it SSSD. Minima are then found there. Asymmetric matching using a reference image is also presented by Kang et al. [70]. The global energy is minimized with the graph cuts algorithm, initialized by Winner Takes All. Mulligan and Daniilidis [107] designed a trinocular algorithm, where images from cameras in general non-parallel positions are pairwise binocularly rectified (center,left and center,right). They construct tentative disparity maps for both image pairs via selecting N highest correlations. The final disparity map is then determined as the maximum of the sum of corresponding correlations among N hypotheses. Fully symmetric trinocular matching algorithm is presented by Buehler et al. [20], where the L-shaped rectification is used. The global energy function is minimized via finding the minimal surface as a generalization to the minimal path in a planar graph. In our paper [181], we propose a simple L-shaped trinocular rectification and confirm both qualitatively and quantitatively the dramatic

improvement when a symmetric trinocular matching algorithm is applied compared to binocular or a join of all binocular matchings. The correlation table was calculated as a mean MNCC (2.13) from all three pairs and the CSM algorithm [129] was used for matching.

Multi-view stereo has been intensively studied recently. Let us name a few works, which also give an extensive references to literature, which an interested reader can follow. Strecha [146] employs PDE-based technique. A mechanism where a surface patch is grown is used by Furukawa and Ponce [47] or by Habbecke and Kobbelt [58]. Vogiatzis et al. [166] uses a volumetric formulation and the solution (optimum partitioning of the 3D space) is found by a minimum graph cut. Labatut et al. [90] propose an optimization of a tetrahedral network obtained by a Delaunay triangulation from an initial 3D point cloud.

2.5 Performance testing and evaluating stereo algorithms

Naturally, it arose a necessity to test and evaluate quantitatively the performance of various stereo algorithms. Several attempts have been made.

There exist two main approaches of the evaluation: (1) image prediction based methods, where matches are verified according to their self-consistency, Leclerc et al. [91] and (2) ground-truth based methods, Bolles et al. [15], Scharstein et al. [136] or our papers [82, 84]. The Middlebury evaluation site of Scharstein et al. [136] has become extremely popular among researchers in stereo. However, the drawback is that ground-truth disparity maps are given to researchers which causes that certain algorithms overfit to this dataset, which is not very representative to real-world problems.

In [82, 81] we presented a methodology for evaluating stereo algorithm performance and defined several error statistics, besides specifying weaknesses of the approach [136]. This study is focused on comparison of the algorithms having a strong prior model (class A) and the algorithms which are based on sufficient discriminability of correspondences (class B). Another our paper [84] presents an ROC curve-based evaluation of algorithms. It is intended for algorithms which do not output fully dense disparity maps. The ROC-like curve is computed when spanning the space of a tested algorithm parameters. Hereby, the method evaluates the algorithm as a whole, not only its particular parameter setting. An explicit performance comparison among various algorithms is thus possible.

Evaluation methods also exist for multi-view stereo algorithms. The methods provide the images with camera calibration. Certain statistics based on the discrepancy between the supplied and ground-truth triangulated mesh are measured. The popular is again the Middlebury site, described by Seitz et al. [139] which contains two small laboratory test objects. Recent benchmark method of Strecha et al. [148] brings several large scale outdoor scene with high-resolution images. In both cases, the ground-truth is not provided to users, the evaluation is kept on-line at the respective web sites.

3 Program of the thesis

After reviewing the state-of-the-art methods and getting experience from the algorithm evaluation, we decided to study and develop methods of class B.

As we have seen, methods of class A using a strong prior model suffer from several types of artifacts. They tend to smooth out occlusion boundaries and they are prone to produce illusory results in the regions, where

1. data is insufficient to establish correct correspondences

   (consider the sky in Fig. 2.2(a),(b), Fig. 2.3(a),(b))

2. data contradicts prior model

   (consider continuity prior at occlusion boundaries in Fig. 2.2(a),(b), Fig. 2.3(a),(b))

3. the correct explanation of the images is no correspondence

   (consider the gap between the leftmost trees in Fig. 2.3(a), 2.3(b) in which we see completely different part of the background).

The reason for the artifacts is probably the inability of the smoothness prior term to capture well the natural statistics of real complex scenes. This is true even in cases where the occlusions are explicitly modeled. This is probably partly due to difficulty of constructing or learning such a complicated prior, and partly due to algorithmic motivation. The class of functionals which are reasonably solvable by graph cuts [78] or other discrete optimization technique is limited.

The other reason the most of class A methods are not much suitable for accurate reconstruction of complex 3D scenes using high-resolution images with large range of disparities is their computational inefficiency. This holds true for discrete methods (A.1) which work with isotropic 2D prior explicitly. The discrete optimization takes a long time beside large memory requirements.

Furthermore, our conception is that the matching algorithm should have the 'reject' option. It should not assign any disparity in the regions, where the data are not reliable enough, but it should identify such regions and assign the disparity for unambiguous regions only. This is demonstrated in Fig. 2.2(c), 2.3(c), 2.2(d) and 2.3(d). If needed, higher model-based (or even cognitive) process may then be used to interpret the data depending on the application.

3.1 A standard approach

A classical approach to binocular stereo is a two stage process. First, a quality of correspondence hypotheses is evaluated. This is the measurement stage. Then, the matching algorithm selects the final matching pixels from the set of evaluated correspondence hypotheses.

A space of all correspondence hypotheses is called a *disparity space*. Its construction is sketched in Fig. 3.1. A pair of rectified images $\mathbf{I_l}(x_l, y)$, $\mathbf{I_r}(x_r, y)$ is assumed, where the horizontal coordinates (image columns) are $x_l \in \mathcal{X}_l$, $x_r \in \mathcal{X}_r$ and the vertical coordinates (image rows) are $y \in \mathcal{Y}$. The disparity space is defined as a Cartesian product $\mathcal{D} = \mathcal{X}_l \times \mathcal{X}_r \times \mathcal{Y}$. Each of its discrete cells represents a correspondence hypothesis. The quality of the correspondence hypothesis is evaluated by a *similarity (correlation) statistic* between *image signatures* of examined pixels. The image signature is a vector representing a local image neighbourhood of a pixel. A cross-section of the disparity space along the last dimension (an image row is fixed) is called a *correlation table*.

The most trivial image signature is the pixel intensity, and the similarity statistic is then their square or absolute difference. However, the *discriminability* of the trivial similarity statistic is very low. The discriminability is a property of the similarity statistic which assigns high values to true matches while keeping low values to the others. A formal definition of discriminability is given in (4.57). Therefore to increase the discriminability, a larger neighbourhood is considered and/or more sophisticated similarity statistics are constructed. Unfortunately, larger neighbourhood causes other artifacts, especially near occlusion boundaries as discussed in the previous chapter.

The correlation table for the highlighted row in Fig. 3.1(a) is shown in Fig. 3.1(c). Here, the image signatures are the raw intensities in a window of 5×5 pixels, see Fig. 3.1(a), and MNCC (2.13) is used as a similarity statistic. The correlation table is encoded in color, such that higher similarity pairs are displayed in warmer colors, lower similarity pairs in colder colors. Notice the red segments mostly belong to correct correspondences (from left to right): to the tree, to the top part of the monument, to the right small pinnacle, and to the image border.

The task of the matching algorithm is to take the evaluated disparity space as the input and to establish the matches in order to either optimize a certain criterion (class A methods) or according to a certain principle (class B methods). The matching algorithms which do not use a 2D prior explicitly, work per scan-lines usually. This is possible in cases, where the matching problem is decomposable into row independent problems. These algorithms per each row compute a correlation table and perform the matching straight away without storing the entire disparity space in a memory.

3.2 Where are the problems?

The standard approach described above suffers from several problems which are dealt with in this thesis:

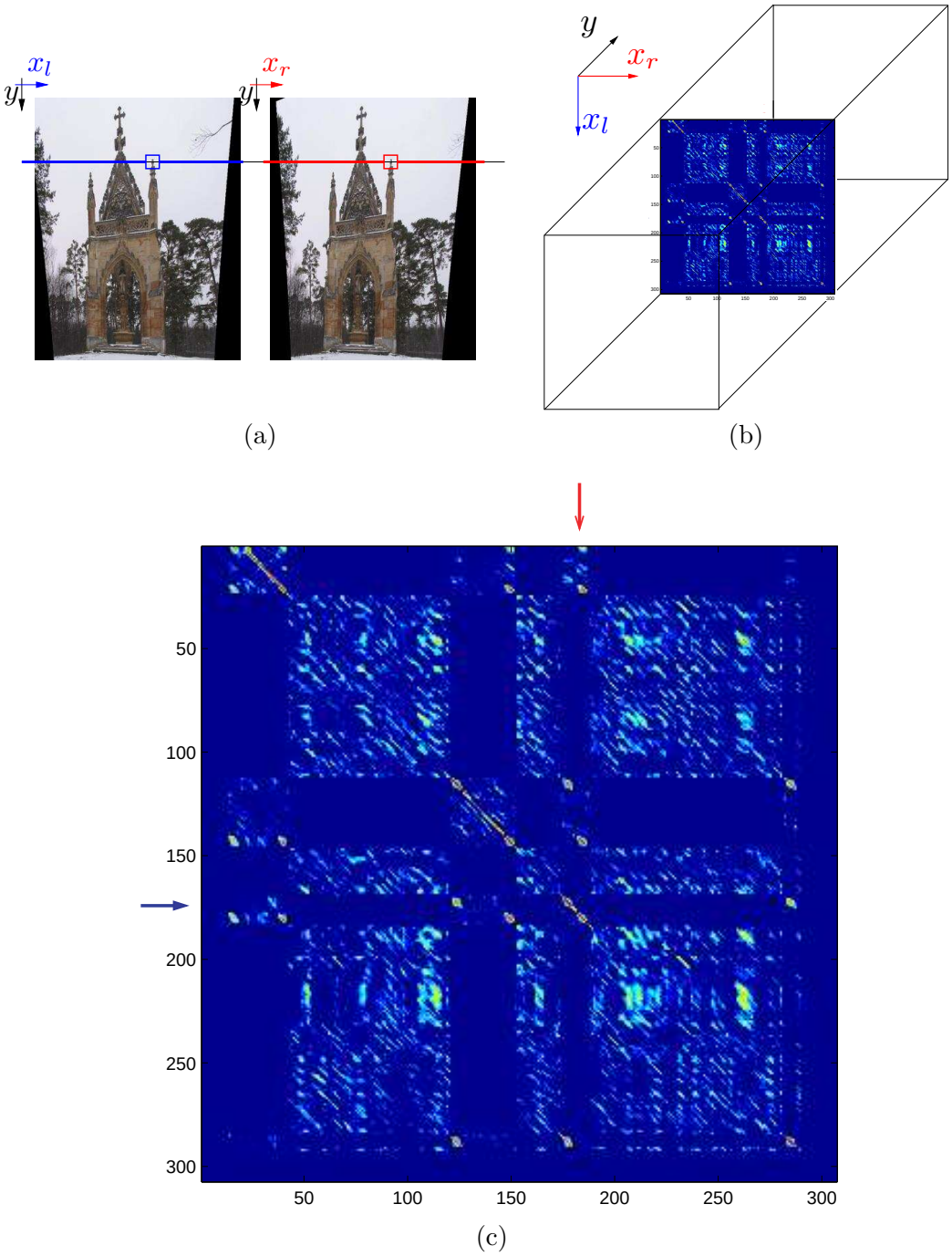


Figure 3.1: A construction of the disparity space: A pair of rectified images (a), visualization of the disparity space (b), correlation table (zoom) for the highlighted epipolar line (c). Correlation values are encoded in color; higher correlations in warmer colors, lower correlations in colder colors.

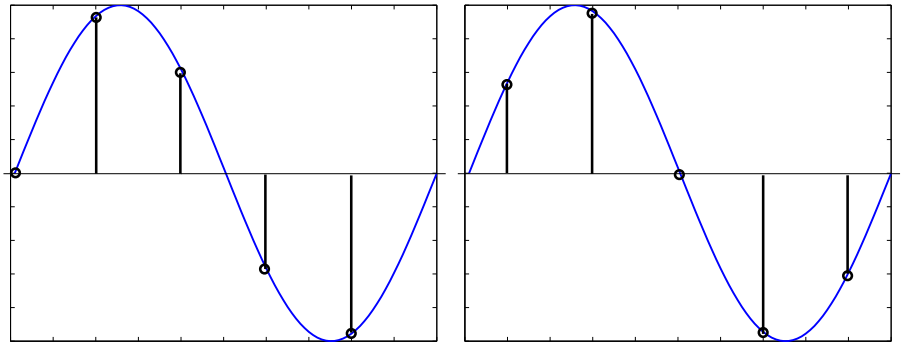


Figure 3.2: Discretization artifacts: Sampling of the right image is half a pixel shifted. Notice, the values differ significantly.

- **Discretization artifacts**

  A sampling of the images causes a corruption of standard correlation statistics. The motivating one dimensional example is shown in Fig. 3.2. There are two sine waves, where the sampling of the right signal is half a pixel shifted with respect to the left signal. The values differ significantly, although we are much below the Nyquist frequency. The standard MNCC statistic (2.13) is 0.8 only. Increasing the frequency of the signal further intensifies the discretization artifacts, which manifests as decreasing the MNCC statistic. For the Nyquist frequency, MNCC drops to zero, since the signals become orthogonal.

- **No invariance to affine distortion**

  Using a standard similarity statistic, e.g. (2.10)–(2.13), computed from square windows assumes the fronto-parallel planar setup. It means, the cameras are observing a plane which is parallel to their image planes. However, the violation of this setup which occurs by rotating the observed plane, causes an affine distortion of local image neighbourhoods which again corrupts the correlation statistic. The geometric distortion of the square matching window is shown in Fig. 4.3.

- **Too large number of similarity statistics to be computed**

  A standard approach is to compute the correlation statistics for the entire disparity space. However, the disparity space is large for high-resolution images. Consider a pair of square 1 Mpx images, 1000×1000 pixels. Then one needs to compute all $1000 \times 1000 \times 1000 = 1\text{billion}$ of correlation values, which takes time, despite the existence of an efficient implementation of computing the correlation tables of basic statistics using 'a sliding window', e.g. [24]. Limiting the disparity search-range, if it is known a priori, does not solve the problem. It reduces computational time, nevertheless the space still remains too large for complex scenes of a great depth.

3.3 Contributions of the thesis

A program of the thesis was initiated by the problems identified in a standard approach and described above. The program resulted in several contributions:

1. **Sampling invariant correlation statistic**

   We designed the Complex Correlation Statistic, which is both invariant to image discretization and it also provides an estimate of the sub-pixel translation between the local image neighbourhoods. The value is not a real but a complex number. The magnitude is the invariant similarity, while the phase is the estimate of the sub-pixel shift. The main idea is that the pixel signatures are represented by responses to a bank of complex Gabor filters. The statistic is then computed from the responses in a closed form.

   This work is presented in Chapter 4 and has been published in [25].

2. **Affine insensitive correlation statistic**

   In a second step, we incorporated insensitivity to affine distortion of a local image neighbourhood, which occurs due to a surface slant, into the Complex Correlation Statistic. The basic idea is to decompose the distortion into orthogonal aspects (slant and skew) and design a filter bank which is able to compensate an affine distortion in a limited range.

   This is presented in Sec. 4.3 of Chapter 4.

3. **Global method for sub-pixel disparity correction**

   We proposed a method of a sub-pixel disparity correction which is formulated as a single optimization problem for the entire image. It is a continuous optimization problem based on a simple quadratic criterion. The algorithm finds a sub-pixel disparity map directly by a gradient descent from a given initialization by an integer disparity map.

   Originally the method should have been a reference method to compare the accuracy with methods based on Complex Correlation Statistic. However, despite its simplicity it turned out to give good results. The method is presented in Sec. 4.4 of Chapter 4.

4. **Fast stereo matching algorithm which involves computation of a small fraction of disparity space**

   It turns out, it is not necessary to compute the correlations exhaustively for the entire disparity space. We propose an algorithm which visits only a small fraction of the disparity space (less than 1 per cent) while it still keeps a good performance, in the sense of robustness to occlusions, weak textures and repetitive patterns, as an exhaustive algorithm. Hereby, the speed up achieved is in the order of two magnitudes. The main idea is to generate promising correspondence hypotheses

by growing initial correspondences (seeds). Finally, a robust matching algorithm is applied to select the stable matching [129] among the competing correspondence hypotheses. The algorithm is not dependent on given high-quality seeds, but surprisingly it works with seeds which are generated completely randomly.

The algorithm is described in Chapter 5. This work has been published in [26].

5. **Verification of tentative correspondences based on fast dense matching algorithm**

   Being inspired by the success of the dense matching algorithm based on growing correspondence seeds, we apply the growing mechanism to construct a procedure which grows a given tentative correspondence, collecting primitive statistics, in order to estimate how likely it is a true correspondence or a mismatch. Such a verification is typically used before fitting a model using RANSAC. It has a large impact in difficult matching problems where the ratio of outliers is above 90 per cent, where matching of standard compact descriptors fails, usually due to a complex 3D structure with many occlusions. The procedure is driven by Wald's sequential decision which makes the verification process efficient. The time spent by the verification is negligible with respect to the time spent by matching of tentative correspondences.

   The method is explained in Chapter 6 and originally published in [23].

3.4 How to read the thesis

The rest of the thesis is organized as follows: Chapter 4 presents Contributions 1––3, Chapter 5 Contribution 4, and Chapter 6 Contribution 5. In the last Chapter 7 discussions and conclusions are given.

The technical Chapters 4–6 are placed in the thesis in the same order in which the work emerged. However, particular chapters are written in a way they are not dependent on each other in order to be maximally stand-alone. Thus the order of reading them does not matter.

4 Complex Correlation Statistic and sub-pixel disparity

A traditional solution of area-based stereo uses some kind of windowed pixel intensity correlation. This approach suffers from discretization artifacts which corrupt the correlation value. We introduce a new correlation statistic, which is completely invariant to image sampling, moreover it naturally provides a position of the correlation maximum between pixels. Additionally, we present a version which is insensitive to affine distortion of corresponding neighbourhoods which occurs due to surface slant. Hereby we can obtain sub-pixel disparity directly from invariant and highly discriminable measurements without any post-processing of the discrete disparity map.

The key idea behind is to represent the image point neighbourhood as a response to a bank of Gabor filters. The images are convolved with the filter bank and the complex correlation statistic (CCS) is evaluated from the responses without iterations. The magnitude of CCS measures the image similarity and the phase gives the sub-pixel position.

In this chapter, we also present a simple global method for sub-pixel disparity correction which is posed as a single optimization task and solved iteratively, a complementary approach to sub-pixel disparities from CCS. The criterion consists of a quadratic data term penalizing discrepancies between the target image and the reference image warped according to estimated disparity map, and a smoothness term which penalizes differences in neighbouring disparities also quadratically.

4.1 Introduction

In stereo, we have to recognize corresponding points, i.e. the image points which are the projection of the same spatial point, according to how much the image point neighbourhoods are similar, computing some kind of image correlation statistic.

As discussed, a stereo algorithm usually consists of two essential modules: the measurement evaluation and the matching. The first process computes some kind of similarity (correlation) statistics between all potential correspondences. The second process takes this measurements and establishes matches according to some principle or optimizing some criterion. Both stages are important and dramatically influence the matching results, which could be seen in stereo evaluation works [136, 82, 84].

This chapter is exclusively devoted to the similarity computation stage introducing

a new correlation statistic which can be used by various matching algorithms. We introduce a Complex Correlation Statistic (CCS) which is invariant to image sampling and allows sub-pixel match localization. We also introduce a CCS version insensitive to affine distortion of corresponding neighbourhoods which occurs due to slanted surfaces. As a complementary method for sub-pixel disparity estimate, we present the global method which formulates the problem as a single optimization task.

Birchfield and Tomasi [10] noticed that it is important the correlation statistic be invariant to image discretization and proposed a sampling-invariant pixel dissimilarity. It is a simple extension of Sum of Absolute Differences (SAD) based on a linear interpolation. It works quite well but it tends to fail where there are very high frequencies in the images. The aggregation of this pixel dissimilarity over a window has become popular in area-based stereo, e.g. [164].

Psarakis and Evangelidis [125] proposed an extended normalized correlation statistic invariant to image sampling using locally linearly interpolated image intensities. Their statistic also offer an estimate of sub-pixel displacement.

Szeliski and Scharstein [151] recommended several other matching scores based on interpolated image signals, e.g. to interpolate the images to a higher resolution, compute the usual statistics, aggregate and subsample to the original resolution. The drawback of this approach is that the increase in resolution is finite, which limits the discretization invariance property.

There are several possibilities to achieve sub-pixel matching. Again, the simplest way is to work with interpolated high resolution images. We could have the sub-pixel precision up to the level of interpolation and also the statistic will be less sensitive to image sampling. But, computational expenses increase dramatically.

A possible solution might be to interpolate in the space of correlation statistic (disparity space), like fitting a parabola between three values in the correlation table to find where the extreme of the statistic is. These methods were studied by Shimizu and Okutomi [140, 141, 142], where they formulate which interpolant is suitable for which statistics, and proposed a method to compensate a systematic error using that method. Nehab et al. [109] reported an improvement when fitting a symmetric model in the correlation table to find the maximum.

In theory, a sub-pixel disparity map could be also obtained from any probabilistic labeling framework, e.g. [149, 153], where the final disparity is determined as a mean value from all integer disparity labels. However, it has not been analyzed properly whether such an estimate is accurate.

Sub-pixel matching precision can also be achieved using so called phase-based techniques. These techniques have been described in Sec. 2.2. They exploit the Fourier shifting theorem which says that a shift in the spatial domain corresponds to a shift of the phase in the frequency domain.

The dense stereo matching under affine distortion is being solved by Birchfield and Tomasi in [11] too. They propose an EM-like approach where the correspondence segmentation is alternated with affine parameters estimation. Ogale and Aloimonos [114] stressed the negative influence of a horizontal slant and proposed a search scheme for

the slant factor beside the disparity.

Our approach is very close to phase-based techniques in the sense we model the image using Gabor filters and exploit the shifting theorem. But we do not estimate the disparity map directly, we embed the estimation of the sub-pixel disparity in the correlation statistic which is evaluated in a closed form without any complicated optimizations. The affine insensitive version of CCS is also computed from responses without any iterative scheme.

The chapter is organized as follows. We introduce our Complex Correlation Statistics, originally published in [25], in Sec. 4.2, its affine insensitive extension is presented in Sec. 4.3. The global method for sub-pixel disparity estimation is described in 4.4. The experiments are given in Sec. 4.5. Finally, Sec. 4.6 concludes the chapter.

4.2 Complex Correlation Statistic

4.2.1 Definition

Let us have rectified images $\mathbf{I_l}(x, y)$ and $\mathbf{I_r}(x, y)$, in which epipolar lines coincide with image rows y in both images. Image similarities for all potential pixel correspondences $(x_1, y) \in \mathbf{I_l}, (x_2, y) \in \mathbf{I_r}$, for current scanline y form so called correlation table $\mathbf{c}(x_1, x_2)$.[1] We define the Complex Correlation Statistic to be a complex number

$$\mathrm{CCS}(x_1, x_2) = Ae^{j\varphi}, \tag{4.1}$$

where the magnitude A is the similarity value which is invariant to image sampling, and the phase φ shows the correct position of the match between pixels, see Fig. 4.1. The thick black line represents the truth matching, i.e. an image of a surface in the disparity space. Green circles mark cells of the correlation table at locations, where magnitude A should be the highest (ideally 1). Blue arrows represent the angle φ pointing towards the correct match position in the horizontal direction. Red arrows are pointing towards the correct match position in the vertical direction. This is the angle $\tilde{\varphi}$ of the complementary correlation $\tilde{\mathrm{CCS}}(x_1, x_2) = Ae^{j\tilde{\varphi}}$. Magnitudes of CCS and $\tilde{\mathrm{CCS}}$ are the same, the only difference is in phases. These statistics are evaluated in each cell of the correlation table. A computational block of this procedure is sketched in Fig. 4.2. The inputs are intensity values of the neighbourhood $\mathcal{N}$ of the left $\mathbf{I_l}$ and right $\mathbf{I_r}$ image at position x_1, x_2, respectively on the yth row. Swapping the inputs f, g of this block causes swapping the CCS to $\tilde{\mathrm{CCS}}$.

In the following sections, we will describe what is inside the CCS-block. First, we give the formula and then we explain it.

[1] For the case of a standard windowed statistic, e.g. SAD, it contains the sum of absolute differences of pixel intensities in windows centered on a common row y at columns x_1 in the left image and x_2 in the right image.

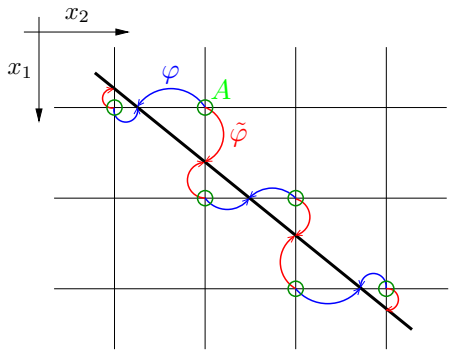


Figure 4.1: Correlation table of the CCS.

Figure 4.2: Computational block.

4.2.2 Procedure for computing the CCS

Both images are convolved with a bank c_i of several Gabor filters tuned to different frequencies to equally sample the frequency spectra, and with the corresponding x-partial derivative filter bank c_{xi}. The CCS is computed from the responses in a closed form. The bank of filters is

$$\begin{aligned} c_i(x,y) &= \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} e^{j(u_{0i}x+v_{0i}y)}, \\ c_{xi}(x,y) &= \frac{\partial c_i(x,y)}{\partial x} = \left(-\frac{x}{\sigma^2} + ju_{0i}\right) c_i(x,y), \end{aligned} \tag{4.2}$$

where $i = 1, \ldots, N$ is the index of a filter tuned to a frequency (u_{0i}, v_{0i}) with a constant scale σ.

The input images $f(x,y)$, $g(x,y)$ are convolved with the filters, so the responses are

$$G_{f_i}(x,y) = f * c_i, \quad G_{g_i}(x,y) = g * c_i, \quad G_{xf_i}(x,y) = f * c_{xi}. \tag{4.3}$$

We estimate the local frequency, i.e. a partial derivative of the phase of the Gabor response

$$u_{f_i}(x,y) = \frac{\Im(G_{xf_i})\Re(G_{f_i}) - \Re(G_{xf_i})\Im(G_{f_i})}{|G_{f_i}|^2}. \tag{4.4}$$

Now, the CCS will be evaluated per scanline y. First, for all the cells of the correlation table and for all the filters $i = 1, \ldots, N$, we estimate the local subpixel disparity

$$\Delta_i(x_1, x_2) = \frac{\arg(\overline{G_{f_i}(x_1,y)} G_{g_i}(x_2,y))}{u_{f_i}(x_1,y)}. \tag{4.5}$$

We denote $a_{f_i}(x,y) = |G_{f_i}|$, $a_{g_i}(x,y) = |G_{g_i}|$ and finally, we aggregate the data using the formula:

$$\mathrm{CCS}(x_1, x_2) = \frac{2\sum_{i=1}^{N} a_{f_i}(x_1,y) a_{g_i}(x_2,y) e^{j\Delta_i(x_1,x_2)}}{\sum_{i=1}^{N} a_{f_i}(x_1,y)^2 + \sum_{i=1}^{N} a_{g_i}(x_2,y)^2}. \tag{4.6}$$

In this subsection we have described completely the procedure for computing the correlation table of the Complex Correlation Statistic. In the next section we will explain it more in detail and show why it works.

4.2.3 Derivation of the formula for the CCS

Locally, we have two signals $f(x, y)$ and $g(x, y)$ related by

$$g(x, y) = f(x + d(x, y), y), \tag{4.7}$$

where the (local) disparity is assumed to be small and linearly varying with x, y

$$d(x, y) = d_0 + d_1 x + d_2 y. \tag{4.8}$$

First, we will show the $d(x, y)$ can be estimated from the responses of the Gabor filters. Let us assume that our signals are real signals consisting of a single frequency (u, v) only

$$\begin{aligned} f(x, y) &= a\cos(ux + vy + \varphi), \\ g(x, y) &= a\cos(u(x + d_0 + d_1 x + d_2 y) + vy + \varphi). \end{aligned} \tag{4.9}$$

The cosine function can be rewritten as

$$f(x, y) = a\cos(ux + vy + \varphi) = \frac{1}{2}a(e^{j(ux+vy+\varphi)} + e^{-j(ux+vy+\varphi)}). \tag{4.10}$$

The response of the Gabor filter tuned to the frequency (u_0, v_0) is a convolution

$$\begin{aligned} G_f(x, y) = f(x, y) * c(x, y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c(s, t) f(x - s, y - t) ds dt = \\ &= \frac{1}{2}a\Big(e^{-\sigma^2\left((u_0-u)^2+(v_0-v)^2\right)} e^{j(ux+vy+\varphi)} + e^{-\sigma^2\left((u_0+u)^2+(v_0+v)^2\right)} e^{-j(ux+vy+\varphi)}\Big). \end{aligned} \tag{4.11}$$

Response magnitude is a sum of two Gaussians centered at signal frequencies (u, v) and $(-u, -v)$. Under the assumption $u_0 \approx u$, $v_0 \approx v$ a part corresponding to the peak of symmetric (negative) frequencies can be neglected. This is always possible except for the tuning frequencies close to zero. Therefore, the responses of both image signals are

$$\begin{aligned} G_f(x, y) = f * c &\approx \tfrac{1}{2}a e^{-\sigma^2\left((u_0-u)^2+(v_0-v)^2\right)} e^{j(ux+vy+\varphi)}, \\ G_g(x, y) = g * c &\approx \tfrac{1}{2}a e^{-\sigma^2\left((u_0-(u+d_1))^2+(v_0-v)^2+2d_2(v-v0)\right)} e^{j(u(d_0+d_1x+d_2y)+ux+vy+\varphi)}. \end{aligned} \tag{4.12}$$

The disparity is determined from the argument of the Gabor responses:

$$\begin{aligned} \arg(\overline{G_f} G_g) &= -(ux + vy + \varphi) + \big(u(d_0 + d_1 x + d_2 y) + ux + vy + \varphi\big) = \\ &= u(d_0 + d_1 x + d_2 y) = u\, d(x, y), \end{aligned} \tag{4.13}$$

which is the formula (4.5), because $u = u_f$ is the local frequency, i.e. the derivative of the response phase, as can be seen by taking partial derivatives of the arguments of (4.12):

$$\begin{aligned} u_f &= \frac{\partial \arg(G_f)}{\partial x} = u, & v_f &= \frac{\partial \arg(G_f)}{\partial y} = v, \\ u_g &= \frac{\partial \arg(G_g)}{\partial x} = u + d_1 u, & v_g &= \frac{\partial \arg(G_g)}{\partial x} = v + d_2 u. \end{aligned} \tag{4.14}$$

The disparity gradient (d_1, d_2) can be estimated from the local frequencies as $d_1 = (u_g - u_f)/u_f$, $d_2 = (v_g - v_f)/u_f$.[2] If we use u_g instead of u_f in the denominators here and in (4.5), we get the complementary disparity $\tilde{d}$, discussed in Sec. 4.2.1.

Numerically, we can estimate the local frequency, i.e. the partial derivative of the phase of the response, from the Gabor and Gabor partial derivatives filters. We denote $R = \Re(G_f)$, $I = \Im(G_f)$, $R_x = \Re(G_{xf})$, $I_x = \Im(G_{xf})$. Then

$$u_f = \frac{\partial \arg(G_f)}{\partial x} = \frac{\partial}{\partial x} \arctan\Big(\frac{I}{R}\Big) = \frac{1}{R^2 + I^2}\Big(\frac{\partial I}{\partial x} R - \frac{\partial R}{\partial x} I\Big) = \frac{I_x R - R_x I}{R^2 + I^2}, \tag{4.15}$$

which is the formula (4.4). Formulas for the other local frequencies can be derived analogously.

Computation of the CCS according to (4.6) is inspired by the Moravec's normalized cross-correlation [105], which works well for windowed data. We use a similar formula regardless of the fact we work with complex Gabor responses. It is not exactly the same, because there is the necessary frequency normalization (4.5). After the normalization, the meaning is as follows: Under the correspondence, the local disparity estimates Δ_i from all the filters should agree and the magnitudes a_{f_i} and a_{g_i} for each filter should also be the same. The formula (4.6) measures how much this is true.

For example, let us have two images where one is shifted from the other with a constant disparity δ which is smaller than one pixel, $f(x, y)$ and $g(x, y) = f(x + \delta, y)$. No noise is assumed. Then all the local disparities in (4.5) are $\Delta_i = \delta$ and from (4.12), we can see that $a_{f_i} = a_{g_i} = a_i$. Substituting this into (4.6) we obtain

$$\mathrm{CCS}(f(x), f(x + \delta)) = \frac{2 \sum_{i=1}^{N} a_i a_i e^{j\delta}}{\sum_{i=1}^{N} a_i^2 + \sum_{i=1}^{N} a_i^2} = e^{j\delta}. \tag{4.16}$$

The phase $\arg(\mathrm{CCS}) = \delta$ and the magnitude $|\mathrm{CCS}| = 1$. Clearly, when either the local disparities Δ_i do not agree or the local magnitudes are different, the $|\mathrm{CCS}| < 1$.

Notice, when the scene is not fronto-parallel, i.e. the disparity is not constant $d_1 \neq 0$, $d_2 \neq 0$, then the magnitudes of the Gabor responses a_{f_i}, a_{g_i} differ in (4.12). It means the $|\mathrm{CCS}| < 1$, but according to experiments we made, it does not cause serious problems for reasonable slants, see experiments in Sec. 4.5.

This derivation is valid only for a mono-frequency signal. Nevertheless it holds quite well for general signals containing all frequencies. Let us assume, there is a constant

[2] In practice, this estimate of disparity gradient is accurate for mono-frequency signals only. For more complex composed signals, this estimate is imprecise.

disparity d between image signals. Any image signal can be expressed as a linear combination of elementary harmonics

$$\begin{aligned} f(x,y) &= \sum_i a_i \cos(u_i x + v_i y + \varphi_i), \\ g(x,y) &= \sum_i a_i \cos(u_i (x+d) + v_i y + \varphi_i). \end{aligned} \tag{4.17}$$

Thanks to the linearity of convolution, we can use the superposition principle. Using the same approximation as in (4.12) the Gabor responses are

$$\begin{aligned} G_f(x,y) &\approx \sum_i A_i e^{j(u_i x + v_i y + \varphi_i)}, \\ G_g(x,y) &\approx \sum_i A_i e^{j(u_i (x+d) + v_i y + \varphi_i)}, \end{aligned} \tag{4.18}$$

where $A_i = \frac{1}{2} a_i e^{-\sigma^2 \left((u_0-u_i)^2 + (v_0-v_i)^2\right)}$. The phases of the responses are

$$\begin{aligned} \phi_f(x,y) &= \arg\big(G_f(x,y)\big) = \arctan \frac{\sum_i A_i \sin\big(u_i x + v_i y + \varphi_i\big)}{\sum_i A_i \cos\big(u_i x + v_i y + \varphi_i\big)}, \\ \phi_g(x,y,d) &= \arg\big(G_g(x,y)\big) = \arctan \frac{\sum_i A_i \sin\big(u_i (x+d) + v_i y + \varphi_i\big)}{\sum_i A_i \cos\big(u_i (x+d) + v_i y + \varphi_i\big)}. \end{aligned} \tag{4.19}$$

These are complicated non-linear formulas with unknown A_i and φ_i. The disparity d cannot be obtained directly. Assuming the disparity is small, we approximate $\phi_g(d)$ with a Taylor series centered at $d = 0$

$$\phi_g(x,y,d) = \phi_g(x,y,0) + \left(\frac{\partial \phi_g(x,y,d)}{\partial d}|_{d=0}\right) d + \epsilon. \tag{4.20}$$

The term $\phi_g(0)$ is equal to ϕ_f. The proportionality term is the local frequency, since

$$\frac{\partial \phi_g(x,y,d)}{\partial d}|_{d=0} = \frac{\partial \phi_g(x,y,d)}{\partial x}|_{d=0} = \frac{\partial \phi_f(x,y)}{\partial x} = u_f. \tag{4.21}$$

Then

$$\phi_g = \phi_f + u_f d + \epsilon. \tag{4.22}$$

When the linear approximation is good enough, i.e. ϵ can be neglected, we obtain the same formula as (4.13)

$$d \approx \frac{\phi_g - \phi_f}{u_f}. \tag{4.23}$$

To summarize the analysis, the source of errors in computing the CCS is in the following:

1. Finite-window convolution,
2. Neglecting the influence of symmetric frequencies,
3. Linear approximation of the multi-frequency signal phase,
4. Noise,
5. Non-constant disparity within a window.

The convolution (4.11) must be performed in a finite (and small) window. The Gabor filter is well localized in the spatial domain. A contribution of the Gabor tails which are cut off is negligible when the half-window size is more than three times larger than Gabor scale σ.

A neglection of the negative frequencies in the Gabor responses (4.12) is possible without loss of accuracy, unless the tuning frequency (u_0, v_0) is not too close to zero. Anyway, the estimate of disparity is numerically ill-posed for low frequencies, since the local frequency is in denominator in (4.5). Therefore our bank starts at 0.2π.

Linear approximation of the phase of the response to the multi-frequency signal in computing the disparity (4.20) works well. This is due to high frequency localization of the Gabor filters, inversely proportional to the scale σ. The other reason it works well is that formula (4.6) aggregates the data from various filter bands and it averages out weak response estimates and symmetrically invalid estimates. Therefore, we do not need to use any mechanism to select stable frequencies as in [44, 175]. Using the CCS statistic instead of relying on a single (stable) frequency proves also in case of noise.

A non-constant disparity within a window remains a problem. Naturally, in this situation CCS statistic behaves better using filters with smaller spatial extent. A version of the CCS insensitive to the linear disparity (slanted plane) will be presented in Sec. 4.3.

4.2.4 Justification of the CCS

A justification of the CCS can be found in an excellent paper by Hel-Or and Teo [61] which unifies the problems of steerability, motion estimation and invariant features using Lie group theory. A family of translations $\{\mathrm{T}(\Delta), \Delta \in \mathcal{R}\}$ by a real shift Δ is a Lie group. Therefore, many properties of the group may be revealed by studying infinitesimal actions of the group.

Considering the linear analysis it is easy to show that for the group action in the image domain, there exists an equivalent group action in the filter domain, i.e. the following responses as scalar products are equivalent

$$\langle \phi(x), i(x-\Delta) \rangle = \langle \phi(x+\Delta), i(x) \rangle, \tag{4.24}$$

where $\phi(x)$ is a function representing the filter, $i(x)$ is a function representing the (one-dimensional) image. It immediately follows from changing the coordinates in the scalar product integral in the left hand side $\int_{-\infty}^{\infty} \phi(t) i(t-\Delta) dt$ such that $t - \Delta = x, dt = dx$. Then it becomes $\int_{-\infty}^{\infty} \phi(x+\Delta) i(x) dx$ which is the right hand side of the equation (4.24).

A space of real harmonic functions $\Phi(x) = [\cos ux, \sin ux]^T$ forms so called *equivariant function space* under translation, since there exists a linear mapping to the space of functions after translation

$$\hat{\Phi}(x, \Delta) = \Phi(x + \Delta) = \mathbf{A}(\Delta)\Phi(x), \tag{4.25}$$

where $\mathbf{A}(\Delta)$ is a 2×2 steering matrix of functions of Δ. To find it, a constant steering matrix $\mathbf{B}$ for infinitesimally small translation is found such that

$$\frac{\partial \Phi(x + \Delta)}{\partial \Delta}|_{\Delta=0} = \mathbf{B}\Phi(x). \tag{4.26}$$

In our case

$$\begin{bmatrix} -u \sin ux \\ u \cos ux \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & -u \\ u & 0 \end{bmatrix}}_{\mathbf{B}} \begin{bmatrix} \cos ux \\ \sin ux \end{bmatrix}, \tag{4.27}$$

then the steering matrix is

$$\mathbf{A}(\Delta) = e^{\Delta \mathbf{B}} = \begin{bmatrix} \cos u\Delta & -\sin u\Delta \\ \sin u\Delta & \cos u\Delta \end{bmatrix}. \tag{4.28}$$

Denoting the responses as

$$\begin{aligned} R &= \langle \cos ux, i(x) \rangle, \\ I &= \langle \sin ux, i(x) \rangle, \\ \hat{R} &= \langle \cos ux, i(x - \Delta) \rangle = \langle \cos u(x + \Delta), i(x) \rangle, \\ \hat{I} &= \langle \sin ux, i(x - \Delta) \rangle = \langle \sin u(x + \Delta), i(x) \rangle, \end{aligned} \tag{4.29}$$

the steering equation (4.25) becomes

$$\begin{bmatrix} \hat{R} \\ \hat{I} \end{bmatrix} = \begin{bmatrix} \cos u\Delta & -\sin u\Delta \\ \sin u\Delta & \cos u\Delta \end{bmatrix} \begin{bmatrix} R \\ I \end{bmatrix}. \tag{4.30}$$

The motion estimation task is to find unknown translation Δ having the responses $\hat{R}$, $\hat{I}$ and R, I. The steering matrix is a rotation matrix. Denoting $\hat{Z} = \hat{R} + j\hat{I}$ and $Z = R + jI$ the formula for Δ follows from a phase difference of the two complex numbers

$$\Delta = \frac{\arg(\bar{Z}\hat{Z})}{u}. \tag{4.31}$$

This is the same formula as (4.5).

Similarly, we can find the invariant. It is a function of the responses that is constant for all translation group action. The following must hold

$$h\big(\hat{\Phi}(x, \Delta)\big) = h\big(\Phi(x)\big), \forall \Delta \in \mathcal{R}. \tag{4.32}$$

Specifically,

$$h\big(\cos u(x+\Delta), \sin u(x+\Delta)\big) = h\big(\cos ux, \sin ux\big), \forall \Delta \in \mathcal{R}. \tag{4.33}$$

Consequently, the invariant function must vanish for infinitesimally small translations. This results in partial differential equation for $h(\phi_1, \phi_2)$, where $\phi_1 = \cos ux$ and $\phi_2 = \sin ux$

$$\frac{\partial h(\phi_1,\phi_2)}{\partial \phi_1}\frac{\phi_1(x+\Delta)}{\partial \Delta}|_{\Delta=0} + \frac{\partial h(\phi_1,\phi_2)}{\partial \phi_2}\frac{\phi_2(x+\Delta)}{\partial \Delta}|_{\Delta=0} = 0. \tag{4.34}$$

Performing the derivatives, the equation becomes

$$-u\phi_2\frac{\partial h(\phi_1,\phi_2)}{\partial \phi_1} + u\phi_1\frac{\partial h(\phi_1,\phi_2)}{\partial \phi_2} = 0. \tag{4.35}$$

The solution is

$$h(\phi_1, \phi_2) = f(\phi_1^2 + \phi_2^2), \tag{4.36}$$

where f is any scalar function. In case f is a square root, then the invariant is a magnitude of the complex number of responses in (4.29):

$$|Z| = \sqrt{R^2 + I^2} = \sqrt{\hat{R}^2 + \hat{I}^2} = |\hat{Z}|. \tag{4.37}$$

The harmonic base functions $\Phi = [\cos ux, \sin ux]^T$ used are not localized in the spatial domain. The extent of these filters is infinite. However, our images are finite, moreover the constant translation between image signals is a local property, i.e. it is valid in a small neighbourhood only. In order to get the locality, the filters are multiplied by a Gaussian envelope, centered at zero with a standard deviation σ. The result is the Gabor filter. The contribution of pixels distant more than 3σ is negligible, and we can use a finite domain (a window) for integration in computing the responses.

Unfortunately, such a modification of the base functions does not preserve the equivariance property. Consequently the formula for the motion estimation (4.31) and for the invariant (4.37) holds approximately only.

This is the same situation as in a short time Fourier analysis. The original functions are the base functions of the Fourier transform. After the filter multiplication with the Gaussian window in the spatial domain, the signal spectrum is blurred due to a convolution of the signal spectrum with the Fourier image of the window, which is a Gaussian with a standard deviation inversely proportional to the spatial σ. This is a price for spatial localization. It is known the joint localization in the frequency and spatial domain is optimal for Gabor filter [33].

Collecting the approximate (due to frequency localization little dependent) estimates of translation and translation invariant response magnitudes obtained from several frequencies u in the CCS statistic (4.6) produces precise results in terms of accuracy and invariance. This will be shown in the experiments, see Sec. 4.5.

4.2.5 Usage of the CCS and technical notes

For each scanline we get the correlation table of CCS. This is a finite set of the sub-pixel correspondence hypotheses. The task of the matching algorithm is to select a subset. It can be simplified so that we submit a table of magnitudes |CCS| to a common discrete algorithm, which establishes the integer matches. Then the sub-pixel disparity is obtained by adding the phase of the CCS to them.

There are high correlation values |CCS| in the vicinity of the truth (sub-pixel) matches, which are due to the fact that each CCS phase aims at the same point, as in Fig. 4.1. The maximum |CCS| is not sharp as a consequence. It might be a problem for some algorithms. So we rearrange the table of magnitudes. The resolution of the table in the direction of the disparity correction is increased twice by adding 1/2 pixel cells. The correlations including the phase are binned in these new cells. Then the magnitude in each bin is determined from the correlation which has the smallest phase. Magnitudes in empty bins are set to zero.

Our filter bank usually contains 50 filters, 10 in the horizontal, 5 in the vertical frequency direction. Gabor filter in the frequency domain is a Gaussian centered at the tuning frequency with a standard deviation proportional to the scale $1/\sigma$. So, the tuning frequencies are selected uniformly from $\pm[0.2\pi, 0.8\pi] \times [0.2\pi, 0.8\pi]$. Too high and too low frequencies are excluded because of aliasing and because of the approximations in (4.12). The scale of the filters is selected from the range $\sigma \in [2, 5]$, depending on the scene complexity. This is a parameter of the method.

4.3 Affine Complex Correlation Statistic

In a previous section we have analyzed the case for the two corresponding 2D signals $f(x, y)$ and $g(x, y) = f(x + d(x, y), y)$, where the disparity map between the signals is a plane

$$d(x, y) = d_0 + d_1 x + d_2 y. \tag{4.38}$$

The disparity of this form deforms a local neighbourhood of the corresponding point as sketched in Fig. 4.3. A square matching window becomes a parallelogram in general case. A factor d_1 causes a stretch of the square window, while d_2 causes its skew.

Fronto-parallel plane We have shown that for fronto-parallel case, i.e. a constant small (sub-pixel) disparity $d(x, y) = d_0$, the complex correlation statistic is

$$\mathrm{CCS}(f(x, y), f(x + d_0, y)) \approx e^{jd_0}, \tag{4.39}$$

so the magnitude $|\mathrm{CCS}| \approx 1$ and phase $\arg(\mathrm{CCS}) \approx d_0$ the truth disparity, both up to the precision discussed above. However, if the disparity is not constant, i.e. the plane is slanted $d_1 \neq 0$ or skewed $d_2 \neq 0$, the statistic is corrupted $|\mathrm{CCS}| < 1$ and $\arg(\mathrm{CCS}) \neq d_0$.

The goal of this section is to present a version of the complex correlation statistic which is insensitive to a small stretch d_1 and small skew d_2. One of the basic ideas is

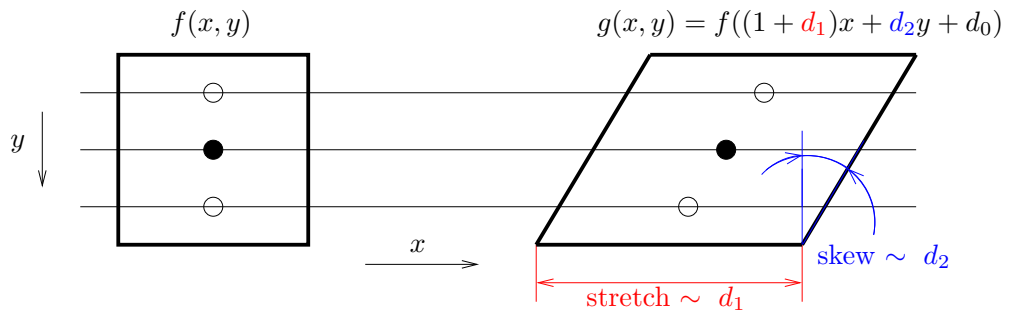


Figure 4.3: Deformation of the local corresponding neighbourhood due to a planar disparity.

that we will not describe 2D signals with responses to 2D filters. Instead, the 2D signal (an image) is decomposed into a set of 1D signals in the rows of the rectified image using 1D filters, the estimation is performed independently, and the results aggregated to create a complex correlation statistic between 2D neighbourhoods.

Skewed plane This is the situation where $d_0 \neq 0$, $d_1 = 0$, $d_2 \neq 0$. This distortion arises due to a plane which is rotated vertically, i.e. around the stereo baseline. It usually occurs in ground-level plane in stereo-images. It causes different displacements in the rows in the matching window, see Fig. 4.3. Therefore, we compute 1D complex correlation statistics for signals in middle row CCS^0 and several neighbouring rows $\mathrm{CCS}^i, i = 1 \ldots n$, of the corresponding windows independently

$$\begin{aligned}
\mathrm{CCS}^0 &= \mathrm{CCS}(f(x, y=0), f(x+d_0, y=0)) \approx e^{jd_0} \\
\mathrm{CCS}^1 &= \mathrm{CCS}(f(x, y=1), f(x+d_0+d_2, y=1)) \approx e^{j(d_0+d_2)} \\
\mathrm{CCS}^{-1} &= \mathrm{CCS}(f(x, y=-1), f(x+d_0-d_2, y=-1)) \approx e^{j(d_0-d_2)} \\
&\vdots \\
\mathrm{CCS}^{\pm n} &= \mathrm{CCS}(f(x, y=\pm n), f(x+d_0 \pm d_2 y, y=\pm n)) \approx e^{j(d_0 \pm d_2 n)}.
\end{aligned} \tag{4.40}$$

These row statistics are aggregated with

$$\mathrm{CCS}_{\mathrm{skew}} = \sqrt{\frac{(\mathrm{CCS}^0)^2 + \mathrm{CCS}^1\mathrm{CCS}^{-1} + \cdots + \mathrm{CCS}^n\mathrm{CCS}^{-n}}{n}}. \tag{4.41}$$

For the true match it is $\mathrm{CCS}_{\mathrm{skew}}(f(x,y), f(x+d_0+d_2y), y) \approx e^{jd0}$. The precision depends on the precision of the row statistics. It holds well for small d_2 and small 'vertical half-window size' n, since required estimate of the displacement in the extreme case is $d_0 \pm nd_2$. If this is a large number the estimate is inaccurate and consequently the statistic $\mathrm{CCS}_{\mathrm{skew}}$ is corrupted. On the other hand larger vertical window helps to

average out errors in the row estimates. We observed that maximum $n = 2$, which works quite well for skews up to $|d_2| < 0.5$. Note that the above formula is insensitive to any row-symmetric transformation, not just the linear skew.

By default, d_2 is not explicitly computed. However, it can be calculated if we consider the linear dependence of the phase of CCS in rows

$$\arg(\mathrm{CCS}^y) = d_0 + d_2 y,\ y = -n \dots n. \tag{4.42}$$

Then d_2 (together with d_0) is estimated by the least-square fit. We used $|\mathrm{CCS}^y|$-weighted least squares for this purpose.

Stretched plane This is the case where $d_0 \neq 0$, $d_1 \neq 0$, $d_2 = 0$. This distortion of the neighbourhoods occurs when the plane in the scene is rotated horizontally, i.e. around the vector perpendicular to the stereo-baseline and the normal of a common image plane of rectified cameras. It manifests by elongating or shortening of the corresponding neighbourhood in the direction of epipolar lines (rows).

As we have seen in (4.12), non-zero d_1 causes that responses of the equivalently tuned filters differ. This happens in 1D case as well. The idea is to compensate for the stretch d_1 with a different filter such that the responses of the original and distorted signal are equal.

Theorem 4.3.1. *Let $f(x)$, $g(x)$ be two signals such that $g(x) = f(x + d_1 x)$ and an arbitrary filter $c(x)$, for $x \in (-\infty, \infty)$. Then the responses (scalar products) are equal if $\langle f(x), c(x)\rangle = \langle g(x), (1 + d_1)c((1 + d_1)x)\rangle$.*

Proof. The scalar product

$$\langle f(x), c(x)\rangle = \int_{-\infty}^{\infty} f(t)c(t)dt.$$

After changing the coordinates to $t = x + d_1 x$ and $dx = (1 + d_1)dt$, we get

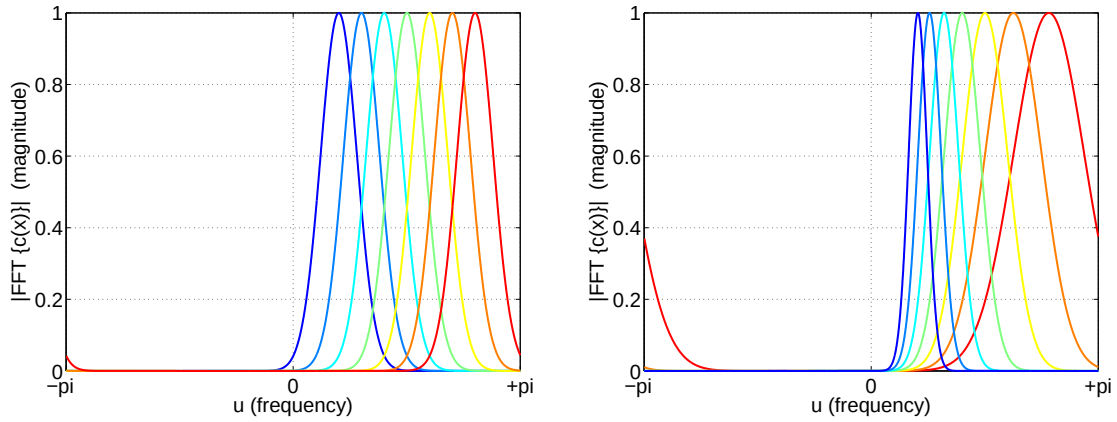
$$\int_{-\infty}^{\infty} f(x + d_1 x)(1 + d_1)c(x + d_1 x)dx = \langle g(x), (1 + d_1)c(x + d_1 x)\rangle.$$

□

The Theorem 4.3.1 says how to shape the filter to compensate for the slat. The resulting filter is the filter with changed coordinates multiplied with a Jacobian of the transformation. This holds exactly in a continuous (Hilbert) space. Unfortunately in the discrete space there are limitations due to the Nyquist limit.

Having the Gabor filter $c(x \mid u_0,\ \sigma)$ with tuning frequency u_0 and scale σ. Then the filter which compensates for the stretch d_1 is

$$(1 + d_1)\ c\big(x \mid u_0(1 + d_1),\ \frac{\sigma}{(1 + d_1)}\big).$$



(a) CCS (b) $\mathrm{CCS}_{\mathrm{stretch}}$

Figure 4.4: The bank of filters. Magnitude spectra.

The $d_1 > 0$ increases the tuning frequency u_0 and makes the filter more narrow in spatial domain, i.e. wider in the frequency domain. So, too large d_1 may cause the Nyquist limit π is broken and filters are corrupted by aliasing.

The stretch-insensitive complex correlation statistic is obtained by searching the maximum CCS for stretch d_1

$$\mathrm{CCS}_{\mathrm{stretch}} = \arg \max_{\mathrm{CCS}(d_1)} |\mathrm{CCS}(d_1)|, \tag{4.43}$$

where $\mathrm{CCS}(d_1)$ means a Complex Correlation Statistic which is computed from Gabor responses with filters shaped according to Theorem 4.3.1 to compensate for the stretch d_1.

$$\begin{aligned} G_{f_i} &= \langle f(x), c_i(x \mid u_{0i}, \sigma_i) \rangle, \\ G_{g_i}(d_1) &= \langle g(x), (1+d_1) c_i(x \mid (1+d_1) u_{0i}, \sigma_i/(1+d_1) \rangle. \end{aligned} \tag{4.44}$$

The problem is that searching the optimum stretch d_1 requires shaping the filters for its compensation and computing the responses (re-convolution). This is very expensive. Moreover it breaks the desired property, that all the responses of a fixed filter bank are computed beforehand and the CCS is computed from them.

In subsequent paragraph, we show how to design a small bank of Gabor filters which has a property that the filters in the bank are of the form that mutually compensate for certain stretch d_1 and the bank still captures well the image spectra. The search in (4.43) is then reduced to the maximum selection from a small number of CCS computation from a small number of precomputed responses.

The bank consists of a central filter with tuning frequency u_0 and scale σ and filters shaped to compensate for a set of stretches

$$\mathcal{D}_1 = \{1 - (1-\delta)^i, i = -N \dots N\}. \tag{4.45}$$

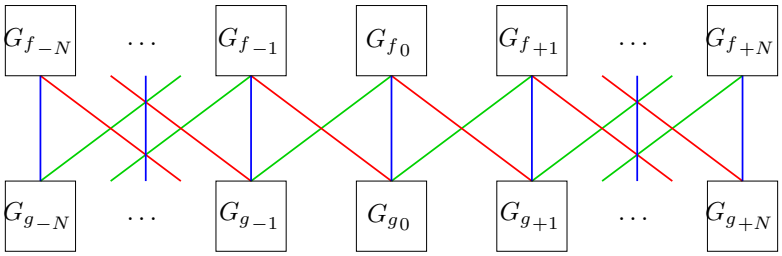


Figure 4.5: Illustration of computing CCS from mutually shifted responses: $\mathrm{CCS}(d_1 = 0)$ in blue, $\mathrm{CCS}(d_1 = 1 - (1+\delta)^{+1})$ in green, $\mathrm{CCS}(d_1 = 1 - (1+\delta)^{-1})$ in red.

The bank consists of $2N + 1$ filters

$$c_i(x) = (1-\delta)^i c(x \mid u_0(1-\delta)^i, \frac{\sigma}{(1+\delta)^i}),\ i = -N \ldots N. \tag{4.46}$$

The signals $f(x)$ and $g(x)$ are convolved with the filter bank. The complex correlation statistic compensating stretch $d_1 \in \mathcal{D}_1$ is computed from 'shifted' responses, as sketched in Fig. 4.5,

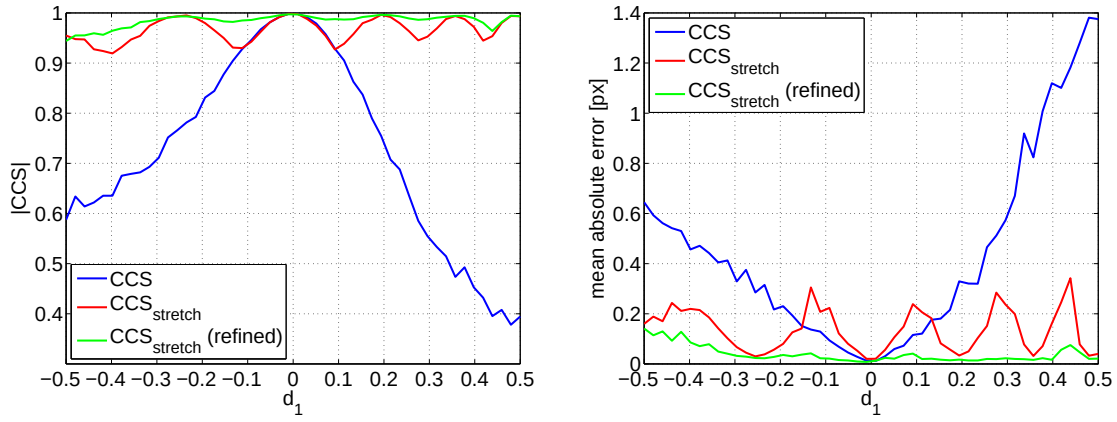
$$\mathrm{CCS}(d_1 = 1 - (1-\delta)^k) = \begin{cases} \frac{2\sum_{i=-N}^{N-k} |G_{f_i}|\ |G_{g_{i+k}}| e^{j\Delta_i}}{\sum_{i=-N}^{N-k} |G_{f_i}|^2 + \sum_{i=-N}^{N-k} |G_{g_{i+k}}|^2}, & k \geq 0, \\ \frac{2\sum_{i=-N-k}^{N} |G_{f_{i+k}}|\ |G_{g_i}| e^{j\Delta_i}}{\sum_{i=-N-k}^{N} |G_{f_{i+k}}|^2 + \sum_{i=-N-k}^{N} |G_{g_i}|^2}, & k < 0, \end{cases} \tag{4.47}$$

where Δ_i is the local sub-pixel disparity is computed from responses mutually shifted in the same sense. Note that for $k = 0$ we have the complex correlation statistic for front-parallel scene. Note also that for large stretches, i.e. large $|k|$, the $\mathrm{CCS}(d_1)$ is computed from a smaller number of filters than for the fronto-parallel case, which makes the CCS slightly less accurate.

The above method with compensated stretches in discrete steps exploits the fact that CCS deteriorates relatively slowly with a different stretch than it is compensated for. There exists a trade-off between the accuracy and the number of filters in the bank and speed of the search for the maximum.

We experimentally constructed the bank with the central filter c_0 with $u_0 = 0.4\pi$, $\sigma = 4$, stretch step $\delta = 0.2$, and $N = 3$. Hereby this filter bank is able to compensate for the slants $\mathcal{D}_1 = \{-0.9531, -0.5625, -0.2500, 0, 0.2000, 0.3600, 0.4880\}$. The spectra of filters in the bank are shown in Fig. 4.4(b). It captures well the image spectra[3] and hereby it is well suitable for matching. Compare with the filter bank with equally sampled spectrum which is used in the standard fronto-parallel CCS, see Fig. 4.4(a). For

[3]The highest-frequency filter shown is slightly corrupted by aliasing as can be seen. This could be avoided by two factors: lower tuning frequency u_0 of the basic filter would shift the bank towards the signal mean value, or smaller extent in the frequency domain would be due to larger scale σ (spatial extent). Both factors are not desirable.



(a) Invariance - CCS magnitude

(b) Accuracy - CCS phase

Figure 4.6: Invariance and accuracy as a function of slant d_1.

the purposes of matching a guaranteed level of discriminability (ability to recognize the correct match) is necessary, otherwise the speed is an important issue since the statistic is computed for the entire correlation table.

On the other hand, once correct matches are identified (up to integer precision), a procedure for computing an accurate estimate of the sub-pixel disparity can be more computationally intensive, since the number of matches is much smaller than the number of all possible matches in the matching table.

Therefore, we propose a procedure for refinement of CCS estimate. First, a more precise stretch d_1 is found by locating a maximum of $|\mathrm{CCS}(d_1)|$ in between stretch steps via fitting a 3-point parabola. Location of its vertex is the refined stretch d_1^*. Then, d_1^* is used to find responses $G_{g_i}(d_1^*)$. These responses are obtained by interpolating the responses G_{g_i} precomputed in discrete steps of d_1, see (4.45). The interpolation[4] of responses is quite precise, since this signal is smooth and cannot change abruptly due to spectral overlap of filters, unlike interpolating raw image intensity signals directly. Finally refined $\mathrm{CCS}^*_{\mathrm{stretch}}$ is computed from responses G_{f_i} and $G_{g_i}(d_1^*)$. The complex correlation statistic refined with this procedure is more accurate, especially for cases where the true stretch d_1 lies in the middle of stretch steps.

This improvement can be seen in Fig. 4.6. This figure shows results of a simple synthetic experiment, where two (white-noise) signals were stretched according to a range of slants d_1. We measured the magnitude of three different complex correlation statistics and the accuracy of their phase as a mean absolute error from ground-truth disparity. All plots are average values from 10 random trials. We can see the standard fronto-parallel CCS possesses a small insensitivity to the stretch distortion. The $\mathrm{CCS}_{\mathrm{stretch}}$ compensate the distortion in discrete steps of $\mathcal{D}_1$, which is clearly visible by peaks and

[4] We used a linear interpolation but in principle any other technique can be used.

valleys at the compensation steps in magnitude and phase accuracy plot respectively. The refined version has the best performance, both in invariance of the magnitude and the phase accuracy of the complex correlation statistic.

General plane This is the situation where all $d_0 \neq 0$, $d_1 \neq 0$ and $d_2 \neq 0$. And it occurs when the plane is in a general position and causes both stretch and skew distortion simultaneously. We call it an affine distortion.

The affine-insensitive complex correlation statistic is computed as skew-insensitive statistic $\mathrm{CCS}_{\mathrm{skew}}$ of stretch insensitive statistics calculated in rows of the matching window,

$$\mathrm{CCS}_{\mathrm{aff}} = \sqrt{\frac{(\mathrm{CCS}^0_{\mathrm{stretch}})^2 + \mathrm{CCS}^1_{\mathrm{stretch}}\mathrm{CCS}^{-1}_{\mathrm{stretch}} + \cdots + \mathrm{CCS}^n_{\mathrm{stretch}}\mathrm{CCS}^{-n}_{\mathrm{stretch}}}{n}}. \tag{4.48}$$

The statistic for the correct match is $\mathrm{CCS}_{\mathrm{aff}}(f(x,y), f(x+d_0+d_1x+d_2, y)) \approx e^{jd_0}$. The accuracy of this composed statistic depends on the accuracy of elementary statistics. It is designed to keep the insensitive properties with reasonable discriminability for slants $|d_1| \leq 0.5$ and $|d_2| \leq 0.5$.

The slants d_1 and d_2 are the x and y components respectively of the disparity gradient. A local surface normal can be computed from them, [128, 102, 71]. The slants d_1 and d_2 are not computed by default for the purposes of matching, but can be easily obtained by the disparity refining procedures of given integer matches as described above.

4.4 Global approach to sub-pixel disparity correction

A fundamental drawback of the above methods for sub-pixel disparity estimation is that they are local in the sense that the estimation is performed in each pixel independently. The only influence is via a partially shared neighbourhood of the corresponding pixels, i.e. a spatial domain of Gabor filters in complex correlation statistics, or a square window in methods like `dispcor`. The `dispcor` type methods search the best (affine) fit of the matching windows in corresponding locations of a stereo-pair, e.g. [56, 35, 128]. For each pixel (x_0, y_0) of a reference image, they independently solve a kind of the following optimization problem

$$(d_0, d_1, d_2)^* = \arg\min_{d_0,d_1,d_2} \sum_{x,y\in\mathcal{N}} \big(\mathbf{I_l}(x_0+x, y_0+y) - \mathbf{I_r}(x_0+x+d_{\mathrm{int}}+d_0+d_1x+d_2y, y_0+y)\big)^2, \tag{4.49}$$

where $\mathcal{N}$ is a small neighbourhood of the corresponding pixels, typically a square window (5×5 pixels in [128]), d_{int} is a given integer disparity, obtained by a matching algorithm beforehand.

There are several sources of error: (1) Computing the criterion involves an interpolation of image intensities at sub-pixel locations, especially for high-frequency image

content. (2) On the other hand, low frequency image content also makes problems, since the optimum of (4.49) is flat and a small noise can cause a large error. (3) The problem of stucking in a wrong local optimum is also present. Estimated sub-pixel disparity map may be quite erratic as a result, see later the comparative experiments in Sec. 4.5.

Therefore, we propose a global formulation, where estimating a sub-pixel disparity map is formulated as a single *continuous* optimization problem which consists of sum of squared error data term and an explicit regularization term to support disparity map smoothness. The regularizer helps to handle both low-textured areas and mitigate the influence of interpolation errors.

Similar problems have appeared in literature. In [124, 94] the task is formulated using the variational calculus. The variational methods have been described in Sec. 2.1.1 in paragraph (A.2). The 3D surface which is sought is expressed as a continuous function, where the regularization term consists of its differential characteristics. The work [147] presents very good results of surface reconstruction from multiple images. The authors formulate a global optimization task with a probabilistic term for occlusion. Similar work is [50] where the task is formulated in Bayesian framework searching for the maximum aposteriori probability 3D surface. They use EM algorithm to optimize. In the E-step the visibility is estimated from current model, in the M-step they use current visibility to optimize the log-likelihoods. In [158], a robust prior term is proposed. The prior term usually penalizes disparities in the neighbourhood quadratically to support smoothness. But, occlusions breaks this assumption. The authors suggest to use a Gaussian kernel in prior term instead of the quadratic term. The term is robust to occlusions because of the low influence of distant neighbouring disparities. However, a statistical efficiency of this prior is lower, the model cannot cope with errors from wrong initialization.

Our situation is simpler, we start from an integer disparity map obtained by a reliable matching algorithm, which delivers also the map of occlusions. We use the following method as a sub-pixel disparity correction which is nevertheless able to correct some mismatches, but not as a complete algorithm to recover 3D surface from images like the methods mentioned above.

The problem to find the optimal (sub-pixel) disparity map $\mathbf{d}$ of the size $M \times N$ pixels is formulated as

$$\mathbf{d}^* = \arg \min_{\mathbf{d} \in \mathcal{R}^{MN}} J(\mathbf{d}). \tag{4.50}$$

The criterion is

$$J(\mathbf{d}) = D(\mathbf{d}) + \lambda \, R(\mathbf{d}), \tag{4.51}$$

where the data term is a sum of squared differences

$$D(\mathbf{d}) = \sum_{x=1}^{N} \sum_{y=1}^{M} \big(\mathbf{I_l}(x, y) - \mathbf{I_r}(x + \mathbf{d}_{x,y}, y)\big)^2, \tag{4.52}$$

the regularization term is a sum over all pairwise horizontal and vertical edges between

adjacent pixels with a quadratic penalty

$$R(\mathbf{d}) = \sum_{x=1}^{N-1} \sum_{y=1}^{M-1} \big((\mathbf{d}_{x,y} - \mathbf{d}_{x+1,y})^2 + (\mathbf{d}_{x,y} - \mathbf{d}_{x,y+1})^2 \big), \tag{4.53}$$

and λ is a weight of the regularization term. To reduce the number of parentheses, we used a simpler notation $\mathbf{d}_{x,y} = \mathbf{d}(x, y)$ for disparities at integer (pixel) locations (x, y). This is a large optimization problem containing the same number of variables as the number of pixels in the reference (left) image. As it will be shown, this problem can be successfully solved by gradient descent initialized by the integer disparity map.

The gradient of the criterion $J(\mathbf{d})$ is stacked in an array of the same size as the disparity map with components

$$\begin{aligned} \delta\mathbf{J}_{x,y} = \frac{\partial J(\mathbf{d})}{\partial \mathbf{d}_{x,y}} = -2\big(\mathbf{I_l}(x, y) - \mathbf{I_r}(x + \mathbf{d}_{x,y}, y)\big) \frac{\partial \mathbf{I_r}(x + \mathbf{d}_{x,y}, y)}{\partial \mathbf{d}_{x,y}} + \\ + 2\lambda\big((\mathbf{d}_{x,y} - \mathbf{d}_{x+1,y})b_R - (\mathbf{d}_{x-1,y} - \mathbf{d}_{x,y})b_L + (\mathbf{d}_{x,y} - \mathbf{d}_{x,y+1})b_B - (\mathbf{d}_{x,y-1} - \mathbf{d}_{x,y})b_T\big), \end{aligned} \tag{4.54}$$

where b_R, b_L, b_B, b_T are binary indicators of right, left, bottom, top image boundary respectively, e.g. $b_R = 0$ for $x = N$, otherwise $b_R = 1$.

Note that calculating the data-term (4.52) and the gradient (4.54) involves interpolation of image intensities. We used bicubic interpolation which is more accurate than linear. The intensity derivative term in (4.54) is approximated as a difference of intensities

$$\frac{\partial \mathbf{I_r}(x + \mathbf{d}_{x,y}, y)}{\partial \mathbf{d}_{x,y}} \approx \mathbf{I_r}(x + \mathbf{d}_{x,y}, y) - \mathbf{I_r}(x + \mathbf{d}_{x,y} + 1, y). \tag{4.55}$$

The disparity map of a scene is not entirely smooth as assumed by the described regularizer (4.53), but it is typically piece-wise smooth due to multiple objects in the scene. Since we have the disparity map along with a map of occlusions and occlusion boundaries, it is easy to modify the optimization problem: All occluded pixels or any other pixels the algorithm did not assign a valid disparity are removed from the sum in data-term, regularization-term, and consequently their gradient component is set to zero. The neighbours of unassigned pixels are handled the same as they are at the image boundary. Similarly, if there is a step in disparity (due to occlusion boundary), the regularization term is 'disconnected' by removing the appropriate penalty edges. One could use the computational molecules of Terzopoulos [154].

To solve the problem (4.50), a quasi Newton method with an approximate scaled identity Hessian matrix was employed. This method turned out to be a good trade-off between accuracy and speed. We modified a Matlab function `fminunc` to work with sparse matrices.

In Fig. 4.7, there are results for several weights of the regularization term λ. The optimization problem was initialized by the integer stereo matching algorithm, the one described in Chapter 5. To ensure surface homogeneity, small gaps in the disparity map were filled by a primitive algorithm which assigns a median disparity of 5×5 pixel

window around the gap. The input images are first photometrically normalized to a unit variance, (1) to eliminate a potential change of contrast between the left and right image, and (2) to ensure weight λ has a uniform impact on the surface smoothness regardless of the intensity scale of images.

We can see that the surface is over-smoothed for $\lambda = 5$ and $\lambda = 0.5$ in Fig. 4.7(a),(b), and under-smoothed for $\lambda = 0.1$ in Fig. 4.7(c). However, when the optimization problem is first solved with $\lambda = 0.5$ and the resulting (sub-pixel) disparity is used as an initialization for another optimization with $\lambda = 0.1$, we get a satisfactory result. This is in Fig 4.7(d), denoted as $\lambda = \{0.5, 0.1\}$ which is a default setting of the method and it is kept for all other images shown later. The intuition is that the optimization with larger λ is able to get closer to the global optimum (of a wrong over-smoothing criterion though), while with smaller λ the criterion is correct however the optimization gets stucked in a wrong local optimum. Thus most of the over-smoothed details is recovered back in this scheme.

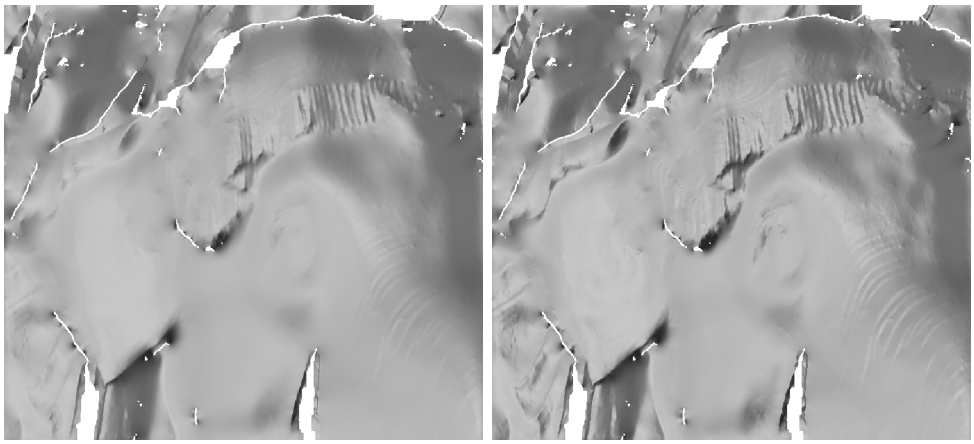
The optimization takes 200 iterations maximum (typically less than 100), which is in time tens of seconds for about 0.5 Mpx images in purely Matlab implementation.

We experimented with initialization by a sub-pixel disparity obtained from previous methods. The results were not qualitatively very different from integer matching initialization. The only difference was in the reduced number of iterations of the optimization process. Furthermore, we also experimented with the case where the λ becomes a function of image intensity, as in conditional Markov Fields. The λ was 'modulated' by a magnitude of image gradient. Nevertheless, the results were rather worse since textured regions stayed under-smoothed.

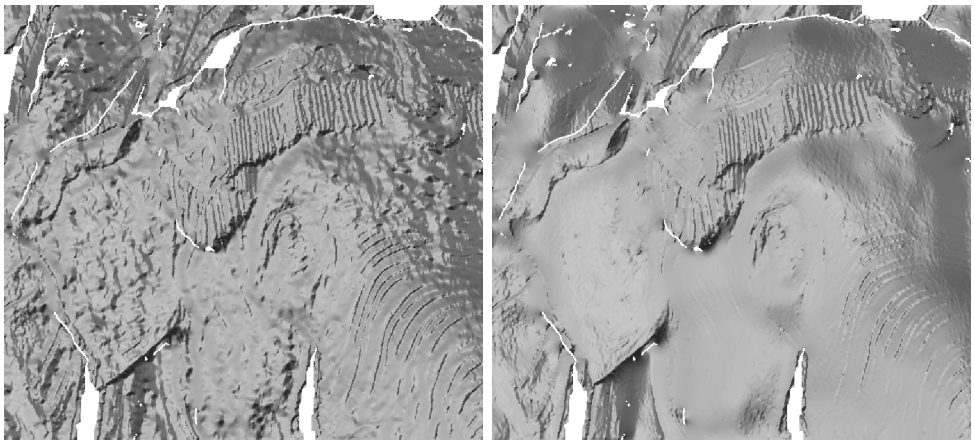
4.5 Experiments

In this section, we experimentally evaluate and compare the performance of the complex correlation statistics described. We demonstrate two important properties of CCS: the magnitude is discriminable and invariant to image sampling and affine distortion, and the sub-pixel disparity estimation of its phase is accurate. The global method for sub-pixel disparity correction is also evaluated.

We made several experiments on both synthetic and real data. The discriminability of the presented CCS is compared to other correlation statistics and the accuracy of the presented sub-pixel estimation is compared to other sub-pixel estimation methods both quantitatively in ground-truth experiments and qualitatively as disparity maps or reconstructed 3D-surfaces.



(a) $\lambda = 5$ (b) $\lambda = 0.5$



(c) $\lambda = 0.1$ (d) $\lambda = \{0.5, 0.1\}$

Figure 4.7: Input images and relighted 3D models for several weights of the regularization term λ.

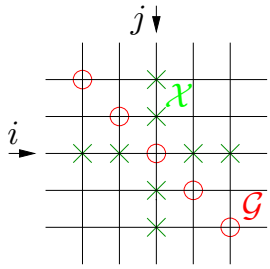


Figure 4.8: The definition of discriminability. Forbidden zone $\mathcal{X}(i,j)$ in green crosses, Ground-truth matches $\mathcal{G}$ as red circles.

4.5.1 Synthetic data

A synthetic stereo-pair of images with a relation $\mathbf{I_r}(x,y) = \mathbf{I_l}(x + \mathbf{d}(x,y), y)$, where $\mathbf{d}(x,y)$ is the prescribed disparity map, was generated using a harmonic basis such that

$$\begin{aligned}\mathbf{I_l}(x,y) &= \sum_{i=1}^{N}\sum_{j=1}^{N/2} a_{ij}\cos\Big(u_i x + v_j y + \varphi_{ij}\Big),\\ \mathbf{I_r}(x,y) &= \sum_{i=1}^{N}\sum_{j=1}^{N/2} a_{ij}\cos\Big(u_i\big(x + \mathbf{d}(x,y)\big) + v_j y + \varphi_{ij}\Big),\end{aligned} \tag{4.56}$$

where frequencies u_i are equally spaced values from interval $[-\pi,\pi]$ and frequencies v_j are equally spaced values from interval $[0,\pi]$. The image size in pixels is $N \times N$. Magnitudes a_{ij} and phases φ_{ij} form together a space of $N \times N$ parameters which were chosen randomly from uniform distribution over intervals $[0,1]$ and $[0,2\pi]$ respectively.

This representation generates a precise stereo-pair and it easily allows to control the frequency content of the images, via setting the relevant a_{ij} magnitudes.

For forthcoming experiments we use 50×50 pixel images. If not stated otherwise, their spectra are limited to $f_h = 0.6\pi$, by setting magnitudes $a_{ij} = 0$ of frequencies u_i, v_j higher than f_h, to avoid aliasing in shrunk images which may occur due to the prescribed disparity map.

In the following experiments, the Gabor scale parameter of CCS was set $\sigma = 2$. For $\mathrm{CCS}_{\mathrm{aff}}$, the central filter has the scale $\sigma = 4$, and the size of the window of other statistics was set to 5×5 pixels.

Discriminability Discriminability is a property which intuitively says that correlation statistic assigns high values to the true corresponding pairs while keeping low values of all other potential matches.

We define the discriminability as a probability that the ground-truth match has all X-zone competitors [129] of lower correlation, see Fig. 4.8, and it is estimated from

$$\mathit{discriminability} = \frac{\mathrm{card}\{(i,j) \in \mathcal{G} : \forall (k,l) \in \mathcal{X}(i,j), c(k,l) < c(i,j)\}}{\mathrm{card}\ \mathcal{G}}, \tag{4.57}$$

where (i, j) is a cell in the correlation table, $\mathcal{G}$ (red circles) is the set of the ground-truth correspondences and the $\mathcal{X}(i, j)$ (green crosses) is the forbidden zone [87] for (i, j), and $c(i, j)$ is the correlation value. The estimation was averaged from 100 random trials over texture generation for each stereopair.

This definition of discriminability is insensitive to the scale of the statistic c. We only need the similarity property: higher similarity implies higher value of c.

We will compare the discriminability of CCS and CCS_{aff} with other (window) statistics: the Sum of Absolute Differences (SAD), the Moravec's Normalised Cross Correlation (MNCC) [105] and the sum of Birchfield-Tomasi sampling insensitive pixel dissimilarities [10] over a window. The SAD and BT is redefined to be a difference of the original statistic from 1 to have the similarity property.

We measured the discriminability for a constant subpixel disparity: $\mathbf{d}(x, y) = d_0,\ d_0 \in [0, 0.5]$ px (fronto-parallel scene), for a horizontally slanted plane (stretch distortion): $\mathbf{d}(x, y) = d_1 x,\ d_1 \in [0, 0.5]$, and for a vertically slated plane (skew distortion): $\mathbf{d}(x, y) = d_2 y,\ d_2 \in [0, 0.5]$. In case of fronto-parallel scene, we also measured the discriminability as a function of the top frequency limit f_h.

The results for the fronto-parallel case are shown in Fig. 4.9(a): The worst case occurs where the true disparity d_0 is 0.5 pixel. The SAD and MNCC tend to fail towards this point, BT has small problems at this point too, although it should be invariant to image discretization, but due to interpolation used in BT, it gets worse as more high frequencies are present, see Fig. 4.9(b). We did not observe any failure of CCS and CCS_{aff} in any of the 100 trials, which corroborates its invariance to image sampling.

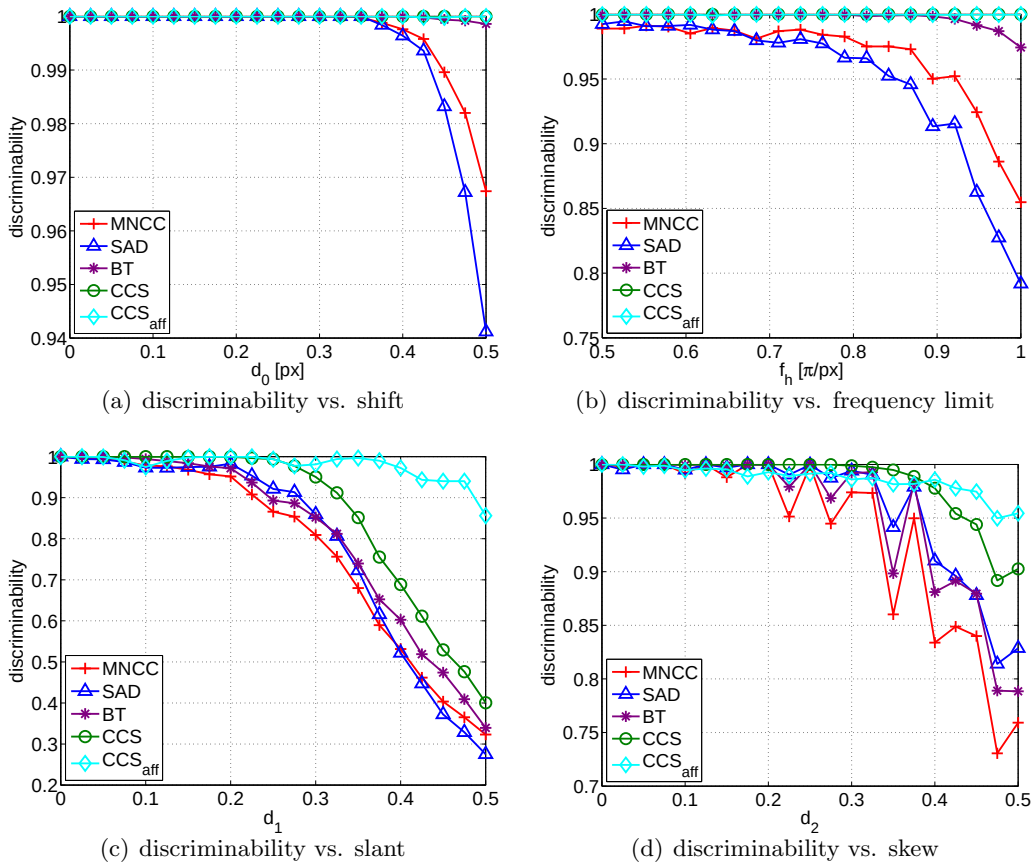
Unlike the others, the CCS and CCS_{aff} have no problems with high frequencies up to the Nyquist limit, see Fig. 4.9(b). These plots show the discriminability where $\mathbf{d}(x, y) =$ 0.5 px versus the upper frequency present in the image spectra.

The results for the horizontally slanted plane are shown in Fig. 4.9(c). The only statistic which is insensitive to this distortion is CCS_{aff}. Although the stretch-compensation steps are noticeable, the discriminability stays high. None of the other tested statistics is invariant to the slant. The discriminability decreases with higher slants d_1 as expected.

The vertical slant (skew distortion) is less harmful compared to the horizontal slant, see Fig. 4.9(d). The CCS_{aff} has the highest discriminability for the largest skew d_2 among all tested statistics. It is not completely invariant due to imprecise row estimates for large displacements. To summarize the CCS is the first or second highest discriminability statistic for all measurements.

Accuracy By matching accuracy we mean that the estimated disparity is close to the ground-truth disparity. We measure the mean absolute error of this difference, but matches whose error in disparity is higher than 1 pixel are considered outliers and excluded. The results are average values from 10 random trials over a texture generation in (4.56) and the plots have errorbars of a standard deviation.

All tested algorithms start from truth integer matching, which was obtained by rounding the ground-truth disparity map to the closest integer. We compare the accuracy obtained from a phase of CCS and CCS_{aff}, from our global sub-pixel correction, from the



(a) discriminability vs. shift

(b) discriminability vs. frequency limit

(c) discriminability vs. slant

(d) discriminability vs. skew

Figure 4.9: Discriminability experiment results.

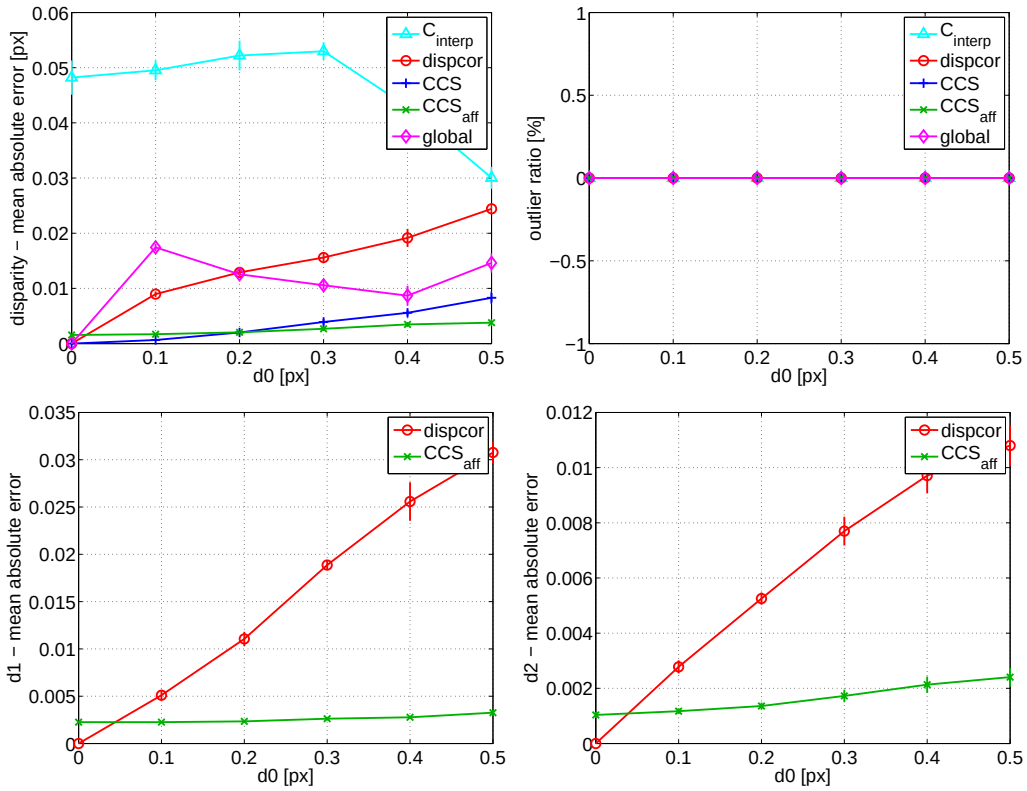


Figure 4.10: Accuracy: Fronto-parallel plane (sub-pixel shift).

independent fitting of an affine window `dispcor` [128], and from finding the maximum correlation via fitting a 3 point parabola into a MNCC statistics C_{interp}.

For all algorithms, we show the mean absolute error of disparity, the percentage of outliers (pixel with errors grater than 1 pixel). For CCS_{aff} and `dispcor`, we show also the errors of the estimate of disparity gradient components $d_1 = \partial \mathbf{d}(x,y)/\partial x$ and $d_2 = \partial \mathbf{d}(x,y)/\partial y$. The other algorithms do not output the disparity gradient components directly.

Results in Fig. 4.10 shows the fronto-parallel plane, where the accuracy is plotted as a function of the sub-pixel shift d_0. Both CCS and CCS_{aff} produce most accurate results. The global method is slightly worse, the peak for $d_0 = 0.1$ is probably caused by the stucking in the local optimum close to initialization. The error of `dispcor` is increasing towards the worst case for $d_0 = 0.5$, where the signals are the most disturbed by discretization artifacts. The simplest method C_{interp} has the highest error. It is reported to be biased towards integer values [142]. Notice that the estimate of disparity gradient components is much more accurate using CCS_{aff} than using `dispcor`. This is due to discretization invariance.

Results for slanted plane in a horizontal and vertical direction are shown in Fig. 4.11

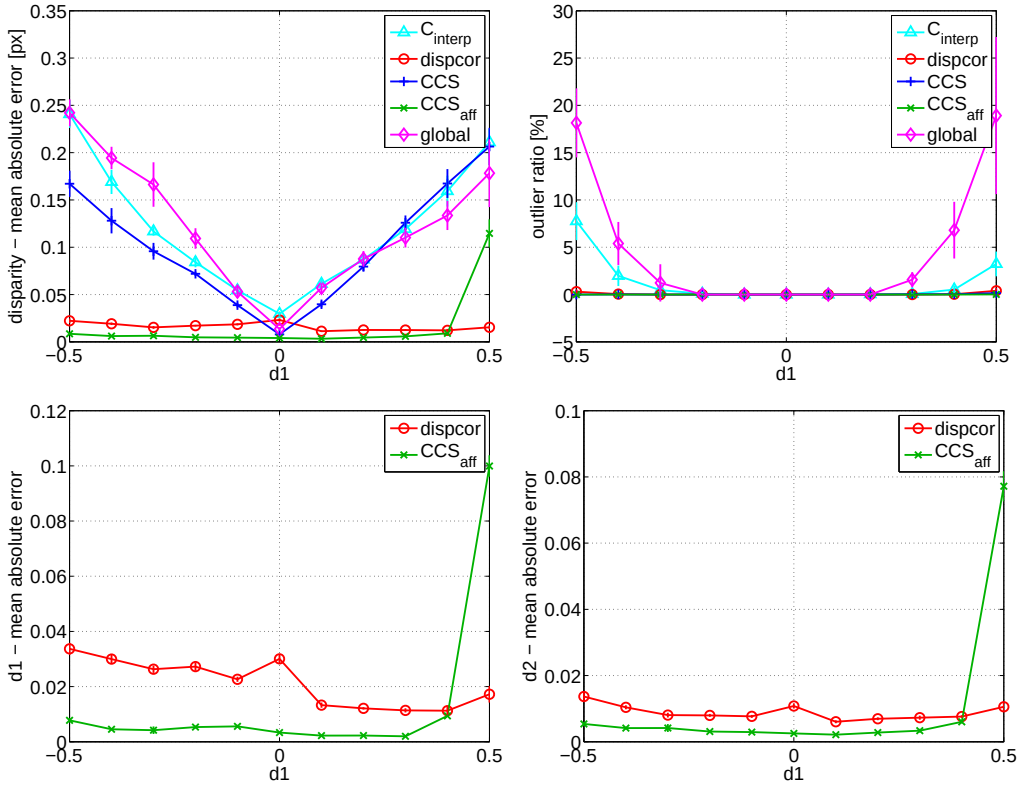


Figure 4.11: Accuracy: Horizontally slanted plane (stretch).

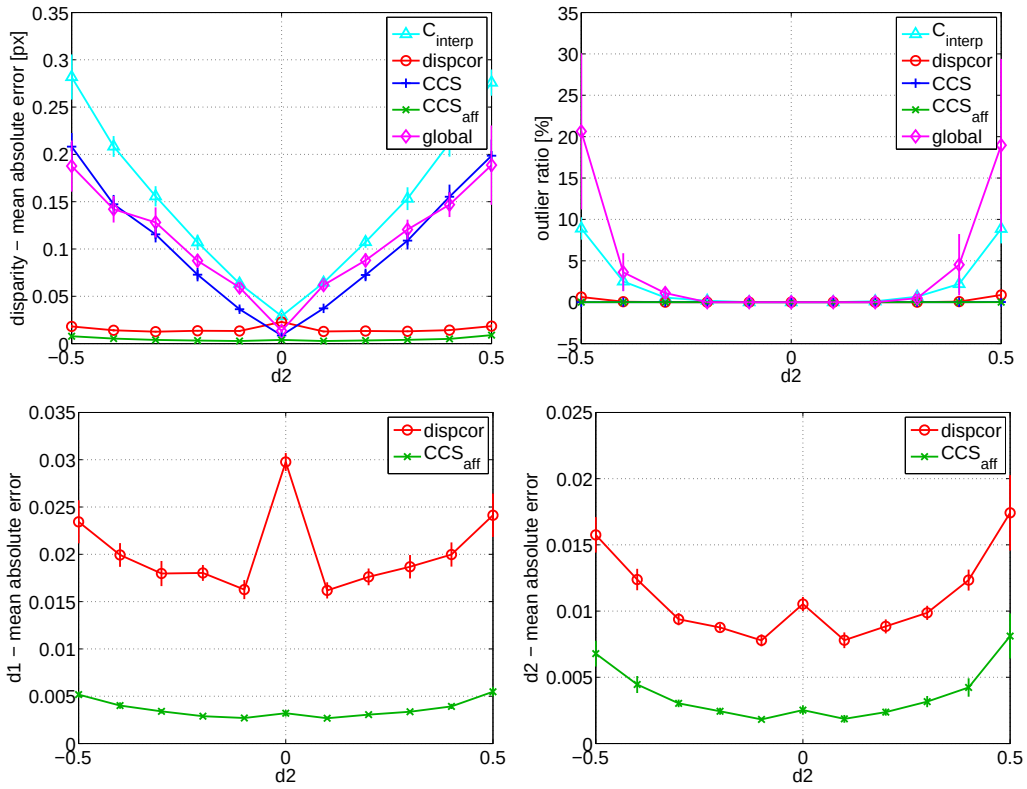


Figure 4.12: Accuracy: Vertically slanted plane (skew).

and Fig. 4.12 respectively. We can see in both cases, that the fronto-parallel model is not preserved in C_{interp}, CCS and the global method which causes the accuracy drops with increasing slant. The estimate using CCS_{aff} is consistently more precise than `dispcor`. The only exception is the case $d_1 = 0.5$, where the slant is out of the compensation interval. The same holds for the estimate of disparity gradient components.

The situation where the geometrical model is not preserved for any of the algorithms is shown in Fig. 4.13. The disparity map is not planar, but spherical. The accuracy is plotted as a function of a radius of the sphere. The smaller radius (the larger curvature) the more the model is violated. We can see, the accuracy of the algorithms do not differ that much. The most accurate is the `dispcor`. The problems of CCS_{aff} are caused by the large spatial extent of its filters compared to a small window of `dispcor`. The deviation from a planar model is small in a small distance from the origin.

Another situation where the geometrical model is not preserved is in Fig. 4.14. This is a step in disparity of up to 5 pixels. Note that the error is higher for all the algorithms compared to previous tests. The model is seriously violated in this case. The disparity estimates of C_{interp}, `dispcor` and CCS_{aff} are very erratic in the neighbourhood of the step. This manifests also by a number of outliers. The CCS tends to smooth out the step

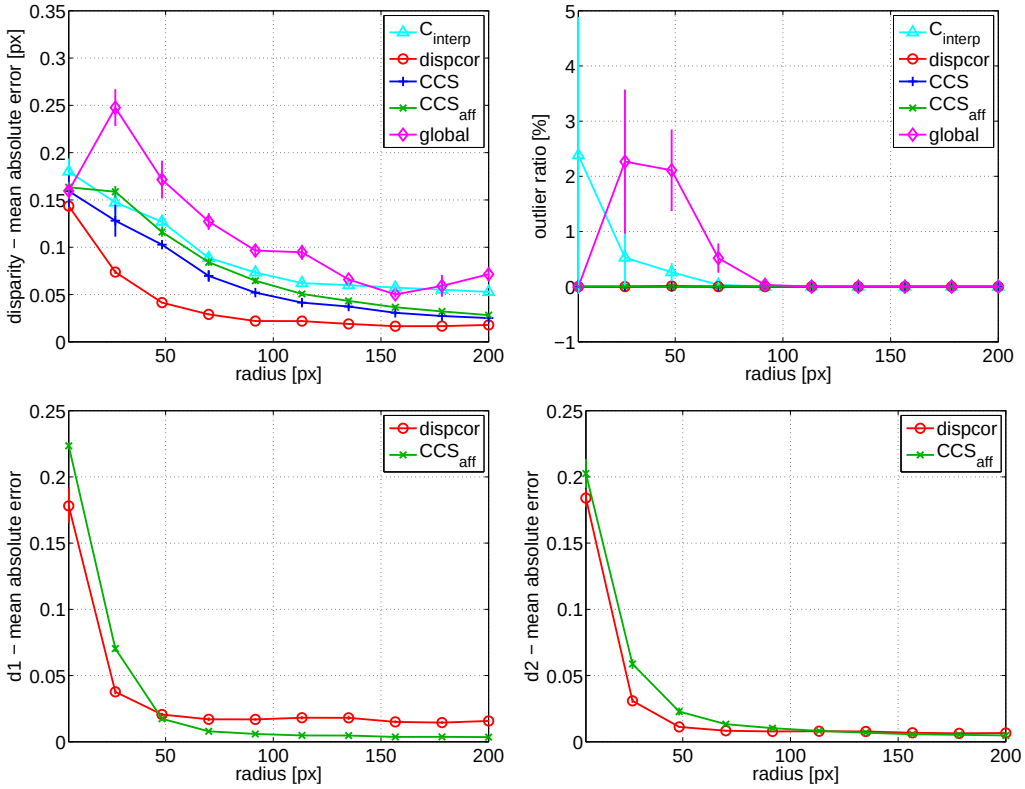


Figure 4.13: Accuracy: Constant curvature (sphere).

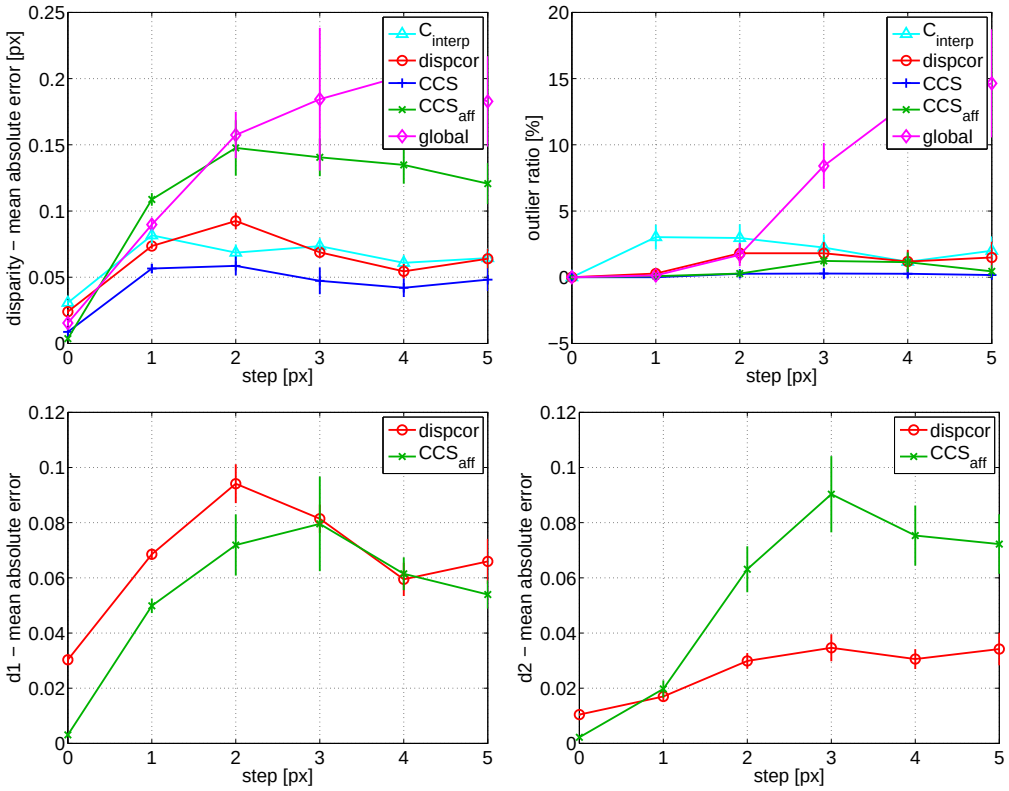


Figure 4.14: Accuracy: Step in disparity in horizontal direction

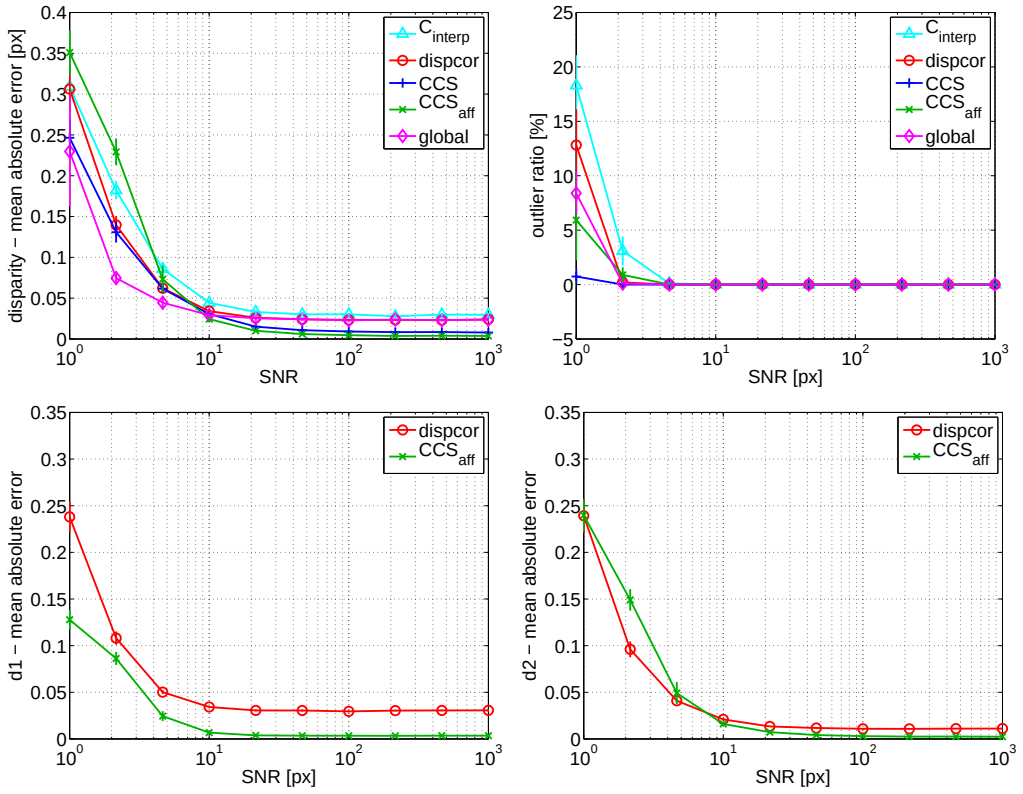


Figure 4.15: Accuracy: Additive Gaussian Noise (SNR), fronto-parallel plane ($d_0 = 0.5$).

consistently without random errors. Similar behaviour occurs in the global method, but it is very over-smoothed producing large errors. Note that for this experiment, the global optimization problem was not 'disconnected' due to occlusion boundary as discussed in Sec. 4.4.

The algorithms were also tested under contamination with additive Gaussian noise up to signal to noise ratio of SNR = 1, see Fig. 4.15. None of the algorithm is extra sensitive to noise. The ordering of the algorithms is more or less preserved from the case without noise. Geometrically it is a fronto-parallel plane $\mathbf{d}(x, y) = d_0 = 0.5$. The error does not grow much up to SNR = 10.

The last experiment demonstrates the accuracy of presented algorithms as a function of image frequency content. The highest frequencies of image spectra are limited in a range from 0.1π to 0.9π. Geometrically, the disparity map is the same as in the previous experiment, a fronto-parallel plane. The results are shown in Fig. 4.16. The algorithms `dispcor`, and the global method deteriorate with increasing higher frequency content of the images. The CCS and $\mathrm{CCS}_{\mathrm{aff}}$ behave inversely, i.e. they perform better when the higher frequencies are present. This holds for the disparity gradients too. The $\mathrm{C}_{\mathrm{interp}}$ deteriorates for both very high and very low frequencies. The reason for such

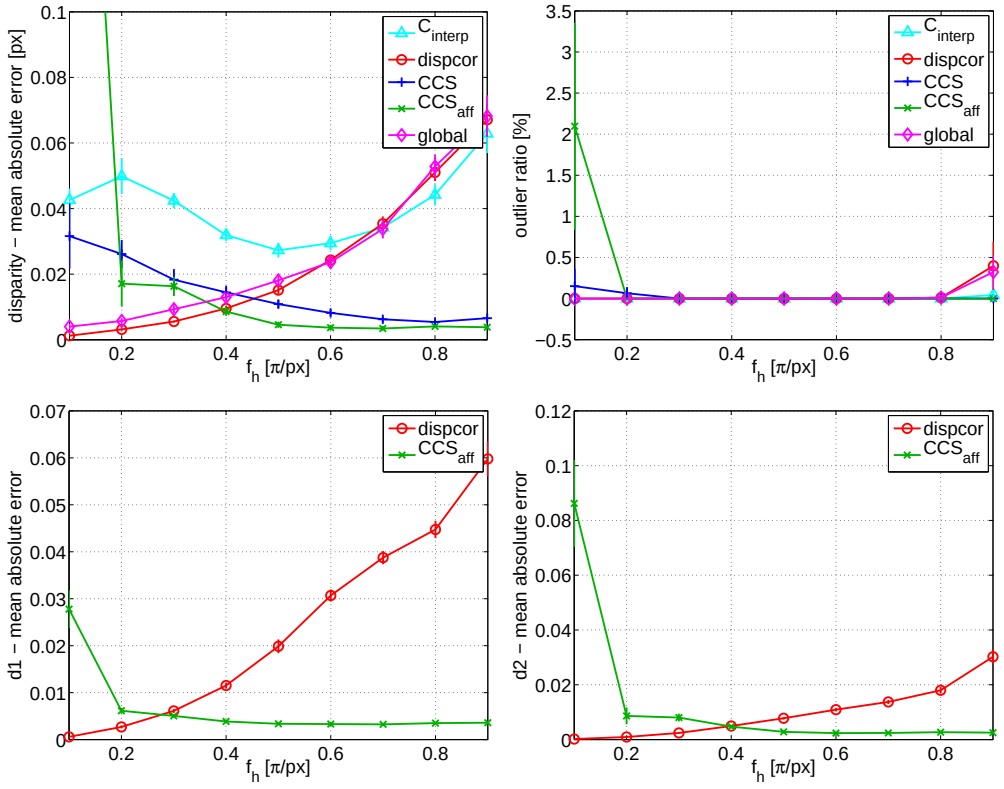


Figure 4.16: Accuracy: Limiting the upper frequency of the image signals, fronto-parallel plane ($d_0 = 0.5$).

behaviour is that image interpolation/optimization-based methods (`dispcor` and global) have difficulties interpolating higher-frequency signals. Interpolating a low frequency signal is more accurate and the optimization problems tend to be closer to unimodal and correct. The correlation-interpolation C_{interp} is corrupted the same way for high frequencies. On the other hand, low frequencies probably cause that neighbouring integer MNCC statistics have very similar values and the sub-pixel parabola vertex becomes badly conditioned. Very low frequency images represent a problem for CCS and CCS_{aff} methods. The reason is that very few filters give non-zero responses, which causes the statistic be aggregated from just a few independent estimates. The situation is worse for CCS_{aff}, where the filters in the frequency domain are steeper, see Fig. 4.4, which causes the CCS_{aff} deteriorate abruptly for the lowest upper frequency. The inverse character of the accuracy of CCS-based and image interpolation/optimization-based approaches is interesting, however a low frequency spectra is rather uncommon, since real images contain also some degree of higher frequency noise which disturbs the weak low-frequency components. Moreover, low frequency content makes the (preceding) matching problem difficult due to low discriminability.

4.5.2 Real data

We will report three experiments where the images were captured by real cameras: (1) a precise laboratory test, (2) matching experiments on standard Middlebury stereo images [136] and some outdoor scenes, and (3) an experiments demonstrating the sub-pixel accuracy of the presented methods.

In the first two experiments we compare the performance of the following algorithms: MNCC-dc, CCS, and CCS_{aff}. Algorithm MNCC-dc means that the integer matching is found from MNCC table by Confidently Stable Matching algorithm[5] [129]. Afterwards, a sub-pixel refined disparity is found by the affine window fitting `dispcor`. In CCS and CCS_{aff} algorithms, the integer matching is found from the rearranged magnitude table of the respective statistics by the same matching algorithm [129], while the sub-pixel disparities are taken from their phases, as described in Sec. 4.2.5.

Laboratory test scene The algorithm performance was tested on the CMP dataset [82]. The scene consists of thin stripes in front of a planar background, see Fig. 4.19 (first row). The algorithms are evaluated on an image series of varying texture contrast as described in [82] in detail.

We observed the Mismatch Ratio MIR (matches which differ more than one pixel from the ground-truth), the False Negative Ratio FNR (the sparsity of matching) and the accuracy as the root means square (RMS) error. See [82] for a complete definition of the error statistics. Results are shown in Fig. 4.17. We can see the performance of CCS is the best for all texture contrasts in all observed quality measurements. In a comparison with MNCC-dc, the CCS has about twice less mismatches, the FNR is lower

[5]The principle of this matching algorithm is described in Chapter 5.

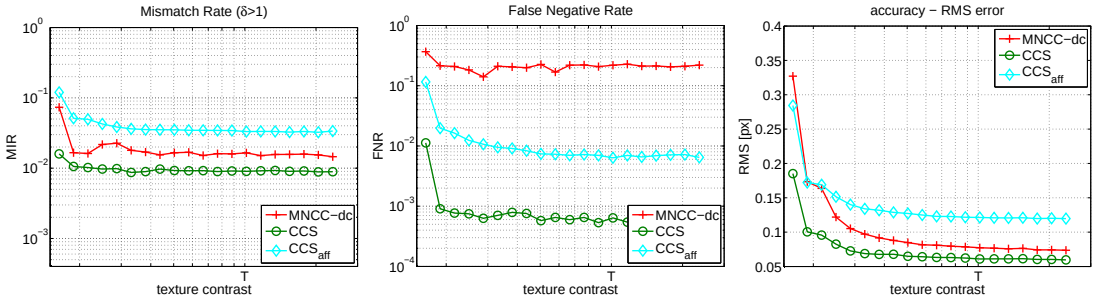


Figure 4.17: Laboratory test scene: matching error evaluation plots.

by the order of 2 magnitudes (!), and in the sub-pixel accuracy, it has from 1.5 times to twice (for weak textures of low contrast) lower RMS error. The CCS_{aff} is not so good. Compared to MNCC-dc, the CCS_{aff} is only better by a magnitude in FNR, while worse in MIR and in accuracy.

The reason for CCS_{aff} failure on this scene is most probably too large spacial extent of the filters violating the planar assumptions close to the occlusion boundaries. This effect overweighs the benefits of its affine insensitivity, the scene is slightly horizontally slanted, but $d_1 \approx 0.06$ only.

A significant improvement occurs in FNR for CCS-based algorithms, in case of CCS it is by the order of 2 magnitudes. The reason is that for a standard windowed correlation, besides the loss of discriminability where the disparity is around 0.5 pixel offset, there are gaps in the disparity map where the disparity changes, see the unassigned iso-disparity contours in the disparity maps in Fig. 4.19. The matching algorithm cannot decide which of the two competitive matches having approximately the same correlation to select, and according to the stability principle [129] it decides to reject both of these matches. While in the case of CCS, this undecidable situation does not occur, since all such competitive matches point towards the same match between pixels.

The disparity maps for the best texture contrast are shown in Fig. 4.19 (first row). The whole evaluation method is described in [82], where additional matching error statistics are defined and results for other algorithms are reported.

Standard and outdoor scenes We show results on the the standard Middlebury test scenes and results on outdoor scenes.

In both Middlebury scenes in Fig. 4.18 and the scenes in Fig. 4.19 we can see, that generally CCS-based methods gives denser results. The undecided iso-disparity contours in MNCC-dc are not present in CCS-based methods. The CCS has slightly less mismatches compared to MNCC-dc, see the mismatches at half-occluded region in Map scene and in the closest plane in Venus scene. For both methods, the artifacts around occlusion boundaries are comparable in all scenes, since the size of MNCC window is $3\sigma + 1$ pixels, where σ is the Gabor filter scale used in CCS. The CCS_{aff} produces more mismatches, and the occlusion artifacts are stronger than MNCC-dc and CCS. The reason is the large

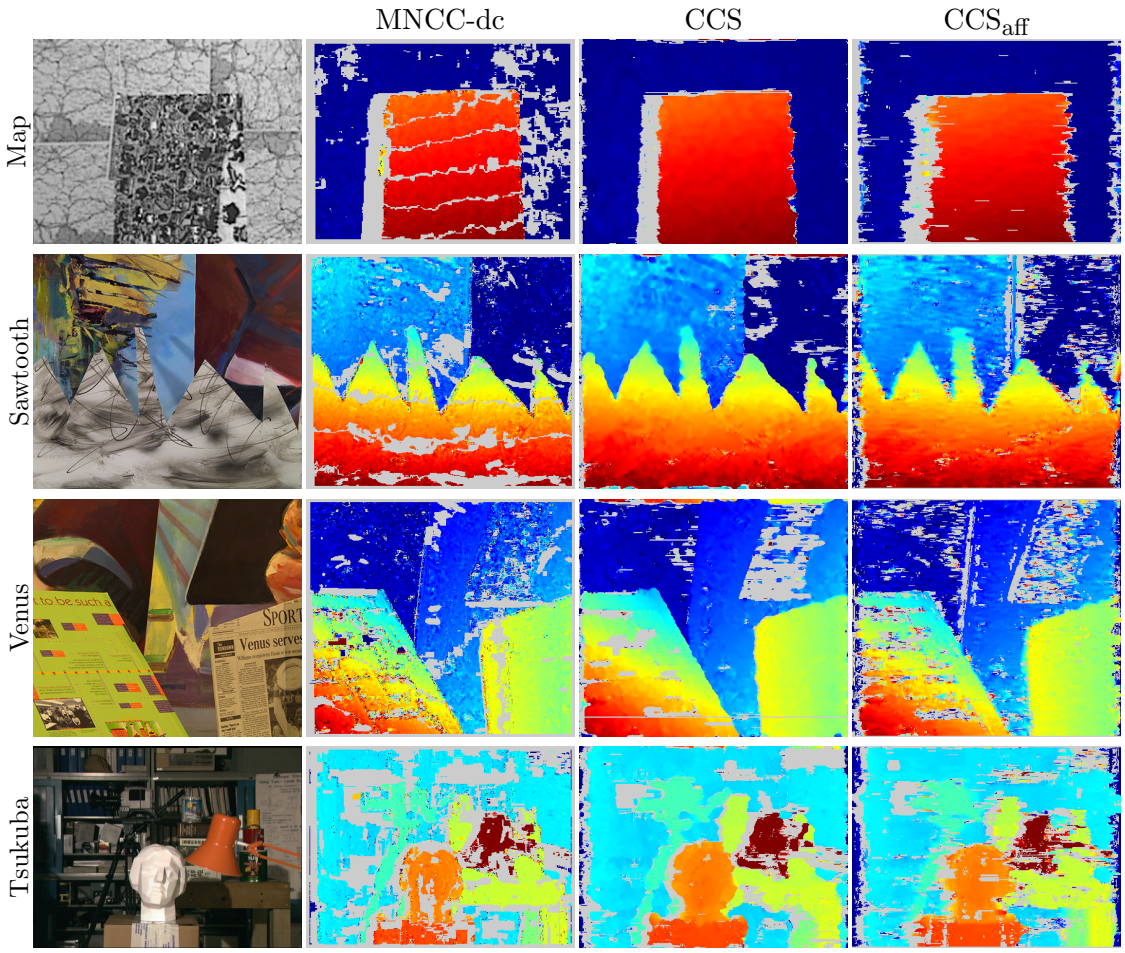


Figure 4.18: Middlebury scenes: left images and disparity maps. Disparity maps are color-coded, warmer colors represent larger disparities, gray color stands for unassigned disparity. Names of the scene are listed in the left.

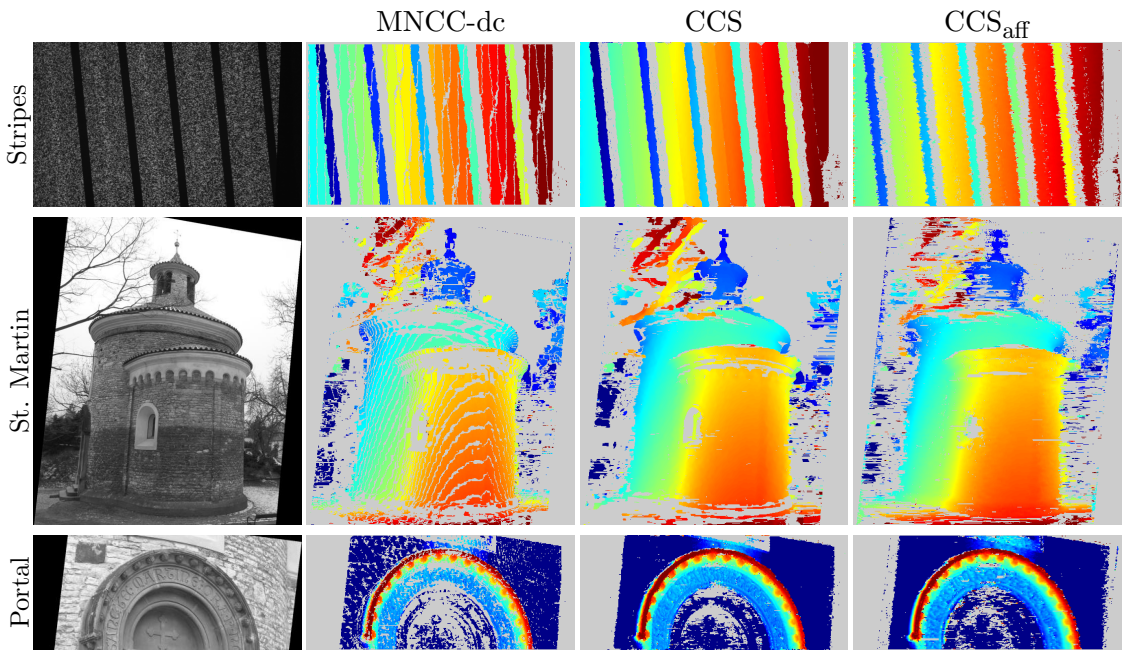


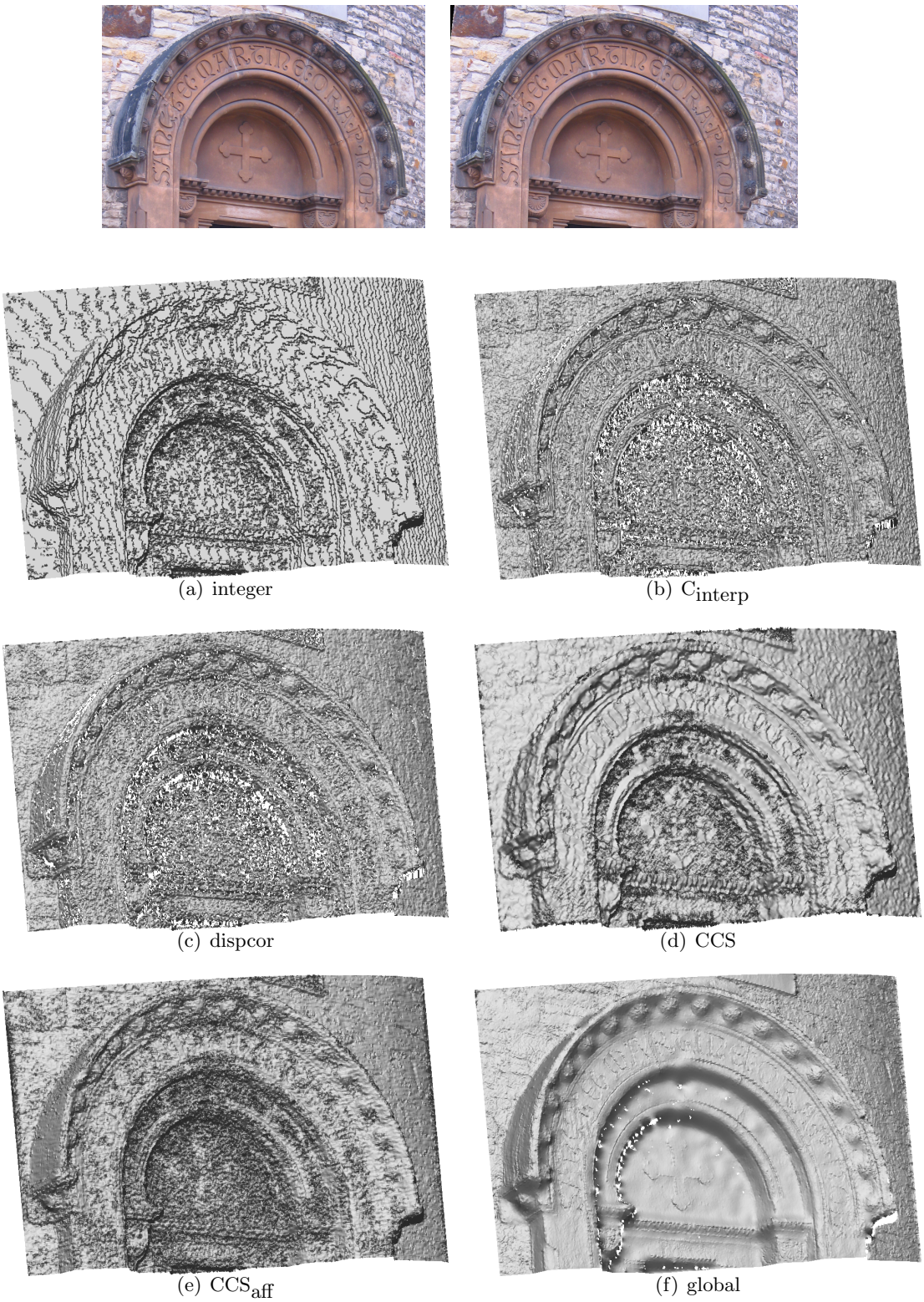
Figure 4.19: Real scenes: left images and disparity maps.

spatial extent of filters used in CCS$_{aff}$ in horizontal direction, in the vertical direction the extent is small (3 pixels only), which makes the matching support region anisotropic and produces streaking artifacts near occlusions boundaries, e.g. Map scene.

The influence of the shape of the matching support region is best viewed in St. Martin scene, around the cross on top of the rotunda. Square support in MNCC-dc, smooth circular support in CCS (non-zero entries in the Gaussian envelope), and elliptical support smooth in horizontal and crisp in vertical direction in CCS$_{aff}$.

The Middlebury scenes are rather piecewise planar, while St. Martin scene with non-planar surfaces, half occluded regions and thin objects is more complex. Note an inscription above the door in Portal scene. Importantly, the letter-to-background difference in disparity is less than one pixel. The inscription is clearly visible in the CCS disparity map (also in CCS$_{aff}$), while not in the MNCC-dc map, which is very noised.

Sub-pixel disparity estimates on real scenes In the previous real experiments, we tested discriminability of the statistics rather than the sub-pixel accuracy. The goal of the present experiment is to compare the presented sub-pixel disparity estimates independently on the integer matching. Therefore, the integer disparity map is common to all methods: C$_{interp}$, `dispcor`, CCS, CCS$_{aff}$ and the global method. The integer disparity map is obtained using a Growing Correspondence Seeds matching algorithm described in Chapter 5, followed by a post-processing that fills small gaps in the disparity map by a simple algorithm which assigns a median disparity of 5×5 pixel window around

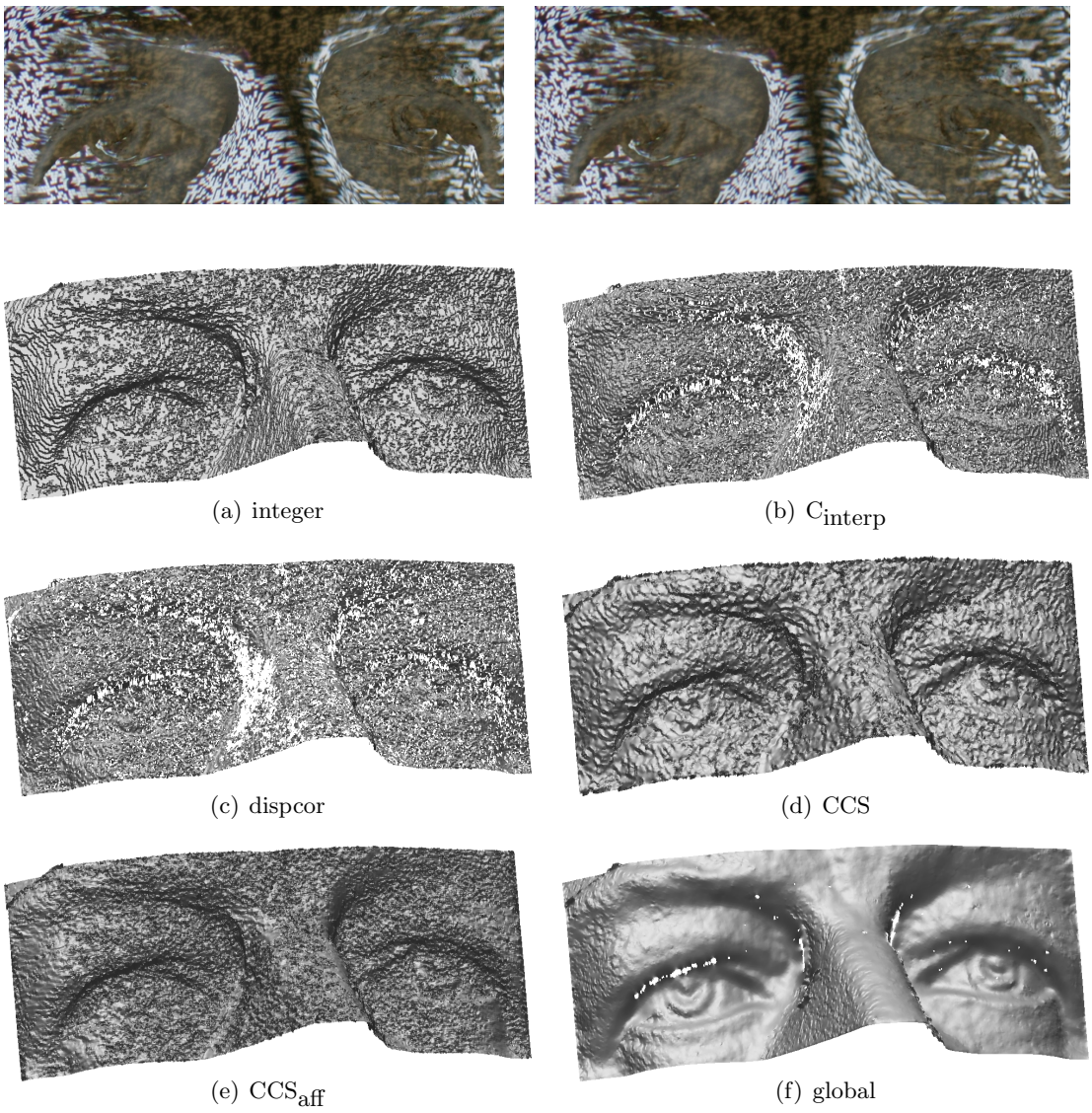


(a) integer (b) C_{interp}

(c) dispcor (d) CCS

(e) CCS_{aff} (f) global

Figure 4.20: Portal. Input images and reconstruction results as relighted 3D models.

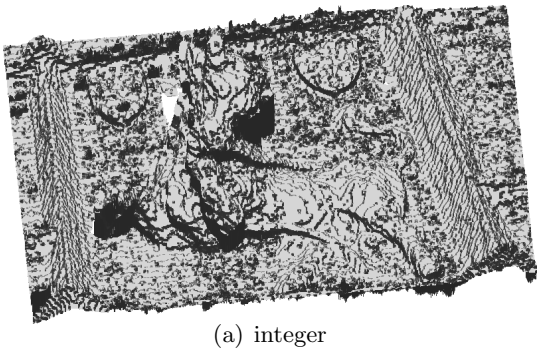


(a) integer (b) C_{interp}

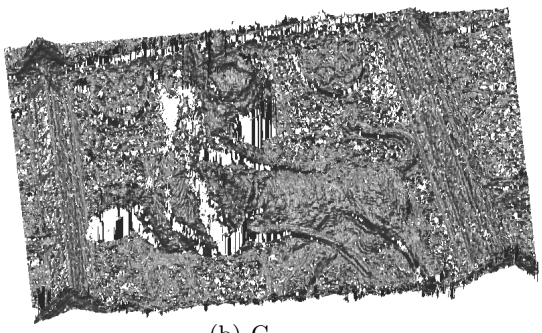
(c) dispcor (d) CCS

(e) CCS_{aff} (f) global

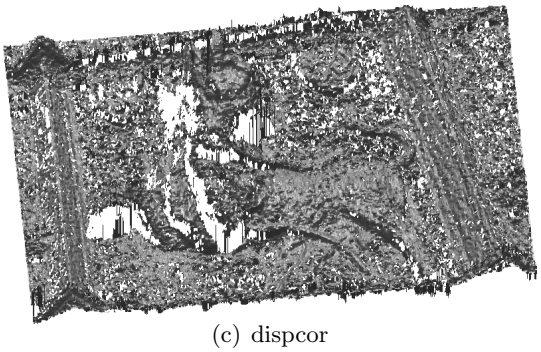
Figure 4.21: Wilson. Input images and reconstruction results as relighted 3D models. The plaster model (c) Albin Polasek Foundation.



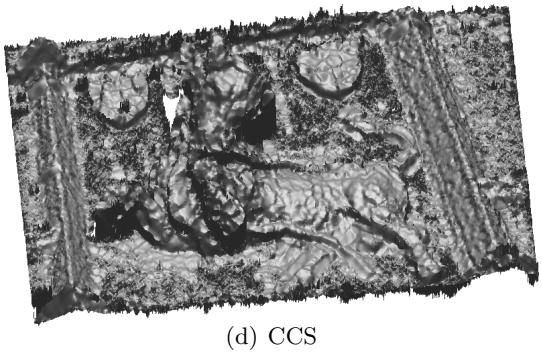
(a) integer



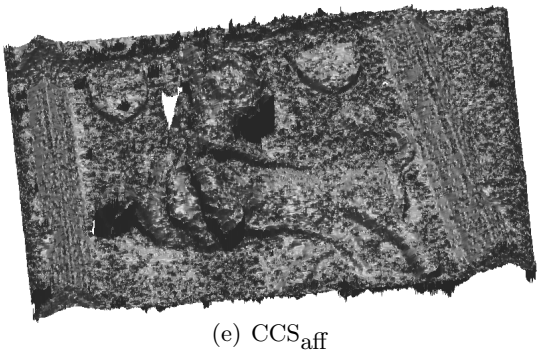
(b) C_{interp}



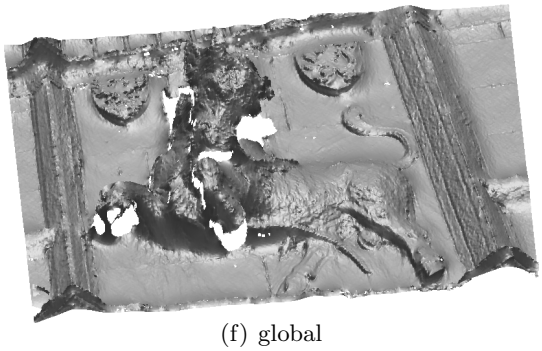
(c) dispcor



(d) CCS



(e) CCS_{aff}



(f) global

Figure 4.22: Bull. Input images and reconstruction results as relighted 3D models.

Figure 4.23: A result of the recent PDE-based method by Zimmer et al. on the portal scene; a reproduction from [185].

the gap in order the surface is visually more compact.

We show results as relighted 3D models for three scenes: Portal in Fig. 4.20, Wilson[6] in Fig. 4.21, and Bull in Fig. 4.22. The reconstruction results are shown for the integer matching with no sub-pixel disparity correction and all the methods tested in the ground-truth synthetic experiment in Sec. 4.5.1.

Generally, we can see that clearly the best results visually are produced by the global method. The surface is smooth without disturbances, nevertheless small details are well captured, e.g. the inscription, rosettes, stone bricks in Portal scene; the eyes in Wilson scene; the heraldic symbols or the relief in Bull scene. Moreover the global method is the only method which can cope with the incorrect integer matching. This can be seen in weakly textured regions of Portal scene, see the integer matching in Fig. 4.20(a).

On the other hand, `dispcor` or C_{interp} result in similar and very noisy surfaces. Additionally C_{interp} suffers from artifacts that the sub-pixel disparity is biased towards integer values and the integer contours are still recognizable.

The CCS outputs locally sleeker results than `dispcor`, however the surface details are fattened and not that precise as in the global method. The CCS_{aff} behaves similarly, although some details are retained, the surface looks more disturbed. Notice the CCS_{aff} performs better than CCS in slanted planar segments of the surface, e.g. the lower left part of the arch in Portal scene.

In Fig. 4.23, we show a reconstruction result of the recent PDE-based method by Zimmer et al. [185] on the Portal scene. The authors used our images. These results

[6]The stereo-pair is taken from a plaster model of a head of Woodrow Wilson's statue. A random texture was projected to the plaster model to ease the stereo matching. The statue was standing near to main train station in Prague since 1928 until 1941 when it was destroyed by Nazis during the occupation. American Friends of the Czech Republic, a Washington, DC based non-profit organization, has undertaken the project to reestablish the monument to American President. Center for Machine Perception was asked to deliver a virtual 3D model of the statue from the plaster model of the head and historical photographs.

do not achieve visual accuracy of results of the global sub-pixel disparity correction, see Fig. 4.20(f). In a way, they are comparable to all other sub-pixel methods we show. In their results, there are no mismatches, however the quality of surface details is not perfect.

4.6 Discussion and conclusions

In this chapter, we have introduced a new Complex Correlation Statistic, which possesses both invariance to sampling and ability to estimate the position of the maximum. In theory, using the CCS, we obtain sampled continuous solution from a discrete algorithm.

We proposed two versions of Complex Correlation Statistic. One is insensitive to image sampling, assuming locally constant disparity (CCS), while the other assuming locally linear disparity is additionally insensitive to an affine distortion (stretch and skew) of a local image neighbourhood which occurs in rectified stereo images ($\mathrm{CCS}_{\mathrm{aff}}$).

The experiments show that CCS is usable as a good alternative to a window MNCC statistics, benefiting from its sampling invariance and direct sub-pixel disparity estimate. On the other hand, the experiments did not confirm a practical usefulness of the affine version $\mathrm{CCS}_{\mathrm{aff}}$.

We performed several experiments in real images, however the advantage of affine insensitivity did not manifest significantly. It probably means that narrow baseline stereo images are matchable with a simpler model. An alternative approach to invariant/insensitive correlation statistic for dense matching is a local image rectification (change of coordinates) according to local affine frames inherited from matching discriminative features used in the preceding wide baseline stereo to find the epipolar geometry. This strategy is applied in Chapter 6.

The problem of $\mathrm{CCS}_{\mathrm{aff}}$ is the large spatial extent which results in a higher sensitivity to the model validity (non-linearity) and consequently, it leads to problems in both discriminability and sub-pixel accuracy.

The spatial extent of these statistics is a fundamental problem and cannot be avoided. The filters cannot be smaller, since the smaller extent in the spatial domain implies a larger extent in a frequency domain. This makes the filters more dependent decreasing the discriminability and it is harmful for the local disparity estimates too. The loss of discriminability with decreasing spatial extent (window size) occurs in window intensity statistics as MNCC also. There is a natural tradeoff between the discriminability/accuracy and a level of artifacts due to model violation.

As a complementary method for sub-pixel disparity estimate, we proposed the global method which solves the task as a single optimization instead of many independent problems.

A surprising observation from our study is that, despite the global method for sub-pixel disparity estimation does not perform excellently in synthetic ground-truth experiments compared to other methods, the real scenes show the best results visually. Small variations around the true surface are probably much more disturbing visually than smooth

surface which is not always in a correct position.

The computational complexity of CCS statistics is higher than the complexity of standard windowed intensity statistics. Besides the complex arithmetics, we cannot use the 'sliding window' algorithm, which exploits precomputed partial sums [24]. Nevertheless there are no iterations and optimizations within, we have a closed-form formula for the CCS and very simple algorithm for $\mathrm{CCS}_{\mathrm{aff}}$. So, using a special (but simple) parallel hardware, we can even have a real time implementation.

The global method for sub-pixel disparity estimation can be also speeded-up significantly using an implementation in a modern GPU, since the bottleneck here is the image warping in the criterion computation.

5 Efficient sampling of disparity space

In this chapter, a simple stereo matching algorithm is proposed that visits only a small fraction of disparity space in order to find a semi-dense disparity map. It works by growing from a small set of correspondence seeds. Unlike in known seed-growing algorithms, it guarantees matching accuracy and correctness, even in the presence of repetitive patterns. The proposed algorithm is able to work with complex scenes with a rich 3D structure of a great depth and many complicated occlusions, see Fig. 5.1. This success is based on the fact it solves a kind of a global optimization task. The algorithm can recover from wrong initial seeds to the extent they can even be random. The quality of correspondence seeds influences computing time, not the quality of the final disparity map. We show that the proposed algorithm achieves similar results as an exhaustive disparity space search but it is two orders of magnitude faster. This is very unlike the existing growing algorithms which are fast but erroneous. Accurate matching on 2-megapixel images of complex scenes is routinely obtained in a few seconds on a common PC from a small number of seeds, without limiting the disparity search range.

5.1 Introduction

Traditional area-based dense stereoscopic matching algorithms perform exhaustive search of the entire disparity space, i.e. they need to compute a correlation statistic for all putative correspondences [136]. Although efficient implementations for computing most of the commonly used statistics (SSD, SAD, NCC) are known, by using a 'sliding window' [164, 22], this is still one of the most expensive phases of stereo matching.

To avoid visiting the entire disparity space, algorithms were proposed that greedily grow corresponding patches from a given set of reliable seed correspondences. Such algorithms assume that neighboring pixels have similar disparity, not exceeding disparity gradient limit [123] or a similar constraint.

The principle of growing a solution from initial seeds had long been known in segmentation [59]. The first algorithms using this principle in stereo were proposed in photogrammetric community: by Otto and Chau [118], O'Neill and Denos [117], and by Kim and Muller [74].

Later, Lhuillier and Quan [93] employed the uniqueness constraint and proposed an algorithm both with and without the epipolar constraint in which images play a sym-

Figure 5.1: Complex scenes and matching results as disparity maps obtained by the proposed GCS-2 algorithm.

metric role. This work has later been used in a 3D surface reconstruction pipeline [94]. The basic idea is to grow contiguous *components*[1] in disparity space from initial correspondence seeds sorted in decreasing value of image similarity and to stop the growth process at image pixels where uniqueness constraint would be violated. The growth occurs in the neighborhood of previous matches in disparity space. This creates new seeds that are put to the priority queue. A decision on match acceptance is never revised. This inherent greediness of the algorithm may cause a complete failure in the presence of repetitive texture in the scene, as will be discussed later.

Independently, Chen and Medioni [27] proposed an alternative scheme in which the growth is not constrained by uniqueness. The best-first strategy is not used and the seeds are taken in arbitrary order. When a match of better image similarity is found at a given pixel, it overrides the previous match but it does not grow further, hence the correction is only local and there is no ability to follow a new disparity component of high image similarity. If the seeds in the queue are processed in a different order, (very) different results are obtained. Moreover, images in this algorithm do not have a symmetric role, which means the resulting matching violates uniqueness constraint.

The disadvantage of both approaches is that the decision on a match is local in the sense that other matches do not influence it (no global optimization is involved).

The following two works use variations of the above methods which means they suffer from the same drawbacks, especially in the presence of repeated structures. These methods are interesting not because of the baseline growth mechanism but because of improvements in other respects.

The first, by Zeng et al. [179, 180] uses the best-first strategy in a multi-image algorithm that replaces the small pixelwise growth increments by an optimal choice of a whole surface patch extending the currently found 3D segment. Final selection is done by marching cube tracing in 3D which may not be able to recover from the earlier stage errors, especially in complex scenes.

The second, by Megyesi et al. [102] uses the Otto and Chau's algorithm with adaptive affine deformation of the domain of image similarity statistic, where the affine parameters are estimated from surface normals which are propagated within the growth process.

In stereo literature, also *progressive algorithms* have been proposed [183, 168, 53]. Earlier matches guide the subsequent matching by postponing ambiguous decision until enough confidence is accumulated to resolve the ambiguity, see Sec. 2.1.3. Although this may look like a growth from seeds, it does not explicitly use spatial coherence: the growth does not necessarily occur in the immediate neighborhood of previously accepted matches. These algorithms never revise the decision on match acceptance, as well. They need to visit a large fraction of disparity space (satisfying a set of constraints) before making the first decision. This is unlike the true growing algorithms (whose decision is *online*).

A problem common to all known growing algorithms is their reliance on high-quality seeds. In dense stereo, this means there must be at least one seed in each true disparity

[1] The term *disparity component* has been coined in [16].

component, which is very hard to fulfill. To our surprise, it turned out that there is a suitable algorithm that has the ability to recover from errors, which means that good-quality initial seeds are not needed: In fact, even quite complex scenes can be matched from a few *random* seed correspondences, as will be discussed in Sec. 5.3.

To overcome drawbacks of the discussed methods, we temporarily forego uniqueness constraint and propose an algorithm which, under a theoretically well-grounded rule, keeps growing disparity components regardless of their overlap in disparity space. The rule facilitates the efficiency of the growth process by stopping it at loci where the final, optimal result cannot be improved. As a result, it is not necessary to visit a large fraction of disparity space to obtain optimal solution. We then solve a global optimality task by a fast robust matching algorithm that selects among the competing components in disparity space. This leads to a significant improvement in the quality of the result at only a small computational cost and it is equivalent to the ability to revise decisions made at the growth phase.

The chapter is structured as follows: Sec. 5.2 formulates the dense stereoscopic matching task as a global discrete optimization problem and describes an efficient disparity component growth algorithm. Experimental validation comparing the proposed algorithm with a baseline algorithm is done in Sec. 5.3, where we discuss efficiency and the ability to deal with repetitive image structures correctly and the ability to find a large number of disparity components from a small set of seeds, even if they are random. Sec. 5.4 concludes the chapter and hints some extensions of the work.

This chapter is a revised and extended version of paper [26].

5.2 Matching algorithm

To help understanding the proposed algorithm, we make a thought decomposition of the matching task to two phases: (1) an unconstrained growth of disparity components from an initial set of seeds and (2) an optimal matching working with the set of components found in the first step. We will then show that some of the work of the second phase can already be done during the component growth without loosing optimality of the algorithm, while significantly speeding up the first phase.

To simplify the description of the algorithm, we assume a pair of horizontally rectified stereo images is used. Generalization to unrectified images is possible but it will not be discussed here. We therefore assume we are working with matching table. It represents a 3D discretized disparity space in which each element (x, x', y) denotes a possible correspondence $(x, y) \leftrightarrow (x', y)$, see Fig. 5.2. Each matching table element (x, x', y) may be associated with some parameters θ modeling relative distortion of image neighborhoods. The parameters θ may be updated during the growth process to accommodate to the slant of the 3D surface, as in [102, 118, 117, 74]. This is important in wide-baseline stereo. For simplicity, we omit the distortion model here.

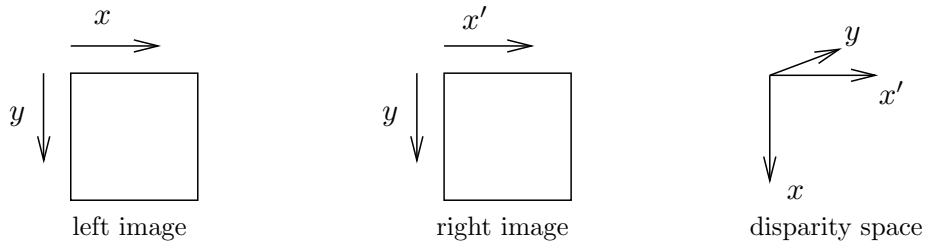


Figure 5.2: Coordinate system used. The y is a common row coordinate in rectified images. The x, x' are the column coordinates in the left and right image, respectively.

5.2.1 Disparity component growth

Suppose we are given an unsorted list of disparity seeds $\mathcal{S}$. Each seed is a point in disparity space, $\mathbf{s} = (x, x', y)$. Its neighborhood $\mathcal{N}(\mathbf{s})$ in disparity space consists of 16 points constructed from four sub-sets $\mathcal{N}_1(\mathbf{s}) \cup \mathcal{N}_2(\mathbf{s}) \cup \mathcal{N}_3(\mathbf{s}) \cup \mathcal{N}_4(\mathbf{s})$, see Fig. 5.3(a) where we use the same colors for:

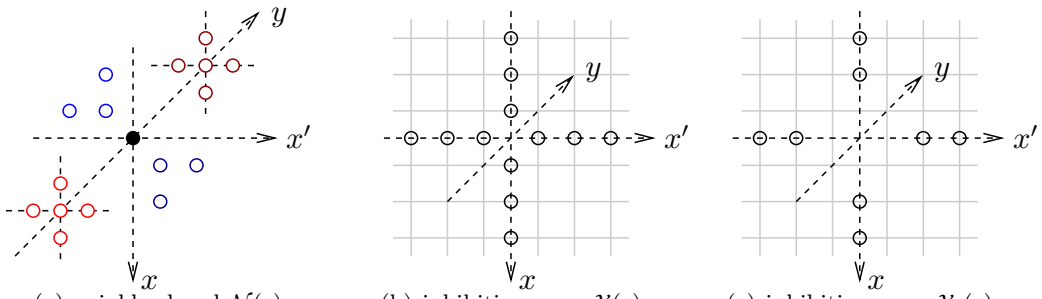
$$\begin{aligned}
\mathcal{N}_1(\mathbf{s}) &= \big\{(x-1,\, x'-1,\, y),\, (x-2,\, x'-1,\, y), (x-1,\, x'-2,\, y)\big\},\\
\mathcal{N}_2(\mathbf{s}) &= \big\{(x+1,\, x'+1,\, y),\, (x+2,\, x'+1,\, y), (x+1,\, x'+2,\, y)\big\},\\
\mathcal{N}_3(\mathbf{s}) &= \big\{(x,\, x',\, y-1),\, (x\pm 1,\, x',\, y-1), (x,\, x'\pm 1,\, y-1)\big\},\\
\mathcal{N}_4(\mathbf{s}) &= \big\{(x,\, x',\, y+1),\, (x\pm 1,\, x',\, y+1), (x,\, x'\pm 1,\, y+1)\big\}.
\end{aligned} \tag{5.1}$$

The neighborhood is selected so as to limit the magnitude of disparity gradient to unity and to improve the ability to follow a disparity component even if the image similarity peak falls in between pixels in the matching table. This improves performance in both the baseline and the proposed algorithms (see below).

Assuming similarity is computed from small image windows around pixels (u, v) and (u', v) by e.g. the normalized cross-correlation, we prepare an empty matching table $\mathcal{T}$ and start growing disparity components by drawing an arbitrary seed $\mathbf{s}$ from $\mathcal{S}$, adding it to $\mathcal{T}$, individually selecting the best-similarity neighbors $\mathbf{q}_i$ over its four sub-neighborhoods $\mathcal{N}_i(\mathbf{s})$:

$$\mathbf{q}_i = (u, u', v)_i = \operatorname*{argmax}_{(x,x',y)\in\mathcal{N}_i(\mathbf{s})} \operatorname{corr}(x, x', y), \qquad i = 1, 2, 3, 4,$$

and putting these neighbors $\mathbf{q}_i$ to the seed list if their inter-image similarity exceeds a threshold τ. Hence, up to four new seeds are created. If we draw a seed from the list $\mathcal{S}$ that is already a member of the matching table, then we discard it. The growth must stop in a finite number of steps by exhausting the list $\mathcal{S}$. The output from the growth phase is a partially filled matching table whose connected regions in 3D represent disparity components grown around the initial seeds. Note that disparity components obtained this way are nothing more than contiguous segments in disparity space. In the extreme case when $\tau = -\infty$ the entire disparity space is filled by a single component grown from the first seed.



(a) neighborhood $\mathcal{N}(\mathbf{s})$ (b) inhibition zone $\mathcal{X}(\mathbf{s})$ (c) inhibition zone $\mathcal{X}_g(\mathbf{s})$

Figure 5.3: Disparity space neighborhood used in this chapter (a). The X-inhibition zone (b) and modified X-inhibition zone with a gap (c) for $\mathbf{s} = (x, x', y)$.

Obviously, such growth is not a very efficient way of selecting high-similarity tentative matches. Instead, we see it only as an elementary mechanism for traveling in disparity space. This phase has been introduced just to show correctness of the entire proposed algorithm which will be described shortly. Note, finally, that the order of selecting the seeds from the list $\mathcal{S}$ is arbitrary, so far.

We refer to the above procedure as *unconstrained growth*. Its output is not a matching. This is the reason why various authors proposed a modification which stops the growth whenever a violation of uniqueness constraint is detected. This already requires drawing the seeds in the order of their image similarity. The result of this modification is shown in pseudo-code[2] as Alg. 1. We call it the *baseline growth algorithm* in this chapter. It is very close to Lhuiller's algorithm [93]. Note that both images have a symmetric role. We will show in Sec. 5.3 that Alg. 1 is fast but its performance is not very good. The reason for its bad performance is that it does not solve any reasonable optimization task. In the next paragraph, we show how to incorporate a formal discrete optimization task into the algorithm. It will turn out that the necessary changes are small and still the performance improves dramatically (at the cost of a small increase in computational complexity).

5.2.2 Matching

Let $\mathbf{s} = (u, u', v)$ be an element of the matching table $\mathcal{T}$ obtained from the unconstrained growth procedure. Let $(:, u', v)$ represent all elements (w, u', v) such that $w \neq u$ and the colon in $(u, :, v)$ has a similar meaning. The set of all $(:, u', v)$ together with all $(u, :, v)$ will be called the X-inhibition zone of $\mathbf{s}$ and denoted as $\mathcal{X}(\mathbf{s})$, see Fig. 5.3(b). Note that $\mathbf{s} \notin \mathcal{X}(\mathbf{s})$ and that $\mathbf{q} \in \mathcal{X}(\mathbf{s}) \Leftrightarrow \mathbf{s} \in \mathcal{X}(\mathbf{q})$. The relation $\mathbf{q} \in \mathcal{X}(\mathbf{s})$ represents the relation of occlusion [130]. We say element $\mathbf{q} \in \mathcal{T}$ is a *competitor* to $\mathbf{s} \in \mathcal{T}$ if $\mathbf{q} \in \mathcal{X}(\mathbf{s})$ and it has better image similarity, i.e. $\text{corr}(\mathbf{q}) \geq \text{corr}(\mathbf{s})$. Element $\mathbf{q}$ is a *strict competitor* to element $\mathbf{s}$ if $\mathbf{q} \in \mathcal{X}(\mathbf{s})$ and $\text{corr}(\mathbf{q}) > \text{corr}(\mathbf{s}) + \mu$, where μ is called the *stability margin*.

[2]The X-inhibition zone $\mathcal{X}(\mathbf{q}_i)$ in Step 1.7 is defined early in Sec. 5.2.2.

Algorithm 1 The Baseline Growing Algorithm

Require: Rectified images $\mathbf{I}_l$, $\mathbf{I}_r$, initial correspondence seeds $\mathcal{S}$, image similarity threshold τ.

1.1: Compute similarity $\mathrm{corr}(\mathbf{s})$ for every seed $\mathbf{s} \in \mathcal{S}$.
1.2: Initialize empty matching table $\mathcal{T} := \emptyset$.
1.3: **repeat**
1.4: Draw the seed $\mathbf{s} \in \mathcal{S}$ of the best similarity $\mathrm{corr}(\mathbf{s})$.
1.5: **for** each of the four best neighbors

$$\mathbf{q}_i = (u, u', v) = \underset{\mathbf{t} \in \mathcal{N}_i(\mathbf{s})}{\mathrm{argmax}}\, \mathrm{corr}(\mathbf{t}),\ i \in \{1, 2, 3, 4\}$$

do
1.6: $c := \mathrm{corr}(\mathbf{q}_i)$.
1.7: **if** $c \geq \tau$ **and** $\mathcal{X}(\mathbf{q}_i) \notin \mathcal{T}$ **then**
1.8: Update the matching table $\mathcal{T} := \mathcal{T} \cup \{\mathbf{q}_i\}$ and
1.9: the seed queue $\mathcal{S} := \mathcal{S} \cup \{\mathbf{q}_i\}$.
1.10: **end if**
1.11: **end for**
1.12: **until** $\mathcal{S}$ is empty.
1.13: **return** matching as a partially filled matching table $\mathcal{T}$.

The matching task is done by solving a graph-theoretic problem known as the *maximum strict sub-kernel* (SSK) [130, 129]. The basic SSK algorithm, which can be used for our problem class and which we call the *dominant element reduction algorithm* works as follows. Given matching table $\mathcal{T}$, partially filled with disparity components from the unconstrained growth phase, one finds a *dominant element* $\mathbf{s}$ of the table whose image similarity satisfies

$$\mathrm{corr}(\mathbf{s}) > \max_{\mathbf{t} \in \mathcal{X}(\mathbf{s})} \mathrm{corr}(\mathbf{t}) + \mu, \tag{5.2}$$

where μ is the stability margin. If there was no dominant element, the algorithm stops. Next, given the dominant element $\mathbf{s}$, one removes all elements $\mathcal{X}(\mathbf{s})$ from $\mathcal{T}$. In the reduced table, new dominant element is found and the reduction process is repeated. The procedure finishes in a finite number of steps. The fully reduced table $\mathcal{T}$ contains a set that is the largest (i.e. maximum) one-to-one matching that is a SSK. It can be proved that our problem always has at most one maximum SSK and that the above algorithm is able to find it or confirm there is none [130]. The defining quality of a SSK is its *strict stability*, which can be considered a global 'optimality' property in the following sense: Let $\mathcal{T}$ be the set of all elements in the initial matching table. Let $\mathcal{K}$ be a subset of $\mathcal{T}$. We say $\mathcal{K}$ is strictly stable if for every $\mathbf{p}, \mathbf{q} \in \mathcal{K}$ it holds that $\mathbf{p} \notin \mathcal{X}(\mathbf{q})$ (i.e. $\mathcal{K}$ is a one-to-one matching) and every element $\mathbf{s} \in \mathcal{K}$ is strictly stable with respect to $\mathcal{K}$. An element $\mathbf{s} \in \mathcal{K}$ is strictly stable wrt $\mathcal{K}$ if every competitor $\mathbf{q} \in \mathcal{T}$ to $\mathbf{s}$ has a strict competitor $\mathbf{t}$ in $\mathcal{K}$. Hence, SSK is the only set that is strictly stable. Note that a SSK can be incomplete. In the extreme case it can be empty if there was no dominant element. Incompleteness is a necessary prerequisite for *robustness*. See [130]

for a discussion of properties of SSK related to robustness.

The paper [130] formulates the SSK problem in rigorous graph-theoretical language and should be referred to for necessary conditions under which the algorithm is valid (for our problem it is valid).[3] The above algorithm directly performs *guiding* or the *least commitment strategy*, as discussed in [183]: most reliable decisions are made prior to unreliable ones that wait until the set of putative solutions becomes more constrained.[4] The accuracy of stereoscopic matching based on the SSK is studied in [82].

Besides the dominant element reduction algorithm, there is another algorithm for finding an SSK, which is more suitable for our matching problem and which produces an equivalent result [129, 130]. It has two phases: The first phase runs as in the dominant element reduction algorithm but with $\mu = 0$ and with the dominance test inequalities not sharp (in such case we are obtaining *weakly dominant elements*). We call its result a *weak kernel* (WK). It can be shown that maximum SSK is always a subset of a WK, irrespective of the order of processing the weakly dominant elements [130]. In the second phase, the WK is converted to a maximum SSK as follows:

Require: The output $\mathcal{T}$ from unconstrained growth,
the WK $\mathcal{K} \subseteq \mathcal{T}$.

1: **loop**
2: Find an element $\mathbf{q} \in \mathcal{T}$, $\mathbf{q} \notin \mathcal{K}$ such that

$$\mathcal{X}(\mathbf{q}) \cap \mathcal{K} \text{ is empty } \textbf{or}$$

$$\operatorname{corr}(\mathbf{q}) + \mu \geq \max_{\mathbf{t} \in \mathcal{X}(\mathbf{q}) \cap \mathcal{K}} \operatorname{corr}(\mathbf{t}).$$

3: **if** no such $\mathbf{q}$ was found **then**
4: **return** reduced WK $\mathcal{K}$
5: **else**
6: Remove $\mathbf{q}$ and $\mathcal{X}(\mathbf{q})$ from $\mathcal{T}$ and $\mathcal{K}$.
7: **end if**
8: **end loop**

The $\mathbf{q}$ in Step 2 are called the *converting elements*. Obviously, their neighbors in $\mathcal{X}(\mathbf{q}) \cap \mathcal{K}$ violate the strong stability condition. Correctness of this algorithm has been proved [130]. A fast algorithm for simultaneous finding WK and its progressive conversion to SSK, is described in [129].

The formulation given here clearly shows what to do in the disparity component growth procedure: Instead of stopping whenever the uniqueness constraint would be violated

[3] Strict sub-kernel is a general notion valid for oriented graphs, in which the graph structure represents the structure of the underlying problem (the structure of constraints) and the orientation represents evidence (data) [130]. The notion of SSK is related to the well-known Stable Marriage and Stable Roommates Problems [57].

[4] Note that the fact the above algorithm proceeds in this way is only due to the special structure of our matching problem. In more general graph orientations, the SSK algorithm is more complex and the general problem of finding a SSK is NP-complete [130]. It can be shown that when $\mu = 0$ the SSK approximates the max-sum independent vertex set problem within a factor of two (in our problem) [130].

Algorithm 2 The Proposed Growing Algorithm

Require: Rectified images $\mathbf{I}_l$, $\mathbf{I}_r$, initial correspondence seeds $\mathcal{S}$, image similarity threshold τ, and margin μ.

2.1: Compute image similarity for all seeds $\mathbf{s} \in \mathcal{S}$.

2.2: Initialize matching table $\mathcal{T} := \emptyset$, and auxiliary arrays $\mathbf{C}_{\text{best}}(:,:) := -\infty$, $\mathbf{C}'_{\text{best}}(:,:) := -\infty$.

2.3: **repeat**

2.4: Draw the seed $\mathbf{s} = (x, x', y) \in \mathcal{S}$ of the best image similarity $\text{corr}(\mathbf{s})$.

2.5: **for** each of the four best neighbors

$$\mathbf{q}_i = (u, u', v) = \underset{\mathbf{t} \in \mathcal{N}_i(\mathbf{s})}{\text{argmax}}\, \text{corr}(\mathbf{t}),\ i \in \{1, 2, 3, 4\}$$

do

2.6: $c := \text{corr}(\mathbf{q}_i)$.

2.7: **if** $c \geq \tau$ **and** $\mathbf{q}_i \notin \mathcal{T}$ **and** $c + \mu \geq \min\{\mathbf{C}_{\text{best}}(u, v),\, \mathbf{C}'_{\text{best}}(u', v)\}$ **then**

2.8: Update the matching table $\mathcal{T} := \mathcal{T} \cup \{\mathbf{q}_i\}$,

2.9: the seed queue $\mathcal{S} := \mathcal{S} \cup \{\mathbf{q}_i\}$,

2.10: and the best image similarities

$$\mathbf{C}_{\text{best}}(u, v) := \max\{c,\, \mathbf{C}_{\text{best}}(u, v)\},$$
$$\mathbf{C}'_{\text{best}}(u', v) := \max\{c,\, \mathbf{C}'_{\text{best}}(u', v)\}.$$

2.11: **end if**

2.12: **end for**

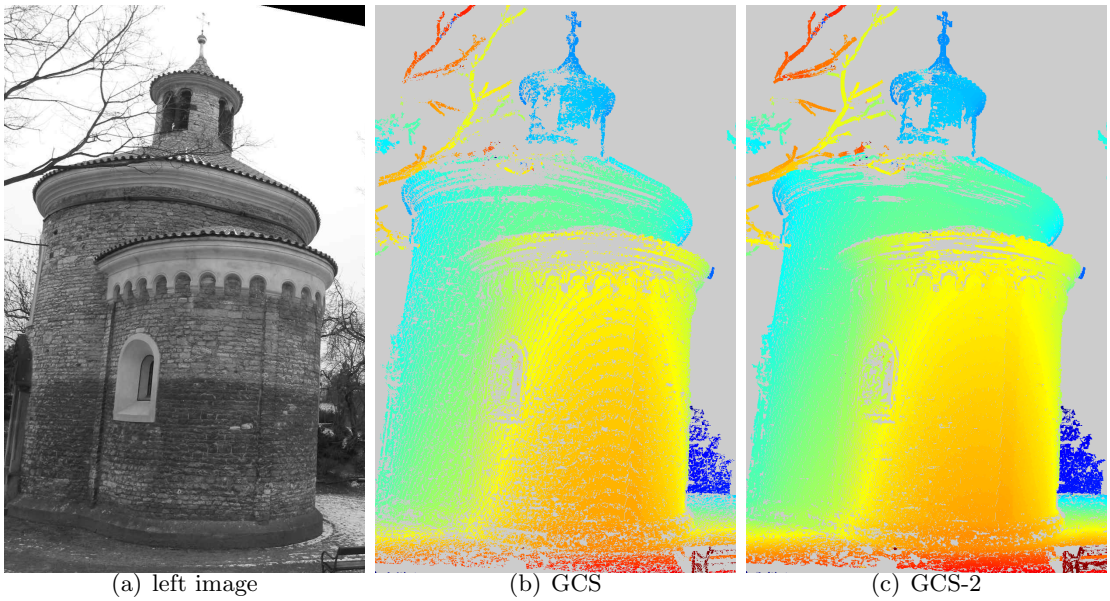
2.13: **until** $\mathcal{S}$ is empty.

2.14: **return** table $\mathcal{T}$ with traced-out disparity components.

as in Step 1.7 of Alg. 1, we have to continue the growth unless image similarities of the overlapping disparity components differ more than μ. Hereby, the components that cannot survive the final competition are not grown further. The resulting modification of Alg. 1 is shown in pseudo-code as Alg. 2. The difference is in Step 2.7 and in some inexpensive bookkeeping in Step 2.10. The difference is seemingly subtle but it has a strong impact on the ability to reject some of the bad matches, including some of the seeds, as will be shown in Sec. 5.3.

Of course, the partially filled matching table output from Alg. 2 is not yet a matching because of the overlap of disparity components in matching table $\mathcal{T}$ which violates the uniqueness constraint. The overgrown components undergo a competition in a subsequent selection process. To this end the output from Alg. 2 is processed by the dominant element reduction algorithm discussed above or by the equivalent WK conversion algorithm. We use an implementation that has been proposed for semi-dense stereo matching [129] and that does not require a fully populated matching table. This final matching is computationally very efficient because of the sparsity of $\mathcal{T}$. Our experiments show it takes less than 15% of the overall computing time.

The resulting composition comprising the growing algorithm Alg. 2 and the final matching competition, we call a Growing Correspondence Seed (GCS) algorithm.



(a) left image (b) GCS (c) GCS-2

Figure 5.4: Impact of the modified inhibition zone. Comparison between GCS and GCS-2 algorithms. The artifacts of unassigned disparity contours between disparity levels present in GCS are removed by a modified inhibition zone in GCS-2 algorithm.

5.2.3 Modification of the occlusion model

The occlusion model of the above matching algorithm is the uniqueness, see the inhibition zone $\mathcal{X}(\mathbf{s})$ in Fig. 5.3(b). However, when the true disparity is around modulo 0.5 pixel, there are two disparity hypotheses. The final robust matching algorithm tends to reject both of them, which causes contour-like artifacts of unassigned disparity between disparity levels, see Fig. 5.4(b). Such competing hypotheses have similar (rather low) correlation, and according to the stability principle, none of them become a part of the solution.

Therefore, we redefine the occlusion model by proposing a modification of the inhibition zone. New inhibition zone $\mathcal{X}_g(\mathbf{s})$ has a gap, see Fig. 5.3(c), which changes the mutual exclusivity in the set of correspondence hypotheses. Both candidates around modulo 0.5 pixel disparity are not competitors any more. The remaining problem is that the uniqueness does not hold then. To obtain one-to-one matching, we use a simple procedure, which aggregates pixels violating the uniqueness and computes the final disparity as a correlation-weighted average. This even produces a rough sub-pixel estimate.

A minimal size of the gap in $\mathcal{X}_g(\mathbf{s})$ is 1 pixel (to all sides) as shown in Fig. 5.3(c). A larger zone gap would probably further reduce the contour artifacts, but it would also decrease the accuracy of the matching.

We will refer to the algorithm which uses $\mathcal{X}_g(\mathbf{s})$ of 1 pixel zone gap as GCS-2.

5.2.4 Implementation notes

The matching table $\mathcal{T}$ of Alg. 1 is implemented as two 2D arrays of the input image sizes, each containing a match index to the other image. The second 2D array is needed for a fast check of the symmetric uniqueness constraint. The correlation map is stored separately. This is possible because there is at most one match per pixel.

In the proposed algorithm Alg. 2, the data structure for $\mathcal{T}$ is more complicated, since the uniqueness does not hold during the growth process. The $\mathcal{T}$ is very sparse, therefore, it is implemented as a 2D array of binary search trees with disparities as the keys. Inserting an element and the test on element's presence both take logarithmic time.

5.3 Experiments

We first demonstrate the principal differences between the baseline and the proposed GCS algorithms on synthetic data and then compare their performance on some real data and on a ground-truth dataset.

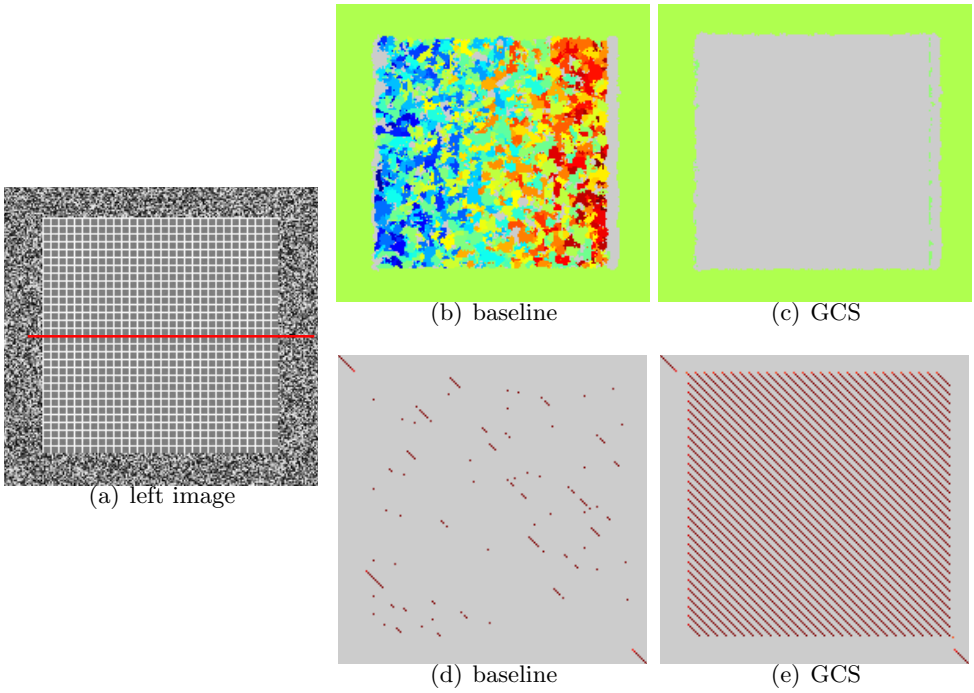
We use Moravec's NCC [105] defined in (2.13) on 5×5 window as image similarity statistic in all experiments. The default parameters of the algorithms are $\tau = 0.6$, $\mu = 0.1$. Unless stated otherwise, we use a simple pre-matcher to obtain initial seeds which we call *Harris seeds*: we take all correspondences of Harris interest points whose image similarity over 5×5 window exceeds a threshold of 0.9. Note that the Harris seeds are not necessarily a one-to-one matching.

CPU times are measured on a PC C2 2.4 GHz. Our code combines Matlab and C++.

5.3.1 Basic behavior of the algorithms

We show that the proposed GCS algorithm can handle repetitive patterns unlike in the baseline algorithm and that it has a greater ability to find all disparity components, even from a small set of random seeds. Synthetic scenes in this experiment were piecewise planar random dot 500×500 pixel stereograms.

Repetitive pattern The scene in Fig. 5.5(a) consists of a foreground plane with a repetitive texture in front of a randomly textured background plane. Harris seeds are used. Disparity map from the baseline algorithm in Fig. 5.5(b) is a set of patchy mismatches. All the patches have high correlation and are too large to be filtered out by any kind of post-processing. The proposed GCS algorithm grows all seeds into mutually competing components shown in Fig. 5.5(e) in a cross-section of matching table $\mathcal{T}$ marked red in Fig. 5.5(a). The cross-section shows similarity values, grey are unoccupied elements in $\mathcal{T}$. The final robust matching algorithm then correctly labels the repetitive area as ambiguous, assigning no disparity there, see Fig. 5.5(c). The bad behavior in the baseline

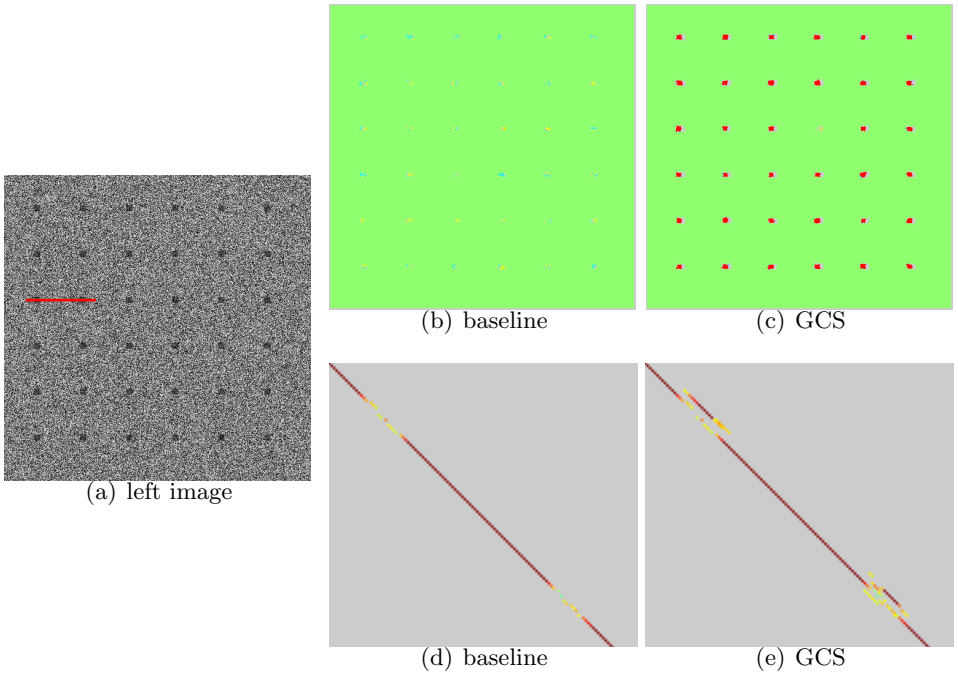


(a) left image (b) baseline (c) GCS (d) baseline (e) GCS

Figure 5.5: The ability of the baseline and proposed GCS algorithms to handle repetitive pattern. Disparity maps (b,c) and cross-sections of disparity space (d,e) for the red line in the image.

algorithm is due to a forced stop by the 'accept the first match for a pixel' in Step 1.7. As a result, the components found in disparity space are of high image similarity but they are only partial and the information on ambiguity is lost, see Fig. 5.5(d).

The capture of all disparity components The scene in Fig. 5.6(a) consists of 36 planar patches of 10×10 pixels in front of a planar 500×500 pixel background. The foreground-to-background disparity difference is five pixels. About 1700 Harris seeds are used. We used $\tau = -\infty$ to promote growth of disparity components in both algorithms. The baseline algorithm assigned correct disparity to no foreground component in Fig. 5.6(b), unlike the proposed GCS algorithm which missed only 1 of 36 in Fig. 5.6(c). We conclude the proposed GCS algorithm has the ability to locate a high-correlated disparity component even if there is no seed in it. This is again due to the ability to temporarily forego uniqueness constraint, the benefit of which illustrates Fig. 5.6(e), as opposed to the baseline algorithm in Fig. 5.6(d): The behavior helps bridge the gap between the components over a set of elements with low image similarity. Clearly, even a seed out of any true component may give rise to a path that ends up on a high-similarity component. Greater μ helps encourage this behavior. When $\mu = \infty$, the entire matching table is visited and all disparity components are found. Nevertheless, our experiments confirm



(a) left image (b) baseline (c) GCS (d) baseline (e) GCS

Figure 5.6: The ability of the baseline and proposed GCS algorithms to find all disparity components (small patches). The disparity space cross-sections (d,e) are shown for the red line in the image.

that $\mu = 0.1$ suffices in practice.

For quantitative evaluation, we performed a randomized experiment. A 500×500 pixel image with a *single* 10×10 pixel patch in a 5-pixel distance from a background is used in a repeated experiment with a set of n uniformly distributed *random* seeds. A thousand trials were performed for each n. The relative frequency of finding the small disparity component is our measure of success as a function of n. The result is shown in Fig. 5.7(a) for the baseline (red) and the proposed GCS (green) algorithms. We can see that the ability of the GCS algorithm to find the small component is indeed much larger for even a small number of seeds. The baseline algorithm has a negligible ability to find a disparity component unless the seed is located directly in it. The non-monotonic behavior of the GCS algorithm is due to a high number of wrong correspondences clogging the matching table and preventing growth over bridges of low image similarity when μ is small.

The probability of hitting the small component with a single random seed is[5] $p = 64/500^3 \approx 5 \cdot 10^{-7}$. The probability the component is hit by at least one of n random seeds is $P = 1 - (1 - p)^n$, which is shown as the blue curve in Fig. 5.7(a). We see, the baseline algorithm experiment correspond with the theory quite well which shows the

[5]The high-correlation component is just about 8×8 pixels for a 10×10 patch with a 5×5 correlation window.

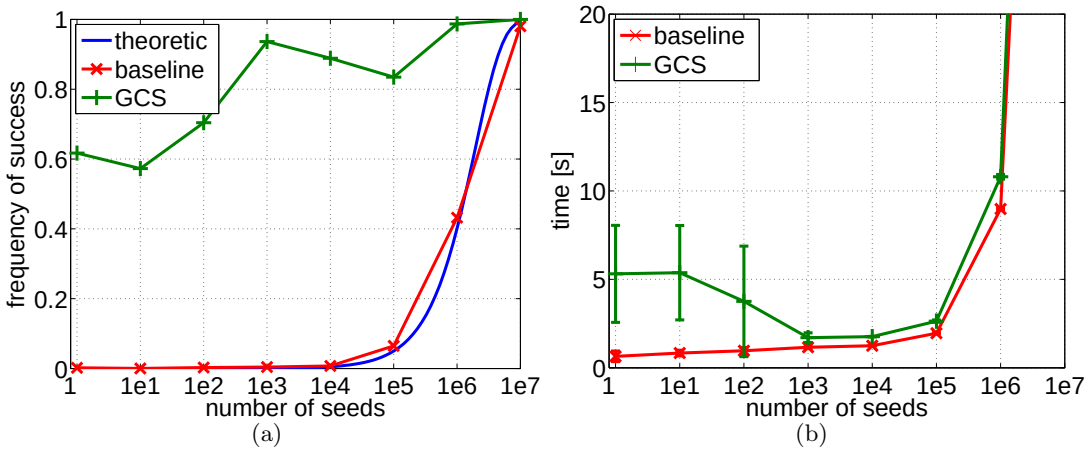


Figure 5.7: Random sampling of disparity space. The relative frequency of finding a small disparity component in the foreground (a) and the total computational time as a function of the number of random initial seeds n (b).

experimental method is correct.

We also measured the CPU time of the algorithm as a function of the number of random initial correspondences, as shown in Fig. 5.7(b), where errorbars show standard deviation. The proposed algorithm is less efficient for a small number of seeds because it has to travel over large subset of the disparity space. Both algorithms are less efficient when the number of seeds is too large. By comparing the plots in Fig. 5.7(a) and 5.7(b) we can see the proposed GCS algorithm finds a solution of similar density in shorter time.

5.3.2 Results on the laboratory test scene

We again evaluated the presented algorithms on the CMP dataset [82] as in Sec. 4.5.2, on the same textured scene consisting of thin stripes in front of the planar background, see Fig. 4.19. The algorithms are evaluated on an image series of varying texture contrast.

As a reference, we show disparity maps from the WK conversion algorithm which works with fully populated matching table, as described in [129] (the variant with no ordering constraint was used). We refer to this as the *exhaustive search algorithm*. We show the results of the baseline algorithms and proposed GCS and GCS-2 algorithms.

The results are shown in Fig. 5.8. We can see, the Mismatch Ratio MIR (matches which differ more than one pixel from the ground-truth) almost does not change at all with varying texture contrast of the images. It is the worst for the exhaustive algorithm. All growing methods have better and identical performance here. The exhaustive algorithm visits the entire disparity space which increases the probability that it encounters wrong correspondence hypotheses with correlations higher by μ than all

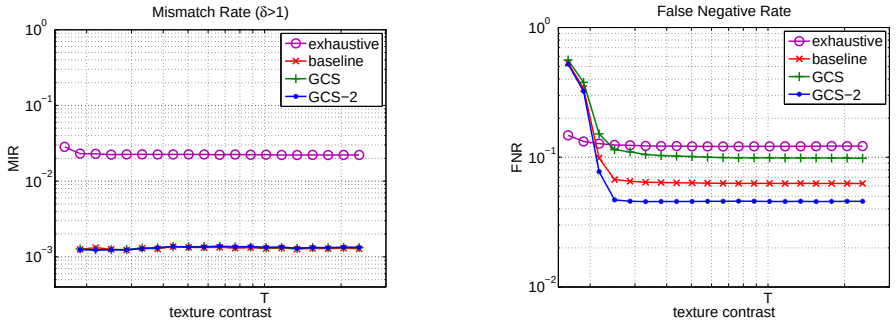


Figure 5.8: Laboratory test scene: matching error evaluation plots.

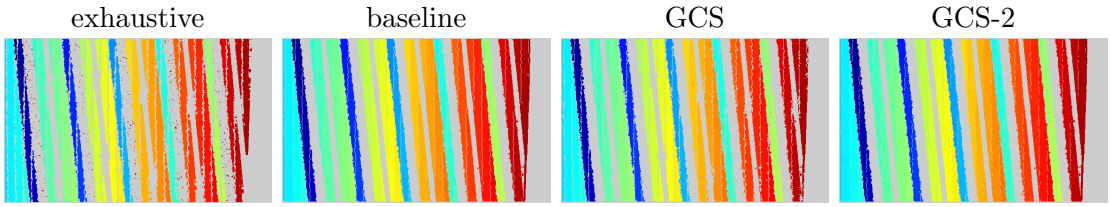
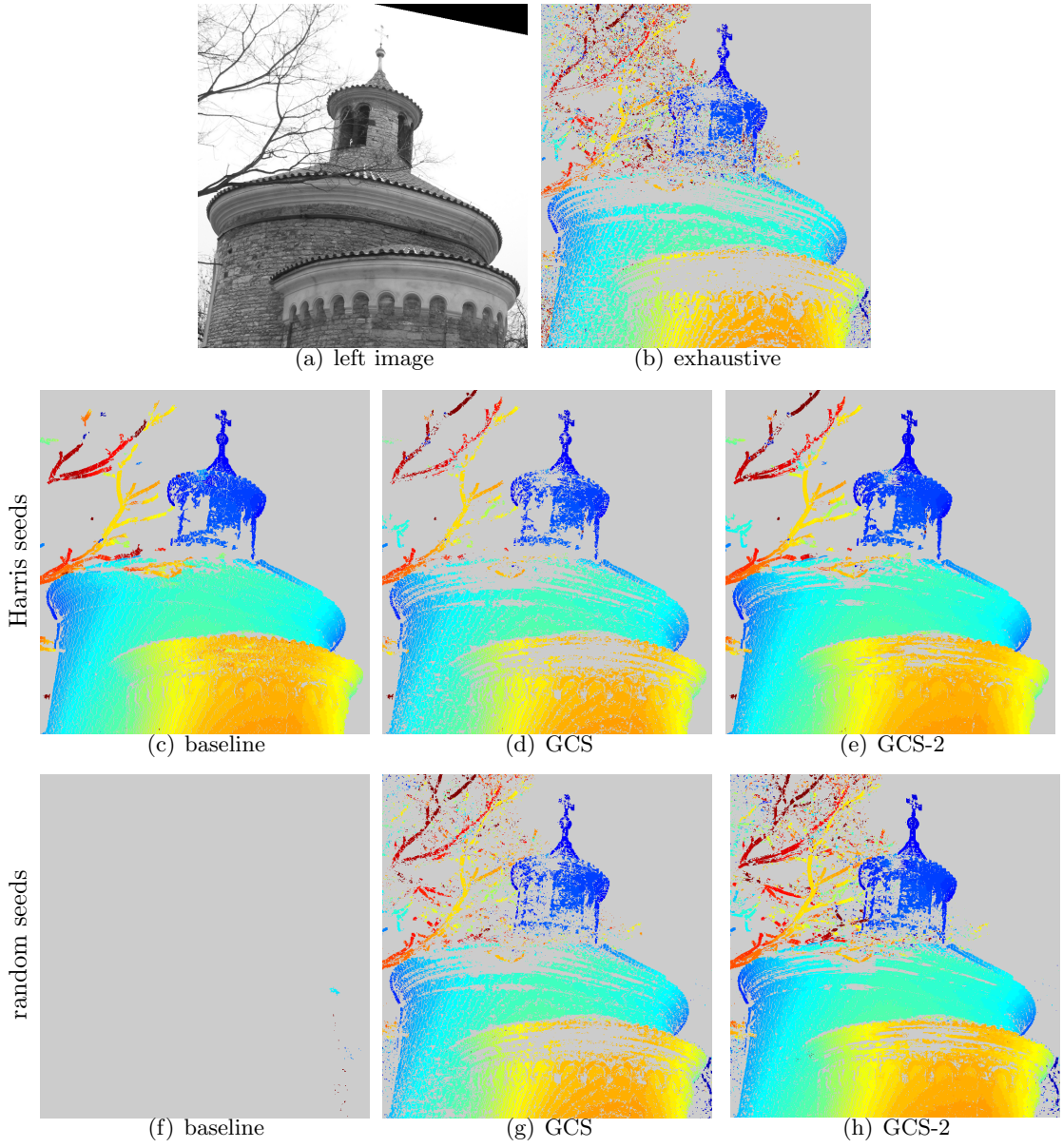


Figure 5.9: Disparity maps of the exhaustive, baseline, GCS and GCS-2 algorithms.

their X-competitors. On the other hand, growing methods visits a small fraction of disparity space following the correct disparity components, from probably mostly correct seeds obtained by the Harris pre-matcher. The probability of escaping from the true disparity component is very low on this kind of texture. The texture is random and non-repetitive, which makes the performance of the baseline algorithm the same as the proposed GCS and GCS-2 algorithms. The False Negative Ratio FNR (the sparsity of matching) is here the best for GCS-2 followed by the baseline algorithm, then by GCS. The worst performance has the exhaustive algorithm for most of the texture contrast. The exhaustive algorithm is the best for weak texture contrasts, while the growing algorithms have difficulties there. The reason is probably that the Harris pre-matcher does not find enough seeds. The baseline algorithm is denser than GCS, since GCS has additional problems with modulo 0.5px disparities. This is solved by GCS-2 (using the zone with a gap) which makes it the best performing algorithm.

Disparity maps of the four algorithms for the best texture contrast are shown in Fig. 5.9, although the differences among individual algorithms are not much obvious.

We show these results because of the performance comparison of the algorithms with CCS-statistics in Fig. 4.17, not because of the demonstration of significant difference among growing algorithms. We can see, the MIR is slightly better for the growing methods, while the FNR is better for CCS-based algorithms. This is also confirmed visually in comparing disparity maps of respective methods, see Fig. 4.19 and Fig. 5.9. Unfortunately, these results are not directly comparable, since the results in Fig. 4.17 are obtained by the exhaustive algorithm with ordering constraint, however results in Fig. 5.8 are constrained by the uniqueness only. Nevertheless, it gives a rough intuition.

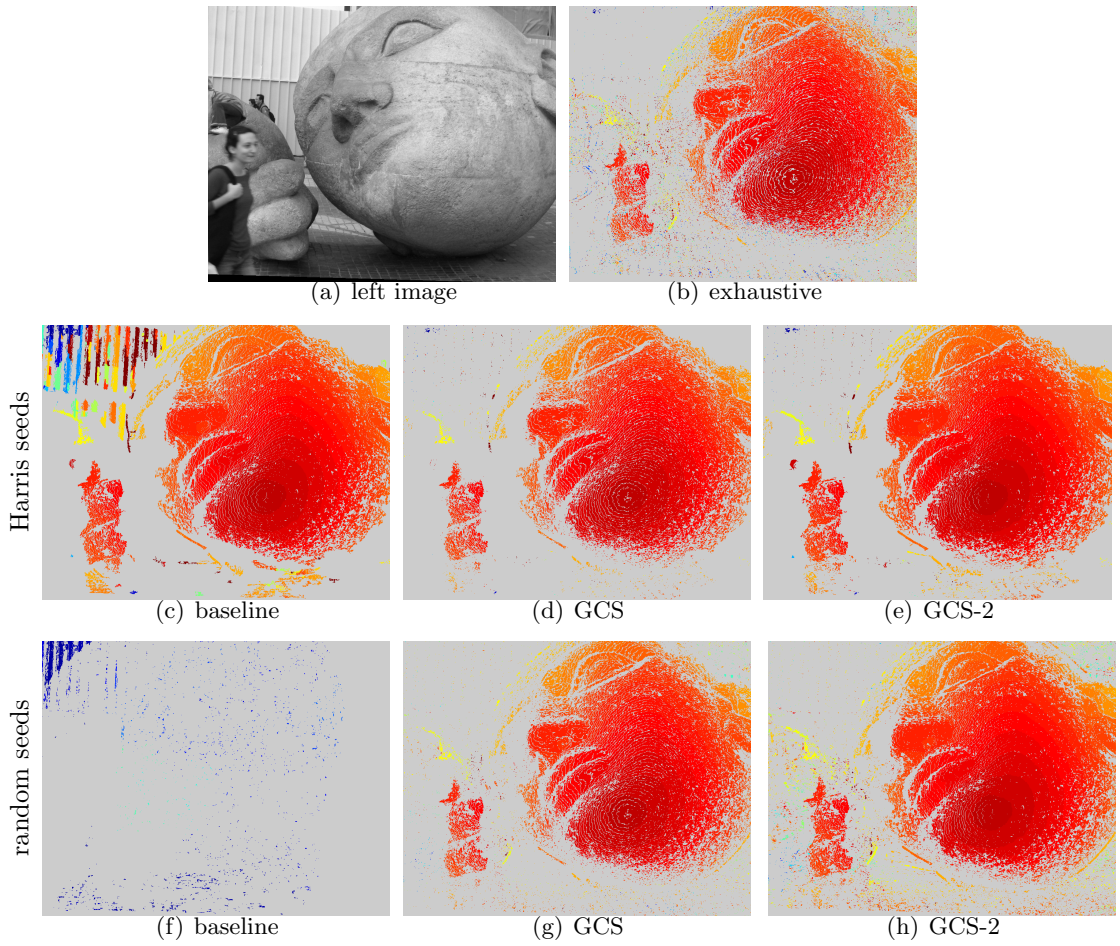


(a) left image (b) exhaustive

(c) baseline (d) GCS (e) GCS-2

(f) baseline (g) GCS (h) GCS-2

Figure 5.10: The St. Martin scene. Results of exhaustive, baseline, GCS and GCS-2 algorithms initialized by Harris and random seeds.

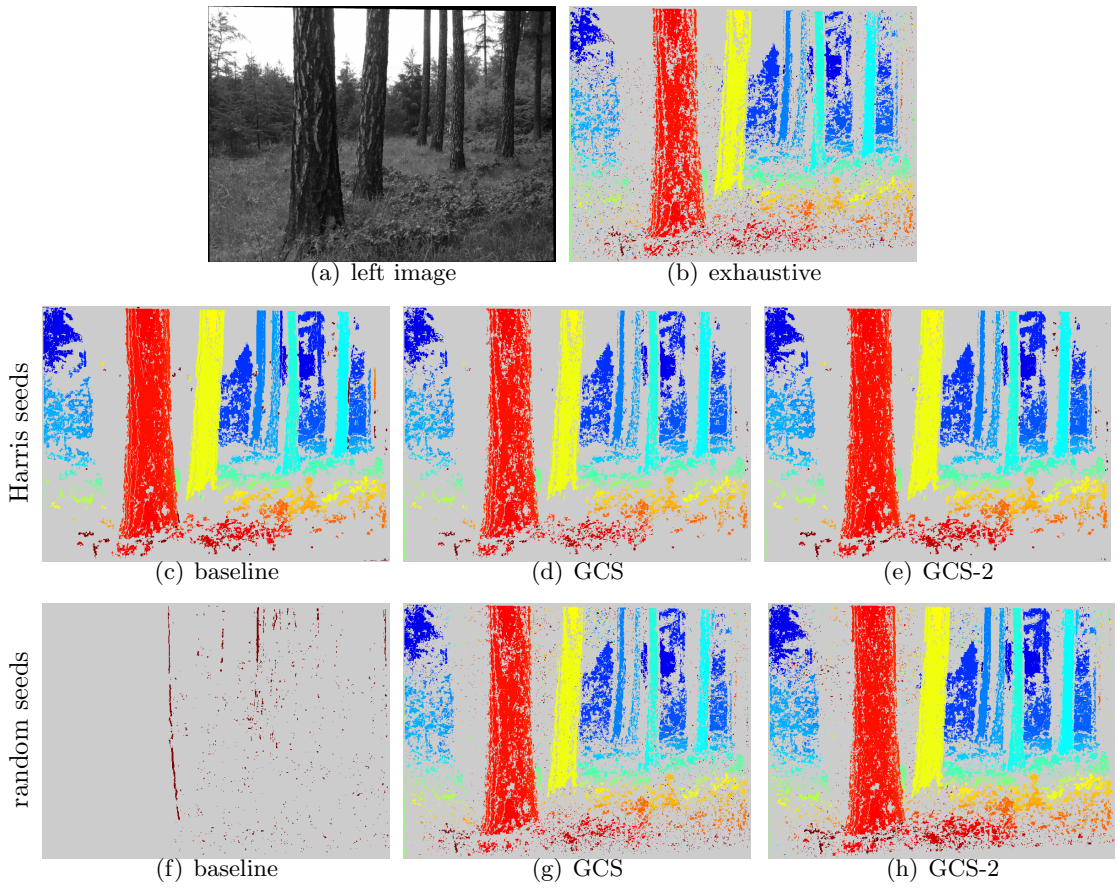


(a) left image (b) exhaustive

(c) baseline (d) GCS (e) GCS-2

(f) baseline (g) GCS (h) GCS-2

Figure 5.11: The Head scene. Results of exhaustive, baseline, GCS and GCS-2 algorithms initialized by Harris and random seeds.



(a) left image (b) exhaustive (c) baseline (d) GCS (e) GCS-2 (f) baseline (g) GCS (h) GCS-2

Figure 5.12: The Larch scene. Results of exhaustive, baseline, GCS and GCS-2 algorithms initialized by Harris and random seeds.

	St. Martin		Head		Larch	
	t [sec]	f [%]	t [sec]	f [%]	t [sec]	f [%]
exhaustive	1023.0	100	889.0	100	182.0	100
baseline Harris	3.8	0.030	2.7	0.032	1.1	0.057
GCS Harris	7.9	0.048	9.1	0.100	1.6	0.076
baseline random	3.9	0.046	2.8	0.045	1.1	0.088
GCS random	31.2	0.320	30.0	0.510	11.9	0.680

Table 5.1: The CPU times t and the fraction f of matching table that was visited. The image sizes for St. Martin, Head and Larch scenes are 1.8 Mpx, 1 Mpx, and 0.5 Mpx respectively.

5.3.3 Results on real outdoor scenes

We tested the algorithms on several complex scenes to confirm the predictions on their behavior from the synthetic-data experiment and to show the proposed algorithm is indeed fast in practice. We did not limit disparity search range for any of the four algorithms in this experiment.

Results are shown in Fig. 5.10, Fig. 5.11 and Fig. 5.12. As a reference we again show the results of the exhaustive algorithm, and results of the three growing algorithms: baseline, GCS, GCS-2. All growing algorithms were initialized by Harris seeds and also by random seeds.

CPU timings are shown in the table in Tab. 5.1, together with the fraction of matching table that was visited by the respective algorithm. The timing is dominated by growing, the final matching in GCS and GCS-2 takes less than 15% of time for both algorithms identically. The timings are thus valid for GCS-2 as well.

Harris seeds On the *St. Martin* scene, see Fig 5.10, the proposed GCS and GCS-2 algorithms, achieves lower density but the holes occur in locations where the baseline algorithm is erroneous (several-pixel errors in disparity in the cornice regions). Note the proposed algorithms grew the tree branches correctly, unlike the baseline algorithm.

There is a strong repetitive pattern in the *Head* scene due to the corrugated iron fence behind the sculpture. The baseline algorithm, see Fig. 5.11(c), suffers from illusions there as well as on the regular cobblestone paving below the sculpture. Such artifacts are highly undesirable for 3D reconstruction, since the incorrect disparity components may represent a large illusory surface of high correlation and cannot be filtered by any post-processing of the disparity map.

The *Larch* is a scene, see Fig. 5.12, of great depth and large occlusions. Results look visually similar, but the baseline algorithm makes more small mismatches at occlusion boundaries.

Random seeds For all scenes, we used 10 random, i.e. wrong, seeds. We set $\tau = -\infty$, and deleted all disparities whose image similarity dropped below 0.6 in the final map.

The experiment shows that the proposed GCS and GCS-2 algorithms can indeed obtain a dense map from a very small number of random initial seeds, showing their quality need not be high to succeed. The bad quality of the seeds just increases the runtime, as clearly visible in Tab. 5.1. In a repeated experiment, the proposed GCS and GCS-2 algorithms always succeeded unlike the baseline algorithm which always failed finding any correct disparity component.

The results from the exhaustive search algorithm are dense, large ambiguous regions are correctly identified, but there are more mismatches unlike in the proposed algorithm. This is caused by a higher number of competing putative correspondences whose image similarity is high due to statistical fluctuations of the MNCC statistic.

The GCS-2 produces denser results than GCS algorithm by removing the contour artifacts around modulo 0.5 pixels thanks to the modified inhibition zone with the gap. It does not yield more matching errors visually. The difference between the proposed algorithms is best viewed in the electronic version of the thesis.

5.3.4 Middlebury dataset

Results on the standard new Middlebury dataset [136] are shown in Fig. 5.13. For both baseline and proposed algorithms, the ROC curves were obtained using the same method as in [54], spanning $\tau \in [-1, 1]$. For GCS and GCS-2, we set $\mu = 0.05$. The exhaustive search algorithm has parameters of a similar meaning. We did not limit the disparity search-range.

We can see the proposed algorithm is consistently better than baseline algorithm producing less errors at the same density. The difference in matching quality is large on the Tsukuba and Teddy scenes, due to ambiguous repetitive structures which are correctly handled by the proposed algorithm. On the Venus, the error of the proposed algorithm for high densities is higher than in the baseline algorithm, since the simple planar scene is well suited for the greedy baseline algorithm.

The GCS-2 algorithm is consistently better than GCS. The GCS-2 does not produce additional errors, it only increases density compared to GCS by reducing the artifacts of undecided contours between disparity levels. This quantitatively confirms the benefits of the modified inhibition zone with the gap.

The behaviour of all involved algorithms can also be viewed in the disparity maps, see Fig. 5.14. The parameters were $\mu = 0.05$, $\tau = 0.6$ for the proposed GCS and GCS-2 algorithms, $\tau = 0.6$ for baseline algorithm. The exhaustive search algorithm was set such that it achieves approximately the same density as GCS algorithm. Notice several mismatches produced by the baseline algorithm in the repetitive region of Teddy and Tsukuba. See the reduction of contour artifacts by GCS-2 compared to GCS, for instance in slanted planes in the Teddy scene, or in the mask in the Cones scene.

We observed that most of the errors in the above algorithms occur at occluding boundaries. Image pre-segmentation as e.g. in [168] would help reduce this error. Naturally,

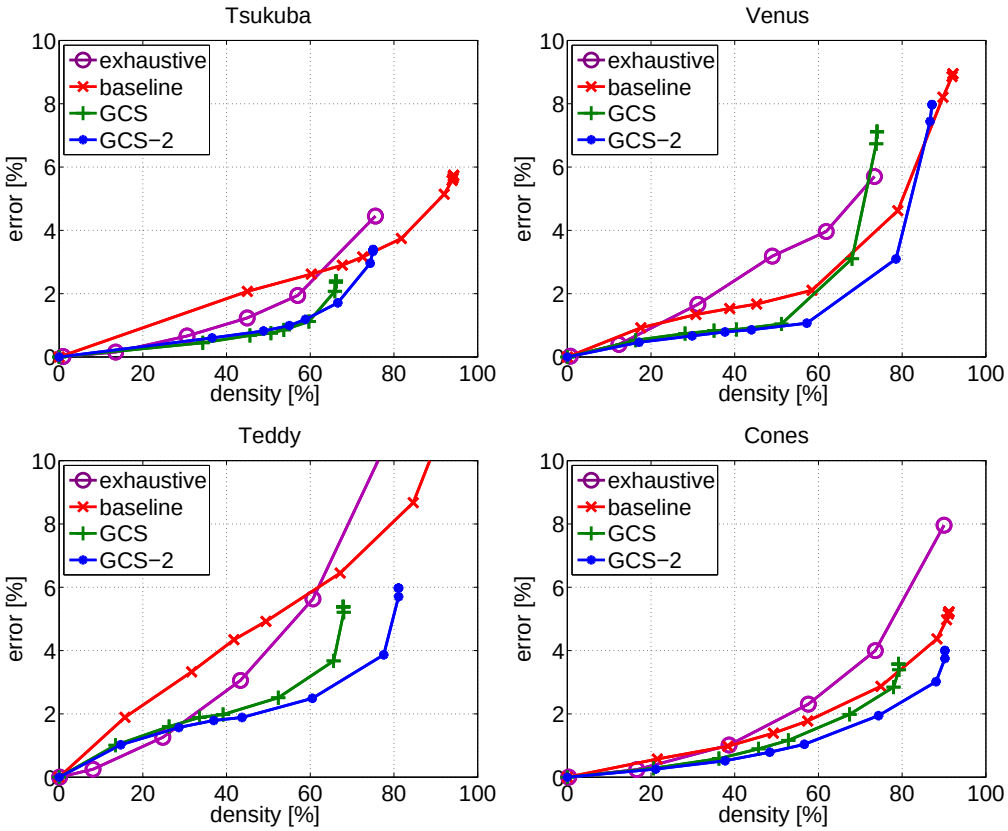


Figure 5.13: Middlebury new test results: ROC curves, error rate in non-occluded regions versus matching density.

these results are inferior to semi-dense methods using a global optimality in the MAP sense, e.g. [53, 54, 163]. But the results are comparable to the exhaustive algorithm [129], confirming the efficiency of the disparity space sampling.

The average time on Middlebury images is about 1 sec. The [53, 54] report faster times, but it is not clear if they count similarity computation and if they limit the disparity search range or not. Our runtime on small images is dominated by Matlab overhead. The time efficiency of the algorithm (due to visiting less than 1% of disparity space) becomes apparent in images above 1 Mpx, whereas Middlebury images are just about 0.15 Mpx.

5.4 Conclusions

We have proposed a novel disparity component growing algorithm that can cope with much more difficult cases (repetitive patterns, complex scene) than similar existing algorithms, that can recover from errors in initial seeds, and that does not require a seed on every component in disparity space. Hence, the seeds need not be salient image features, which opens a way to random sampling of disparity space. The changes against the standard seed growing algorithms are small, but their consequences are deep and allow the resulting algorithm to be well grounded in the theory of robust matching.

Although this has not been demonstrated in this chapter, the algorithm is in no way restricted to narrow baseline stereo images, since whenever a new seed is created, it can inherit an updated set of parameters that describe relative image distortion, as briefly discussed in Sec. 5.2. The algorithm can be easily adapted for multi-image matching.

An implementation of the GCS algorithm is available at `http://cmp.felk.cvut.cz/~cechj/GCS`.

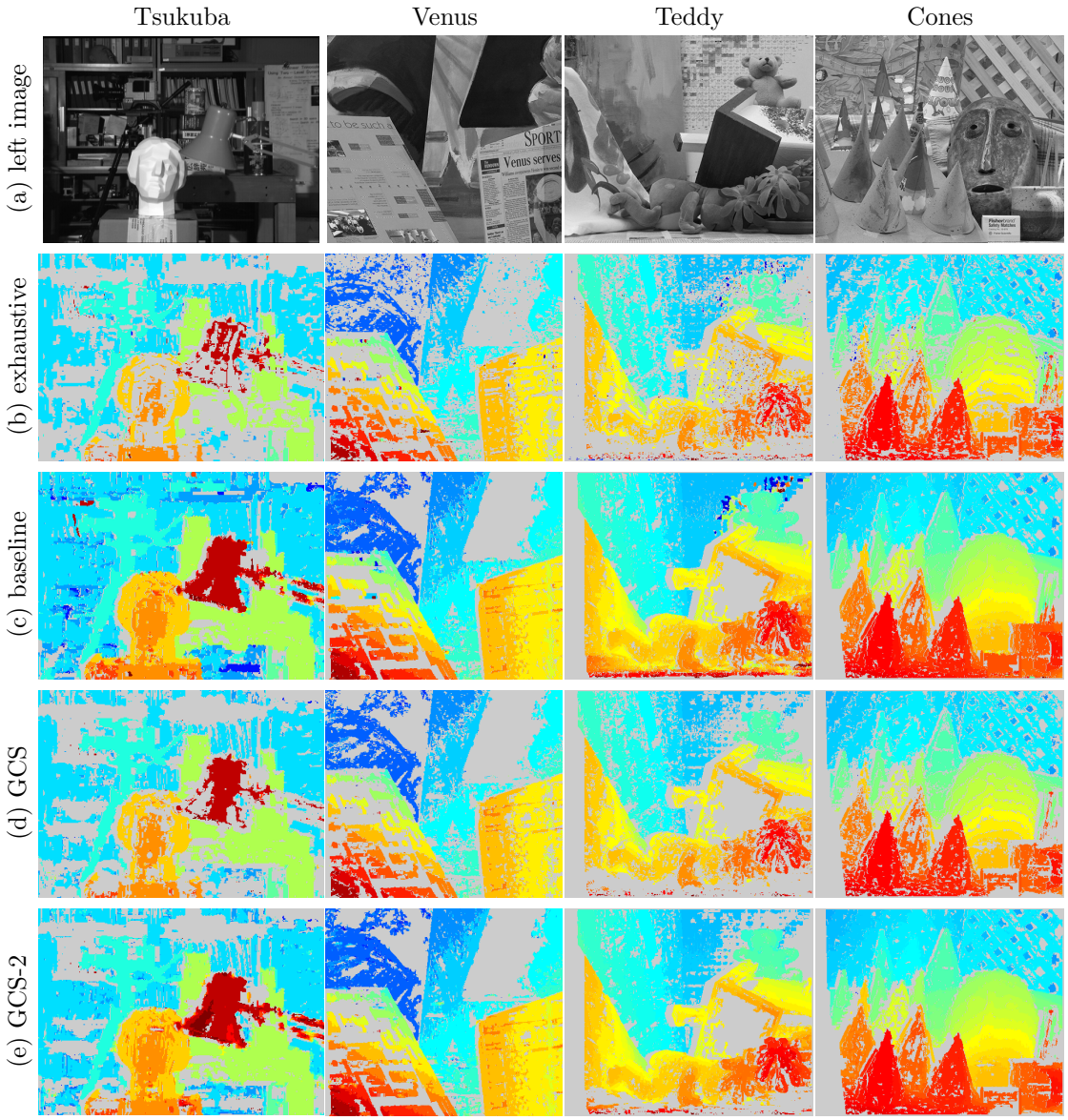


Figure 5.14: Results on Middlebury dataset as disparity maps.

6 Efficient sequential correspondence selection by cosegmentation

Dense stereoscopic matching is not an isolated problem. In order to estimate the epipolar geometry, it is necessary to find sparse correspondences. A question arises, whether a wide-baseline matching and dense matching can interact. In this chapter, we will show a possible mechanism. It turns out that the method goes slightly beyond the scope of 3D reconstruction. A part of the automatic 3D reconstruction from unorganized images is to find clusters of images which overlap. This way, the task comes near to image retrieval. Therefore we will often use this term in this chapter.

In many retrieval, object recognition and wide baseline stereo methods, correspondences of interest points (distinguished regions, transformation covariant points) are established possibly sublinearly by matching a compact descriptor such as SIFT. We show that a subsequent cosegmentation process coupled with a quasi-optimal sequential decision process leads to a correspondence verification procedure that has (i) high precision (is highly discriminative) (ii) good recall and (iii) is fast. The sequential decision on the correctness of a correspondence is based on simple attributes of a modified dense stereo matching algorithm. The attributes are projected on a prominent discriminative direction by SVM. Wald's sequential probability ratio test is performed on SVM projection computed on progressively larger co-segmented regions. Experimentally we show that the process significantly outperforms the standard correspondence selection process based on SIFT distance ratios on challenging matching problems.

6.1 Introduction

Many successful image retrieval, object recognition and wide baseline stereo methods exploit correspondences of distinguished regions[1]. Remember that such correspondences are used in order to estimate calibration of cameras from a collection of unorganized images, as described in Sec. 1.3. Most real-world visual recognition problems are large scale where correspondences between regions from a query (test) image and many database (training) images of objects or scenes are sought. To achieve acceptable response times,

[1] Terms "transformation-covariant regions", "viewpoint invariant features", "interest points", "salient points" and "patches" also appear in the literature. We adopt the term "distinguished region" as a concise shortcut for the self-explanatory but unwieldy "repeatably-detectable transformation-covariant regions".

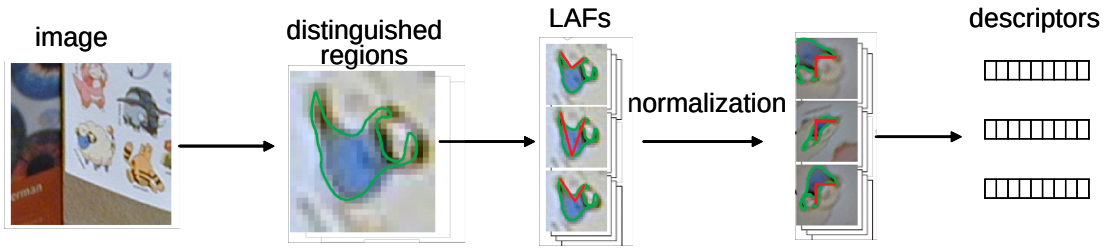


Figure 6.1: Describing a distinguished region by a compact descriptor invariant to local geometric and photometric changes. "LAFs" stands for "Local Affine Frames". *Images by courtesy of Štěpán Obdržálek.*

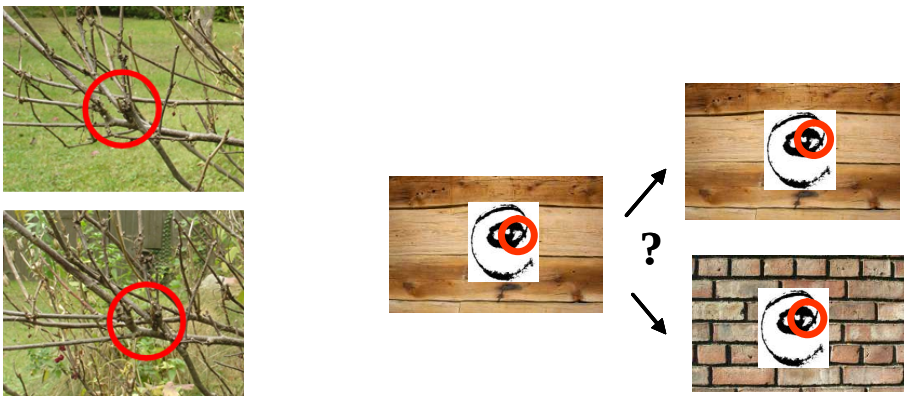
large problems require the time complexity of the region matching process be sublinear in the size of the database; memory footprint of the database representation becomes a concern too. The standard solution is to describe regions with a compact descriptor such as SIFT [95] or some discretization of it (e.g. "visual words" [143]) and to store database image representations in a search tree (k-d [95], metric [92], k-means [111, 121, 29]).

The matching process typically proceeds as follows [159, 100, 95]. Distinguished regions are detected in the image and local affine or similarity covariant coordinate frames are constructed for each region. *Measurement regions*, i.e. image patches of typically rectangular or elliptical shape, are specified in terms of the local coordinate frames. For each region, a descriptor (such as SIFT) is computed from the signal in the measurement region, after both photometric and geometric normalization. Additionally, the descriptor may be compressed, by e.g. quantization.

This process, schematically visualized in Fig. 6.1, has the following main characteristics: (i) all steps are performed in individual images independently, (ii) the shape and size of the measurement region is a fixed function of the shape and size of the distinguished region and (iii) the descriptor has the same form for all regions, e.g. it is a vector in R^d. These properties facilitate fast sublinear region matching, e.g. via search tree or hashing.

However, the fixed size and shape of the measurement region necessarily involves a compromise. In general, the larger the measurement region, the more discriminative the information inside it. On the other hand, large measurement regions may violate the local planarity assumption of wide-baseline matching methods, are more likely to straddle object boundary or to be affected by occlusion. Moreover, they are more sensitive to localization errors of local frames. For "non-compact" objects, such as elongated and wiry ones, the fixed shape has problems.

Consider, for instance, the two images depicted in Fig. 6.2(a). Only a very small circular or rectangular regions around the distinguished region on the branch will not include signal from the background, which is different for the two views. On the other hand, consider the images shown in Fig. 6.2(b). The measurement region inside the circle is too small, any descriptor computed from the region will be close to identical for



(a) measurement region too large and/or with unsuitable shape

(b) measurement region too small

Figure 6.2: Problems with a fixed shape and size of the measurement region.

both images on the right and the correct match cannot be reliably established.

A better estimate of correspondence quality (a prediction of it being correct) can be obtained by looking at both test and training image simultaneously, e.g. by attempting to expand the correspondence domains or to improve the precision of registration. The value of correspondence growing methods has been demonstrated in [162, 39], sometimes with impressive results, e.g. those achieved by the dual bootstrap method [177, 145]. Most approaches to simultaneous cosegmentation and registration focus on the problem of finding the largest corresponding domain and co-domain [177, 39, 93].

Our objective is almost opposite: given acceptable false positive and false negative rates, design the fastest possible test for correctness of a correspondence, based on cosegmentation of regions of growing size. We formulate the problem as sequential decision making, performing Wald's sequential probability ratio test. The test is based on simple statistics of a modified dense stereo matching algorithm which are projected on a single prominent discriminative direction by a linear SVM.

Of course, we do not want to lose the excellent large-scale matching properties of descriptors based on measurement regions of fixed size and scale. We therefore apply the co-segmentation process only to *tentative correspondences* obtained by a sub-linear process, such as kD-tree search. In fact, any such process for generating tentative correspondences can be set to be much more permissive, outputting higher number of correspondences with lower inlier ratios but containing larger number of inliers. We shall see that after filtering by simultaneous cosegmentation, inlier ratios are (more than) recovered and the larger number of inliers leads to higher recognition rates.

On challenging matching problems, *we show that the selection of correspondences based on sequential co-segmentation is very efficient, runs near to real-time and significantly outperforms the standard correspondence process based on SIFT distance ratios*, producing a higher number as well as higher percentage of correct correspondences.

Algorithm 3 SCV: Sequential Correspondence Verification

Require: images $\mathbf{I}$, $\mathbf{I}'$,
correspondence with affine frame $(x, y, \mathbf{A})$,
SIFT ratio s_r,
false positive and false negatives rates (α, β),
model: learned SVM parameters θ_i,
likelihoods $\mathbf{p}_i(q|+1)$, $\mathbf{p}_i(q|-1)$.

3.1: **for** $i = 1$: maximum number of decision stages **do**
3.2: $\mu_i = \begin{cases} 0, & i=1, \\ 10^{i-1}, & i>1. \end{cases}$
3.3: $(\bar{g}, \bar{c}, \bar{u}) = \mathrm{grow}(\mathbf{I}, \mathbf{I}', (x, y, \mathbf{A}), \mu_i)$.
3.4: $q = \mathrm{SVM}(s_r, \bar{g}, \bar{c}, \bar{u}, \theta_i)$.
3.5: $L = \frac{\mathbf{p}_i(q|+1)}{\mathbf{p}_i(q|-1)}$.
3.6: **if** Wald SPRT(L, α, β) is conclusive **then break**.
3.7: **end for**
3.8: **return** likelihood ratio L.

Consequently, combinatorial procedures for estimation of a geometrically consistent subset of correspondences with time complexity sensitive to inlier ratios (polynomial dependence), e.g. RANSAC, should always adopt sequentially terminated cosegmentation as a pre-processing step.

The method scales well: the number of potential correspondences for a query image region can be controlled. If it is constant, the total time complexity of the region expansion process is independent of the size of the database and linear in the size of the input (number of regions in the query image). On a large scale retrieval experiment [111], we observed that the time needed to carry out the sequential procedure is not significant in comparison with the time needed for the initial indexing process for establishing tentative correspondences.

The rest of this chapter is organized as follows. The method is described in Sec. 6.2 and the training data and learning procedure for the sequential classifier in Sec. 6.3. Experiments validation is presented in Sec. 6.4. Conclusions are summarized in Sec. 6.5.

This chapter is a significantly extended and modified version of [23].

6.2 The Sequential Correspondence Verification algorithm

The basic idea of the approach is to distinguish, as fast as possible, correct and incorrect correspondences via a dense matching (pixel-to-pixel) growing algorithm. The requirements of high speed and quality of the decision process are contradictory. We therefore propose a quasi-optimal sequential decision algorithm that minimizes time to decision, given user-specified probabilities of false positive and false negative rates.

The proposed Sequential Correspondence Verification algorithm (SCV) is summarized in Fig. 6.3. It proceeds in decision stages i. In the first decision stage, a fast dense stereo

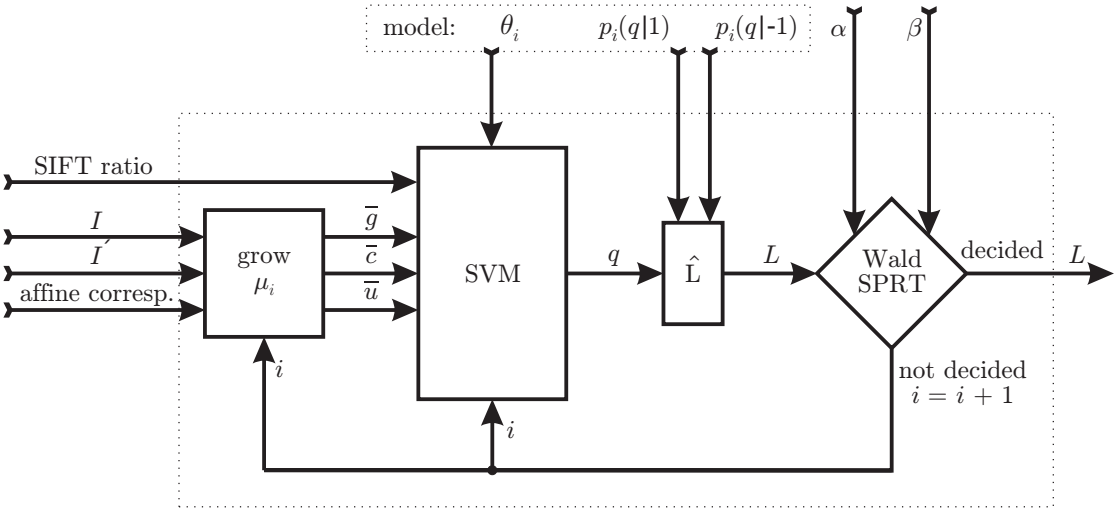


Figure 6.3: The Sequential Correspondence Verification algorithm.

matching growing algorithm, described in Sec. 6.2.1, is initialized by a tentative correspondence of local affine frames. The verification proceeds by attempting to match discriminative neighboring pixels. After a maximum number of growing steps μ_i, the cosegmentation process returns three simple statistics $(\bar{g}_i, \bar{c}_i, \bar{u}_i)$ characterizing the quality of the correspondence: the growth rate $\bar{g}_i$ – the size of the grown region divided by the maximum number of attempted growing steps μ_i, the average correlation $\bar{c}_i$ of the region, and the average number of pixels violating the uniqueness $\bar{u}_i$, i.e. non-bijectivity of matching.

The vector of statistics is projected by SVM to a scalar quality q_i which simplifies the estimation of class conditional probabilities (likelihoods) $\mathbf{p}(q_i|+1)$, $\mathbf{p}(q_i|-1)$ of correct and incorrect correspondence classes respectively. The region statistics are augmented with the first to the second nearest SIFT descriptor distance ratio s_r, a standard measure for selection of tentative correspondence technique [95]. We call s_r the *SIFT ratio*. The Wald's Sequential Probability Ratio Test (SPRT) is performed on the likelihood ratio $L_i = \mathbf{p}(q_i|+1)/\mathbf{p}(q_i|-1)$. If the SPRT test is conclusive, the algorithm terminates and the correspondence is assigned the likelihood ratio L_i of the decision. Otherwise, another decision stage i is performed, i.e. the cosegmentation is continued with exponentially larger maximum number of growing steps μ_i, in Step 3.2 of Alg. 3, potentially producing more discriminative statistics, since it is based on more measurements. Note that $\mu_1 = 0$, which means the decision in the first stage is based solely on the SIFT ratio without growing.

The process continues until the maximum number of decision stages i is reached. In our experiments, we set the maximum number of decision stages to four, so the largest growth is by $\mu_4 = 1000$ steps. Details of the algorithm are described below.

The choice of the three statistics was driven mainly by computational considerations

and we make no claims about its optimality. Beside computational speed, we attribute to the simplicity of the characterization of the growing process the good generalization of the sequential classifier. For instance, the sequential classifier performed almost equally well on correspondences established on a different types of distinguished regions than it was trained on, despite the fact that it is unlike that in general image statistics are the same for different detection processes. Moreover, even after a very modest number of training examples, the classifier performed well on a very large test set. With satisfactory performance, we did not further investigate the feature selection problem (including feature number). Equally ad hoc was the choice of linear SVM, which provided good classification rate, fast learning and execution and trivial implementation.

6.2.1 Growing algorithm

The following simple algorithm explores regions around the input tentative correspondence. The growing mechanism is inspired by [26, 118, 93, 102].

Each initial correspondence defines a local affine mapping from the reference image $\mathbf{I}$ to the target image $\mathbf{I}'$. The mapping generates several pixel to pixel correspondences, the seeds[2]. Seed $\mathbf{s} = (x, y, \mathbf{A})$ is a point (x, y) in $\mathbf{I}$ with associated affine transformation $\mathbf{A}$ which maps the local neighborhood to the other image $\mathbf{I}'$:

$$\begin{aligned} x' &= a_1 x + a_2 y + a_3, \\ y' &= a_4 x + a_5 y + a_6, \end{aligned} \tag{6.1}$$

or simply $(x', y') = \mathbf{A}(x, y)$.

The procedure is presented in pseudo-code as Alg. 4. The inputs are the images $\mathbf{I}, \mathbf{I}'$, the set of initial seeds $\mathcal{S}$ and the maximum number of growing steps μ. The output are three statistics $\bar{g}, \bar{c}, \bar{u}$ which characterize the (in)correctness of the input correspondence.

The algorithm computes the image correlation $\mathrm{corr}(\mathbf{s})$ of all initial seeds $\mathbf{s} \in \mathcal{S}$, Step 4.2, as Moravec's normalized cross-correlation [105] (MNCC), see (2.13), of 5×5 pixel window $\mathbf{w}$ centered at pixel (x, y) in the reference image and window $\mathbf{w}'$ centered at $\mathbf{A}(x, y)$ in the target image, deformed according to the affinity $\mathbf{A}$.

Set $\mathcal{S}$ is organized as a priority queue according to MNCC. A seed is removed from the top of the queue, and for all its 4-neighbors (left, right, up, down) in the reference image, the best correlating candidate in $\mathcal{N}_k(s)$ is found (out of 9 possible positions in the target image), Step 4.6, analogically to (5.1) such that

$$\begin{aligned} \mathcal{N}_1(s) &= \big\{(x-1, y, \mathbf{A}_{c-1,r}) \mid c, r \in \{-1, 0, 1\}\big\}, \\ \mathcal{N}_2(s) &= \big\{(x+1, y, \mathbf{A}_{c+1,r}) \mid c, r \in \{-1, 0, 1\}\big\}, \\ \mathcal{N}_3(s) &= \big\{(x, y-1, \mathbf{A}_{c,r-1}) \mid c, r \in \{-1, 0, 1\}\big\}, \\ \mathcal{N}_4(s) &= \big\{(x, y+1, \mathbf{A}_{c,r+1}) \mid c, r \in \{-1, 0, 1\}\big\}, \end{aligned} \tag{6.2}$$

[2] In our experiments, this is realized by a local affine frames (LAF) constructed on Maximally Stable Extremal Region [113, 99] (MSER) and Hessian Affine points [134]. We take the three points in a LAF as the initial seeds of the growing process.

Algorithm 4 The Growing Algorithm

Require: images $\mathbf{I}$, $\mathbf{I}'$,
initial correspondence seeds $\mathcal{S}$
maximum number of growing steps μ.

4.1: Initialize matching maps $\mathbf{T}(:,:) = 0,\ \mathbf{T}'(:,:) = 0$,
variables $K := G := C := U := 0$.
4.2: Compute the image correlation for all seeds $\mathbf{s} \in \mathcal{S}$.
4.3: **while** $K \leq \mu$ **and** $\mathcal{S}$ not empty **do**
4.4: $K := K + 1$.
4.5: Draw the seed $\mathbf{s} \in \mathcal{S}$ of the best similarity $\mathrm{corr}(\mathbf{s})$.
4.6: **for** each of the best neighbors $\mathbf{t}_k^*$ in $\mathcal{N}_k(\mathbf{s})$:
$\mathbf{t}_k^* = (x, y, \mathbf{A}) = \underset{\mathbf{t} \in \mathcal{N}_k(\mathbf{s})}{\mathrm{argmax}}\, \mathrm{corr}(\mathbf{t}),\ k \in \{1, 2, 3, 4\}$
do
4.7: $c := \mathrm{corr}(\mathbf{t}_k^*),\ c_2 := \max_{\mathbf{t} \in \{\mathcal{N}_k(\mathbf{s}) \setminus \mathbf{t}_k^*\}} \mathrm{corr}(\mathbf{t})$.
4.8: **if** $c \geq \tau$ **and** $c - c_2 \geq \epsilon$ **and** $\mathbf{T}(x, y) = 0$ **then**
4.9: $G := G + 1,\ C := C + c$.
4.10: **if** $\mathbf{T}'(\mathbf{A}(x, y)) = 1$ **then**
4.11: $U := U + 1$.
4.12: **end if**
4.13: Update the matching maps
$\mathbf{T}(x, y) := \mathbf{T}'(\mathbf{A}(x, y)) := 1$ and
4.14: the seed queue $\mathcal{S} := \mathcal{S} \cup \{\mathbf{t}_k^*\}$.
4.15: **end if**
4.16: **end for**
4.17: **end while**
4.18: **return** growth rate $\bar{g} := \frac{G}{\mu}$, average correlation $\bar{c} := \frac{C}{G}$, average uniqueness violation $\bar{u} := \frac{U}{G}$.

where

$$\mathbf{A}_{c,r} = \begin{bmatrix} a_1 & a_2 & a_3 + a_1 c + a_2 r \\ a_4 & a_5 & a_6 + a_4 c + a_5 r \end{bmatrix}. \tag{6.3}$$

If the highest correlation exceeds threshold $\tau = 0.5$ and the difference of the first and second highest correlations is above $\epsilon = 0.01$ and the point is unmatched so far in the reference image, then a new match is found, Step 4.8. Next, the counter for the region size G is incremented and correlation value c is added to sum C. If the pixel in the target image $\mathbf{I}'$ is already matched, the counter for uniqueness violation U is incremented, Step 4.11. The binary matching maps $\mathbf{T}$ and $\mathbf{T}'$ are updated and the found match becomes a new seed. Up to four seeds are created in each growing step.

The process continues until there are no seeds in the queue or the algorithm is stopped when reaching the maximum number of growing steps μ, as tested in Step 4.3.

Examples of region growth in the co-segmentation process are illustrated in Fig. 6.4.

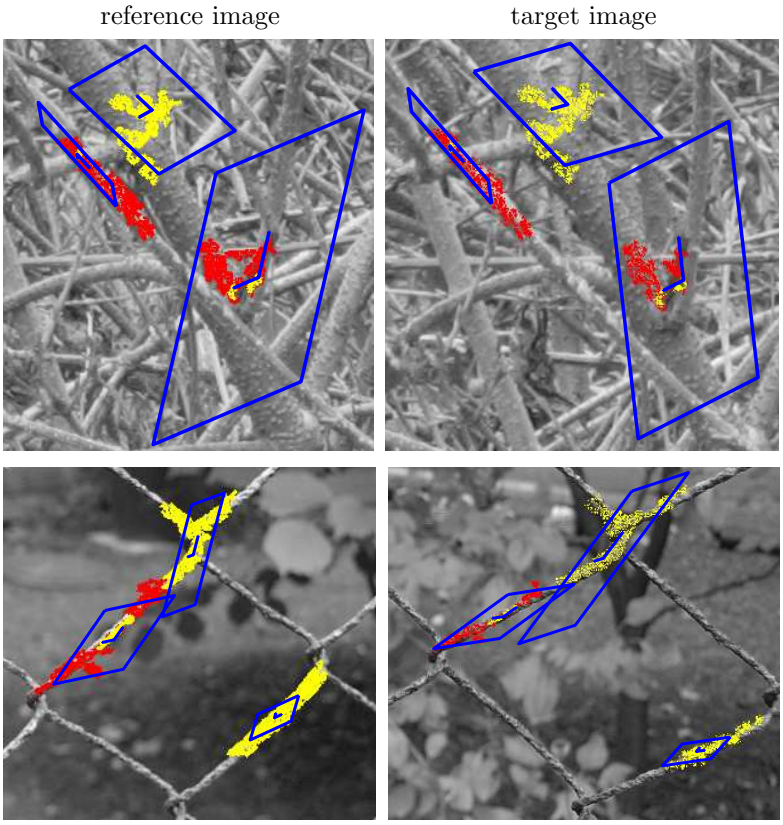


Figure 6.4: Examples of region growth in the co-segmentation process. Selected pixels at the time of decision on correctness (in yellow) and after 1000 growing steps (in red). The red pixels are shown for visualization purposes only, their correspondence is not evaluated by the SCV algorithm. Corresponding local affine frames and measurement regions of SIFT descriptors are in blue. See text for more details.

Local affine frames are shown in blue as a pair of line segments. The three endpoints of the segments are the seed of the growing process. Blue parallelograms delineate measurement regions, i.e. parts of the image where SIFT descriptors are computed. Yellow marks pixels inside the region at the time of the decision on the correctness of the correspondence made by the SCV algorithm. Red marks pixels that would be chosen if the process was left to grow the maximum number of $\mu_4 = 1000$ steps. Note that (i) the SCV decision is often reached after growing over a very small number of pixels and (ii) the shape of the region is data-dependent, preferring areas with edges and high variance of the signal where MNCC response is high. Unsurprisingly, pixel correspondences follow correctly the 3D surfaces (branches of the shrub, parts of the fence); in fact, a small local disparity map is computed.

The measurement regions include large parts of the background that is different in the two images. It might be surprising that the regions in Fig. 6.4 are correctly matched,

given that the first test in the sequential classifier is based on the SIFT ratio. We believe there are (at least) three reasons for the favorable outcome. First, our test on the SIFT ratio is very conservative. Second, the centers of the regions are on corresponding 3D structures and SIFT applies a Gaussian weighting function that reduces influence of the outer parts of the parallelogram where there is no correspondence. Finally, SIFT is an array of histograms of gradients. The strong gradients are in correspondence in both pairs of images, areas without strong edges are irrelevant for the SIFT representation.

Discussion Unlike Vedaldi and Soatto's region growing algorithm [162], Algorithm 4 includes no explicit regularization either of the mapping or of the shape of the cosegmented regions. The reason is that the algorithm grows only in informative areas with distinguishing signal (texture), so regularization is not needed. Areas without texture are ambiguous and do not help to distinguish correct and incorrect correspondences. Growth is restricted to unambiguous areas by requiring MNCC statistic[3] to stay above a threshold τ, and by requiring the distance of the first and second highest correlation to be above ϵ, Step 4.8. Parameters τ and ϵ were set empirically, as a trade-off between reliable growth of correct correspondences and preventing the growth into ambiguous regions. Implicit surface smoothness is enforced. The disparity gradient change is constrained by (6.2), similar constraint is applied in [93].

In wide baseline dense stereo [118, 102, 72], local affine parameters (a_1, a_2, a_4, a_5) representing a window deformation due to surface slant are optimized after each growing step in order to facilitate maximum growth on curved or projectively distorted surfaces. However, our goal is different: for correspondence verification the surface need not be grown too far. Therefore, in our algorithm, the parameters inherited from the initial seed are kept constant, which is significantly faster than the iterative optimization. Experiments show that a small imprecision of the local affine parameters is not critical, possibly due to the fact that effects of transformation errors are subsumed in disparity (gradients).

6.2.2 Statistical correspondence quality

Ideally, correspondence quality would be a function of the probability that a pair of grown patches is a projection of the same 3D surface, as calculated e.g. via MRF on the image grid by global methods in dense stereo [76]. However, finding the MAP solution is computationally intensive even for simple fields. Therefore, we use the efficient growing algorithm as a suboptimal solution and model the correspondence quality on the basis of elementary statistics that were empirically shown to discriminate correct and incorrect correspondences.

The class conditional probability densities of the adopted statistics is shown in Fig. 6.5. We observed that the growth rate $\bar{g}$ is typically larger for correct correspondences than

[3]Note, the MNCC is a zero mean normalized correlation, see (2.13). For areas without texture, after subtracting the mean values of signals in windows, the rest is an uncorrelated noise which results in a low value of the statistic.

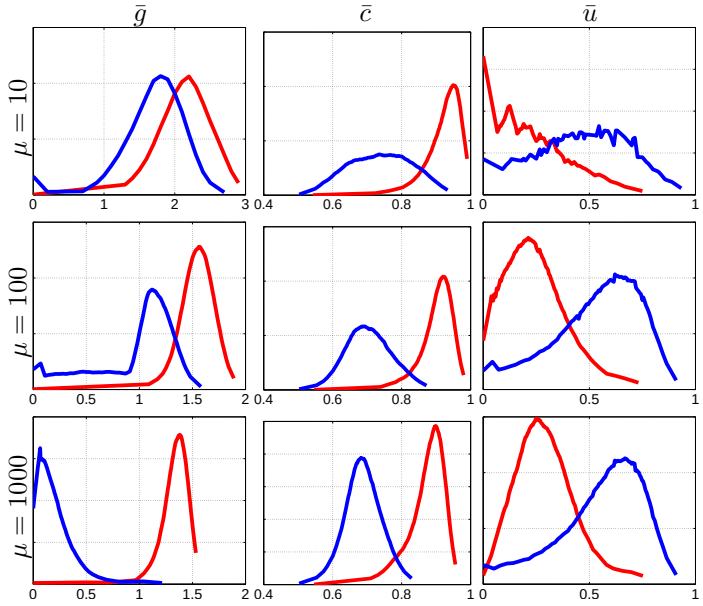


Figure 6.5: Estimated class conditional probability density functions of cosegmentation statistics. From left to right: growth rate $\bar{g}$, average correlation $\bar{c}$, average uniqueness violation $\bar{u}$; from to top to bottom: for 10, 100, 1000 growth cycles μ. Correct correspondences (in red), incorrect correspondences (in blue). Note the gradual reduction in the overlap of the densities, especially in the two leftmost columns.

for incorrect as reported by [162], exceptions include e.g. situations when a correct correspondence lies on a narrow surface or in cases of partial occlusion. The average correlation in the region $\bar{c}$ is also typically higher for correct correspondences, but incorrect correspondences may accidentally have high correlation on repetitive or locally similar structures, especially on small regions. The average uniqueness violation $\bar{u}$ (deviation from bijective matching) when growing the region is also quite discriminative, see Fig. 6.5, right column. Forbidding uniqueness violation as in [93] is not suitable in the wide-baseline setup due to possibly high surface slant or scale changes. The statistics returned by the growing algorithm are combined with the ratio of the first to second closest distance of SIFT descriptors s_r [95], a standard method.

The problem of estimating a high dimensional likelihood ratio is avoided by projecting the four dimensional statistic into a 1D scalar quality $q_i = \mathrm{f}(s_r, \bar{g}_i, \bar{c}_i, \bar{u}_i)$ which expresses a confidence on correctness of the correspondence. This is done using the Support Vector Machine (SVM) trained on a set of exemplar positive and negative correspondences, see Sec. 6.4. The SVM finds a discriminative direction maximizing a margin in combination with a hinge loss (the training data are not separable). Projecting on this direction is an effective feature extraction procedure, suggested already by Vapnik [161] and popularized by e.g. Platt [122]

In consecutive decision stages i, the statistics are more discriminative, with the increase in the maximum number of growth steps μ_i, Step 3.2 of Alg. 3. Thus a different SVM θ_i is trained for each decision stage i. For subsequent stages, the classification error due to the overlap of probability distributions is decreasing, see plot in Fig. 6.9(b).

The likelihoods $\mathbf{p}_i(q|+1)$ and $\mathbf{p}_i(q|-1)$ of positive and negative class respectively were estimated by Parzen window method with a moving average kernel. The likelihoods estimated from our training set (see later) for four decision stages i are shown in Fig. 6.6. In the first stage, there is no growth and the statistic is solely the SIFT ratio. Interestingly, the SIFT ratio threshold of 0.8 suggested for accepting a correspondence by Lowe [95] is confirmed, being close to the equal-error operating point. Note that in the sequential process a significantly stricter test is applied in the first stage: only correspondences having SIFT ratio smaller than 0.4 are immediately accepted as correct, while the others are grown and decided in a later stage of a cascade.

The likelihood ratio L_i given the SVM output q_i is computed from linearly interpolated likelihood estimates.

6.2.3 Wald's sequential decision

Let x be an object belonging to one of two classes $\{-1, +1\}$. In our case, the classified objects are correspondences and the classes are "correct" (1) and "incorrect" (-1). Next, let an ordering on the set of measurements $\{x_1, \ldots, x_n\}$ on x be given. Here measurements x_i are scalar values, oriented distances from SVM decision boundaries after growing step i.

A sequential decision strategy is a set of decision functions $S = \{S_1, \ldots, S_n\}$, where $S_i : \{x_1, \ldots, x_i\} \to \{-1, +1, \sharp\}$. The strategy S makes one measurement at a time. The

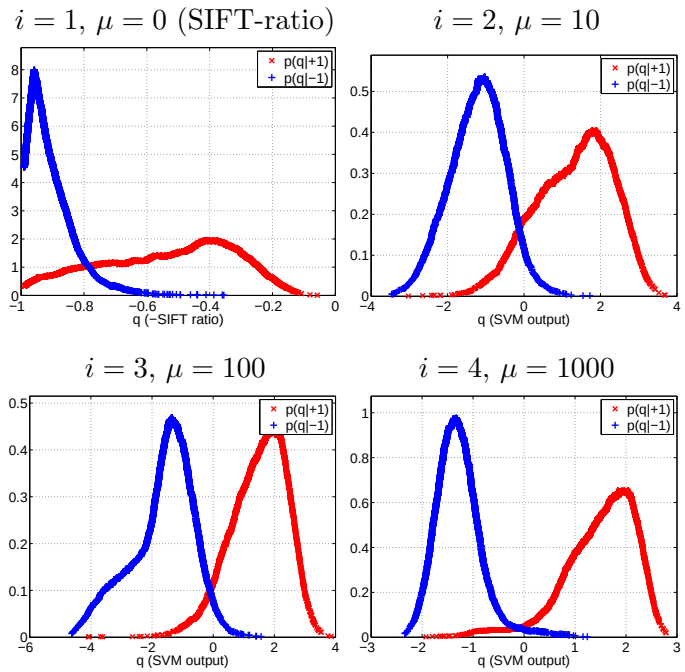


Figure 6.6: Estimated class conditional probability densities for the oriented distance to the SVM hyperplane, ie. for projections on a normal to the maximum margin hyperplane, for correct (red) and incorrect (blue) correspondences, for four stages of the sequential decision process. Note the decrease in the overlap of the distributions with increasing μ.

'♯' sign stands for a "continue" (do not decide yet). If a decision is '♯', x_{i+1} is obtained and S_{i+1} is evaluated. Otherwise, the output of S is the class returned by S_i.

In two-class classification problems, errors of two kinds can be made by strategy S. Let us denote α_S the probability of rejecting a correct correspondence (x belongs to $+1$ but is classified as -1) and β_S the probability of accepting an incorrect correspondence (x belongs to -1 but is classified as $+1$). A sequential strategy S is characterized by its error rates α_S and β_S and its average evaluation time $\bar{T}_S = E(T_S(x))$ where the expectation is over $p(x)$, and $\bar{T}_S$ is the expected evaluation time (or time-to-decision) for strategy S. An optimal strategy for the sequential decision making problem is then defined as

$$\begin{aligned} S^* = \arg\min_S \bar{T}_S \\ \text{s.t.} \quad \beta_S &\leq \beta, \\ \alpha_S &\leq \alpha \end{aligned} \tag{6.4}$$

for specified α and β.

Wald [167] proved that the solution of the optimization problem (6.4) is the *sequential probability ratio test.*

Sequential Probability Ratio Test (SPRT) Let x be an object characterized by its hidden state (class) $y \in \{-1, +1\}$. The decision about the hidden state is based on successive measurements $x_1, x_2, \ldots$. Let the joint conditional density $p(x_1, \ldots, x_m|y = c)$ of the measurements $x_1, ..., x_m$ be known for $c \in \{-1, +1\}$.

SPRT is a sequential strategy S^*, which is defined as

$$S_m^* = \begin{cases} +1, & L_m \geq A \\ -1, & L_m \leq B \\ \sharp, & B < L_m < A \end{cases} \tag{6.5}$$

where L_m is the likelihood ratio

$$L_m = \frac{p(x_1, ..., x_m|y = -1)}{p(x_1, ..., x_m|y = +1)}. \tag{6.6}$$

The constants A and B are set according to the required error of the first kind α and error of the second kind β. Optimal A and B are difficult to compute in practice, but tight bounds are easily derived. It can be shown that setting the thresholds A and B to

$$A = \frac{1-\beta}{\alpha}, \qquad B = \frac{\beta}{1-\alpha} \tag{6.7}$$

is close to optimal.

In the SCV algorithm, we assume that all information about a correspondence is contained in the statistics from the last growth step: $p(q_i|y) = p(q_1, \ldots, q_i|y)$. Therefore only 1D PDFs are needed to carry out the SPRT test. Estimation of scalar PDFs poses no technical problems as discussed in the previous section.

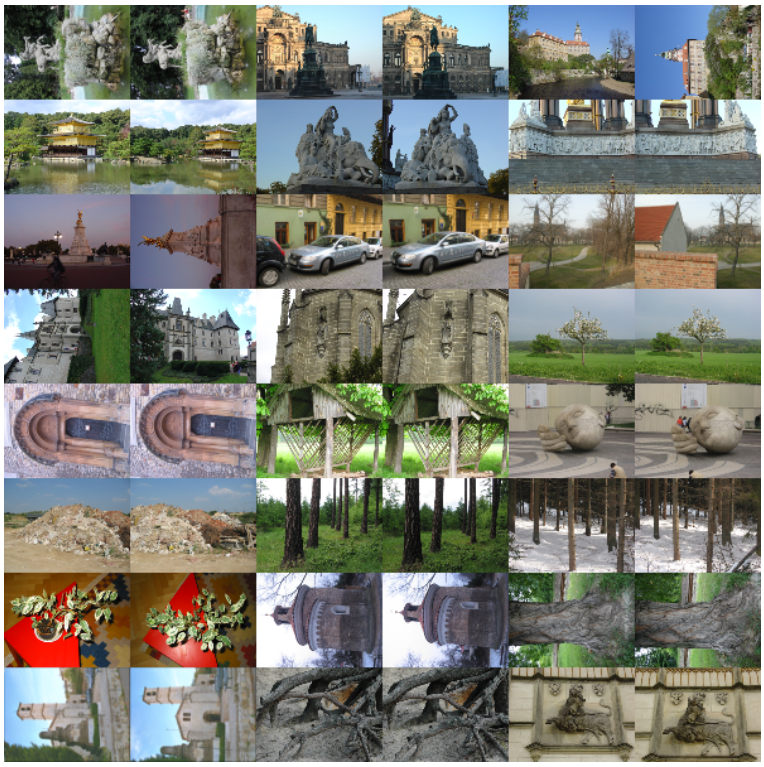


Figure 6.7: The set of training images.

6.3 The training procedure

The set of 24 image pairs used in a training set of correspondences is shown in Fig. 6.7. For all image pairs, MSERs were detected, LAFs constructed [99, 113] and SIFT descriptors computed on normalized patches. Standard wide-baseline matching was performed using SIFTs and a set of tentative correspondences was obtained. Finally, RANSAC was run on each pair of the set to estimate the epipolar geometry. We have manually relabeled correspondences which were accidentally consistent with the epipolar geometry but were in fact incorrect[4]. The remaining inlier correspondences formed the positive subset of the training set, all other correspondences were inserted as negative examples.

This way, approximately 6200 positive and 9800 negative correspondence examples were obtained, which means that tentative correspondences on the training set had on average approximately 40% of inliers.

The ground truth set was split randomly into two equals parts, half for training, half for testing. The learning stage included linear SVM training and probability density estimation via Parzen windowing. For SVM learning, the publicly available Statistical Pattern Recognition Toolbox [45] was used.

[4] A mismatch in tentative correspondence lying on a corresponding pair of epipolar lines cannot be detected by RANSAC.

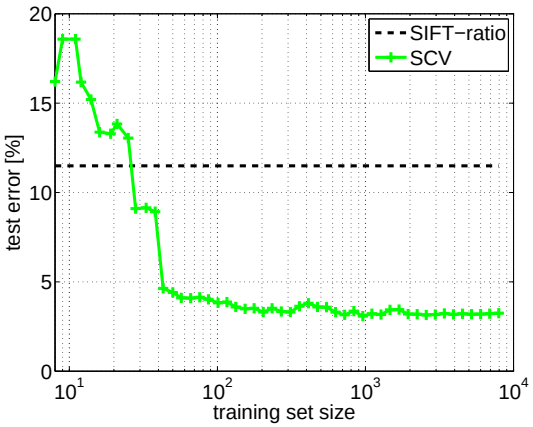


Figure 6.8: Test error as a function of a training set size. Note that only a very small training set with about 10^2 examples is required.

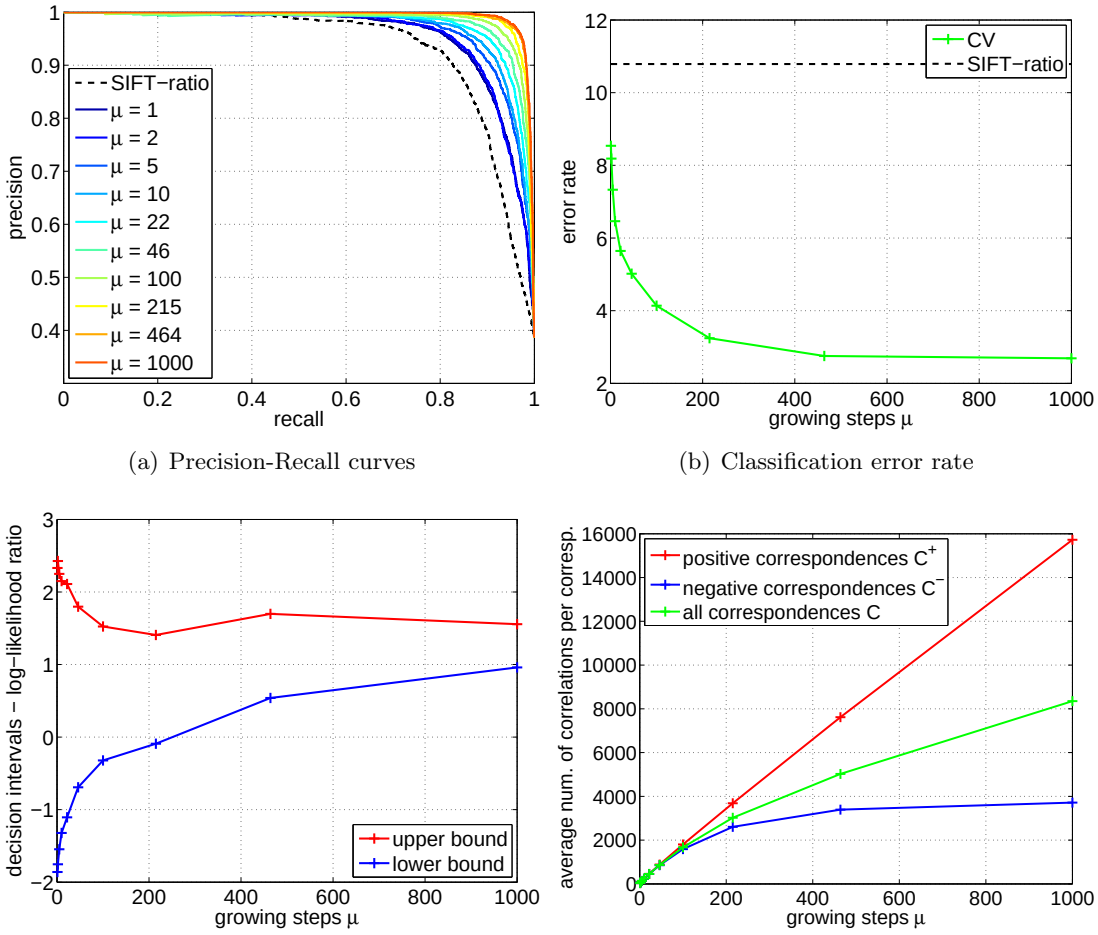
A priori, we had no idea about the necessary size of a training set for the correspondence classification problem. We therefore carried out the following test. A progressively larger portion of the training set was used to estimate the SVMs and likelihoods. The resulting sequential classifier (SCV) was applied to the test set. The observed classification error is plotted as a function of the training set size in Fig. 6.8. For reference, the classification error of the SIFT ratio is plotted too. The error does not improve significantly after about 200 samples, a surprisingly small number. We concluded that the size of our training set is sufficient.

Note that the Wald SPRT is a non-Bayesian technique based on conditional probabilities and its performance is guaranteed in terms of false positive and false negative rates hold for arbitrary prior probabilities. The insensitivity to the prior probability of (in)correct correspondence is an important property of the method, since a wide range of inlier ratios is encountered in practical matching problems. In fact, the SCV procedure is extremely useful for matching problems where tentative correspondences have a very low inlier ratio and direct RANSAC application would require an astronomical number of samples. Such problems differ significantly in terms of the inlier percentage in tentative correspondences from our training set, but in a non-Bayesian setting it does not matter. The training set must only be representative for the conditional probabilities of observations.

6.4 Experiments

6.4.1 Basic properties of the Sequential Correspondence Verification algorithm

We start the performance evaluation of the SCV algorithm by several experiments demonstrating some elementary properties of the algorithm. First, we measured dis-



(a) Precision-Recall curves

(b) Classification error rate

(c) Upper and lower bound of decision intervals in Log-likelihood scale

(d) Average number of computed correlations per correspondence

Figure 6.9: Properties of the SCV algorithm as a function of the number of growth steps μ.

criminability of the SCV algorithm, i.e. its ability to distinguish correct and incorrect correspondences. The discriminability is characterized by a precision-recall curve which is computed as follows. The SCV algorithm assigns likelihood ratio L to all N correspondences in test set. The correspondences are sorted according to their likelihood ratio, $L_{(1)} \geq L_{(2)} \geq \ldots \geq L_{(N)}$. *Precision* is defined as Q_n^+/n, where Q_n^+ is the number of correct correspondences among $L_{(1)}, \ldots, L_{(n)}$. *Recall* is defined as Q_n^+/Q_N^+. The SCV algorithm is more discriminative than a standard ratio of SIFT descriptors, and the difference becomes more prominent with the number of growing steps μ, see Fig. 6.9(a).

When a hard decision on the correctness of a correspondence is required, the likelihood ratio L is thresholded. The classification error rate for the threshold $L = 0$ is plotted in Fig. 6.9(b). In the case of the SIFT-ratio the threshold is 0.8 as discussed before.

The next figure, Fig. 6.9(c), focuses on decision thresholds for Wald's SPRT; the upper and lower bounds of the indecision intervals for Wald's SPRT are plotted for $\alpha = 0.05$ and $\beta = 0.001$ in log-likelihood ratio scale. The undecided interval is shrinking with increasing growth steps μ due to lower error.

Finally, Fig. 6.9(d) shows the average number of computed MNCC correlations per correspondence. This quantity is closely related to the computational complexity of the algorithm. For correct correspondence, the number of correlations C^+ grows almost linearly with growth steps μ, while for the incorrect correspondences, the number of correlations C^- saturates at 4000. It means that for the largest growth $\mu = 1000$, the negative correspondences are about four times faster to decide. This behavior is expected, since the algorithm stops growing when there are no high correlating neighbors and typically finishes by exhausting the seed queue $\mathcal{S}$ before the maximum number of growing steps is reached, see Step 4.3 of Alg. 4.

6.4.2 The SCV efficiently increases discriminability

We show that the SCV algorithm is more discriminative than the SIFT ratio and that the sequential decision-making process speeds the algorithm significantly at the expense of a very small discriminability loss. The comparison of the SCV algorithm was carried out with various settings of Wald's SPRT parameters (α, β), shown in Fig. 6.10. The SCV algorithm outperforms the SIFT ratio for all three settings. The SCV-1 ($\alpha = 0.001, \beta = 0.001$) is the most strict setting which has the highest discriminability. The SCV-2 ($\alpha = 0.05, \beta = 0.001$) allows more false negatives, while the SCV-3 ($\alpha = 0.001, \beta = 0.05$) more false positives, but they both are more efficient in terms of number of window correlations they had to compute.

In Tab. 6.1, three (α, β) settings of SCV algorithm are compared with the non-sequential version (CV), which does not decide until the last stage performing maximally $\mu_4 = 1000$ growing steps. We measured the average number of window correlations per correspondence C which had to be computed, and the percentage d_i of correspondences decided in i-th stage of the algorithm.

These values differ for correct and incorrect correspondences, so besides the mean values d_i, C (which depends on the percentage of correct correspondences in the test

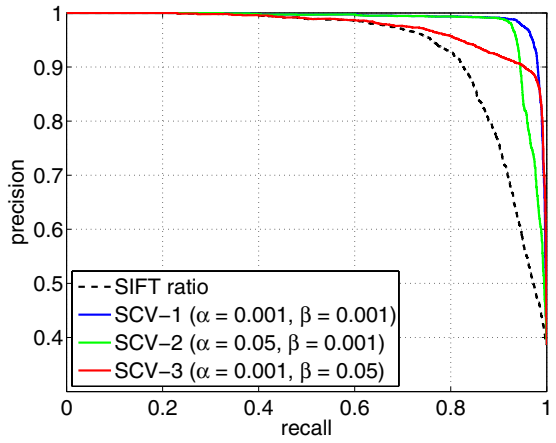


Figure 6.10: Discriminability of the SCV algorithm. The precision-recall curves for SCV with various setting of false positive and false negative rates and for the SIFT ratio alone.

all correspondences

	d_1	d_2	d_3	d_4	$C \times 10^3$
CV	0	0	0	100	8.6
SCV-1	9.7	12.7	9.5	68.1	4.1
SCV-2	9.7	63.1	16.4	10.8	2.1
SCV-3	33.6	9.5	4.2	52.8	2.0

correct correspondences only

	d_1^+	d_2^+	d_3^+	d_4^+	$C^+ \times 10^3$
CV	0	0	0	100	16.2
SCV-1	23.8	23.5	22.9	29.8	4.8
SCV-2	23.8	27.9	24.8	23.4	4.2
SCV-3	79.9	11.3	5.8	3.1	0.4

incorrect correspondences only

	d_1^-	d_2^-	d_3^-	d_4^-	$C^- \times 10^3$
CV	0	0	0	100	3.7
SCV-1	0.8	6.0	1.1	92.2	3.6
SCV-2	0.8	85.1	11.1	2.9	0.7
SCV-3	4.5	8.3	3.1	84.0	3.1

Table 6.1: Efficiency of the algorithm. The d_i is the percentage of correspondences decided in ith stage of the decision process. The C is an average number of window correlation per correspondence.

set), we show the tables for correct correspondences d_i^+, C^+ and wrong correspondences d_i^-, C^- which differ as discussed in the previous subsection.

We see that the sequential decision speeds up the process by factor of two (SCV-1) or more than four (SCV-2, SCV-3) in comparison to the non-sequential algorithm without losing much discriminability. The recall-precision curve in Fig. 6.10 of the non-sequential algorithm (CV) is almost identical to the SCV-1, therefore it is not shown. In the tables we can also verify that the SCV-2 having higher allowed false negative rate tends to decide negative correspondences in lower stages of the sequence speeding up the decision process by factor of more than 5, while the SCV-3 vice-versa, speeding up the decision process of positive correspondences by factor of 12.

Computational complexity The dominant operation in the SCV algorithm is correlation computation, other steps (SVM classification, Wald's SPRT) are negligible. The running time depends on the number of correspondences. Considering an example 1000 tentative correspondences, each requiring on average $C = 2100$ correlations (see Tab. 6.1) we end up with approximately 2×10^6 correlations per image pair. This can be computed on recent CPU in about 0.5 seconds and about 20–100 times faster on a modern GPU.

6.4.3 SCV performance on Hessian affine points

Until now, all experiments have been carried out on correspondences of local affine frames on MSERs. We now show that SCV-algorithm performs equally well for verification of correspondences obtained from Hessian affine points [134].

For all training image pairs in Fig. 6.7, a set of tentative correspondences was generated from Hessian affine points and classified according to the ground-truth epipolar geometry, and split into a training and test set. This is the same procedure as described before for MSERs.

Fig. 6.11 shows that for Hessian affine points, the SCV algorithm (SCV-1) improves the recall-precision curve obtained by SIFT ratio matching. Moreover, the performance is virtually equal for the two cases when the algorithm is trained specifically on Hessian-affine points or when the SCV algorithm trained on MSERs is used. This was not expected a priori, as the images patches around the respective correspondences are quite different. But the (simple) statistics of the growth process initialized from pixel correspondences are preserved.

Interestingly, the ratio of the SIFT descriptors has slightly better discriminability on Hessian affine points than on MSERs. The fact that MSERs are often detected on occlusion boundaries might play a role.

6.4.4 Challenging wide baseline stereo scenes

Results of correspondence selection on difficult wide baseline stereo scenes are shown in Fig. 6.12. These scenes are challenging due to small overlap, high degree of noise in the images (Raglan), complex 3D structure with many occlusions (Forsythia, Fence).

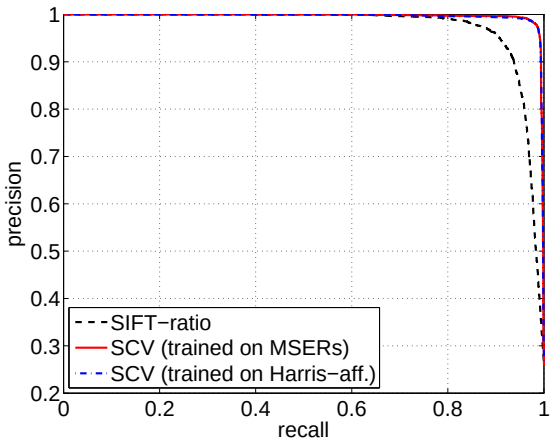


Figure 6.11: Discriminability of the SCV algorithm for Hessian affine correspondences. The SCV algorithm performs equally well when trained on MSER correspondences or specifically on correspondences of Hessian affine points.

In the Orange pair, matching is difficult since the background is locally similar (same grain of wood) but not the same (very different location on the same table). To find the epipolar geometry at all, the matching process generating the tentative correspondences had to be very permissive, so that a sufficient number of correct correspondences was present among tentative correspondences. We allowed more than one-to-one mapping in tentative correspondences which lead to a high number of outliers (about 90 percent).

Plots in the last column of Fig. 6.12 show the precision among the best n retrieved correspondences. This is important for progressive RANSAC procedure [28] which samples tentative correspondences according to preferences defined by the matching processes (approximately speaking in the order as sorted by the matcher of tentative correspondences). For correspondences sorted by the SCV algorithm, in all four scenes, the PROSAC procedure would terminate successfully after a single iteration, since a sufficient number of top correspondences is correct. This is neither the case when the ordering of tentative correspondences is given by the negative ratio of SIFT distances, nor the SIFT distances alone.

On the same images, we compared the sequential algorithm (SCV-2) and its non-sequential version (CV). For all the scenes, the results of SCV-2 are slightly worse than for SCV, which is much faster. The two algorithms evaluated the following numbers of MNCC correlations (SCV-2 and non-sequential CV respectively): 0.5×10^3 and 2.5×10^3 (Raglan), 0.6×10^3 and 5.7×10^3 (Forsythia), 1.2×10^3 and 6.3×10^3 (Orange),and 0.4×10^3 and 4.3×10^3 (Fence) The reason why the decision is even faster here than on the test set in the previous experiment is that there are many more wrong correspondences which are faster to decide and these outliers are quickly decidable in early stages of the sequence.

Omnidirectional images The method was successfully tested on challenging image pairs like Fig. 6.13, obtained by a catadioptric camera. Besides a significant spatial

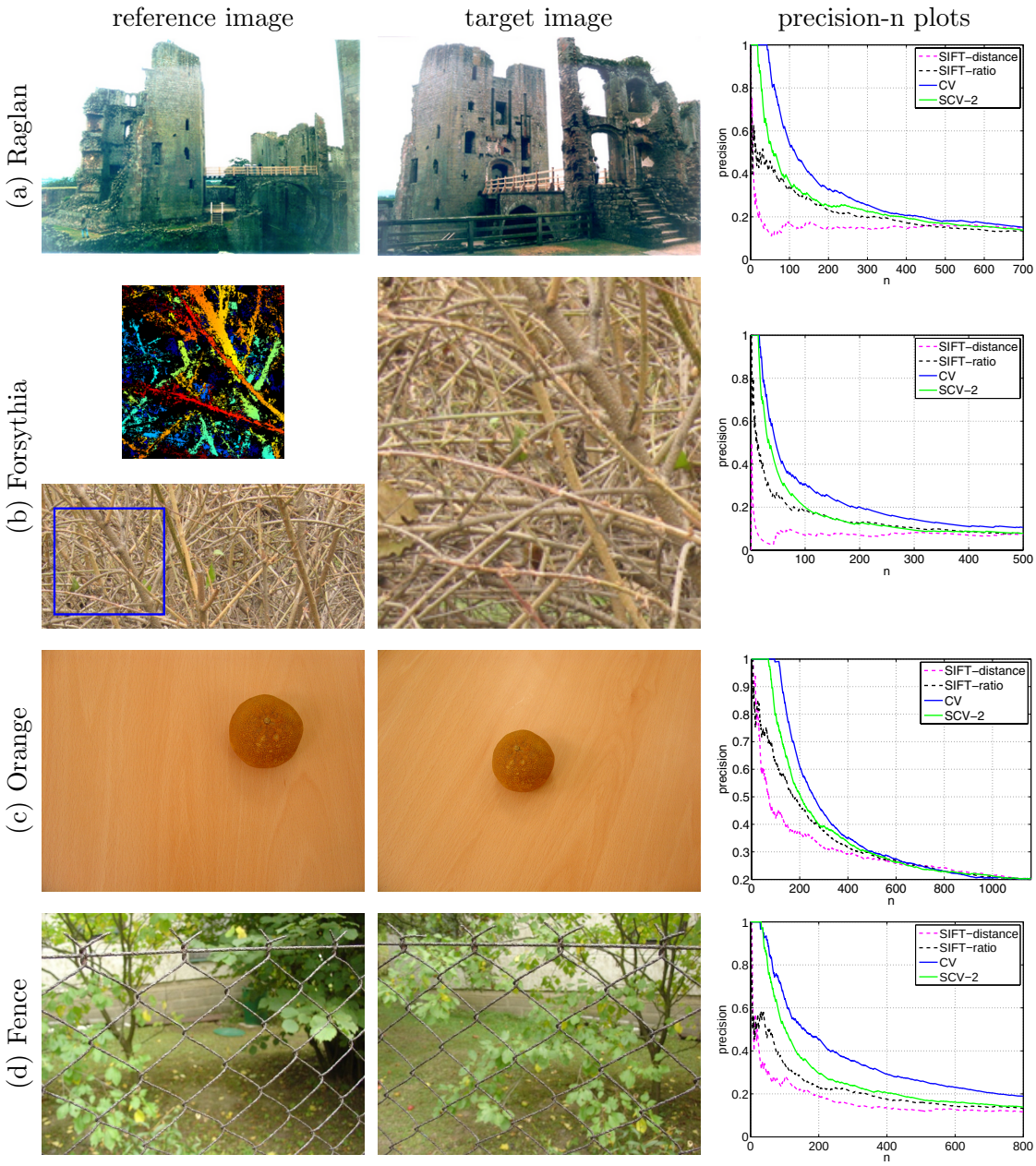


Figure 6.12: Results on challenging wide-baseline scenes. For Forsythia, we show the color coded depth map of a common part (inside the blue frame) to demonstrate the 3D structure of the image pair. Notice that the orange is placed in a different place on the table and no correct correspondence exist on the background. Although not obvious, almost the same part of the fence appears in both images of the Fence scene.

wide baseline setup, the pair has a wide temporal baseline: the first image was captured in the winter when there were no leaves on trees, while the other in the summer at different day-time and in very different lighting conditions and shadows. Despite the difficult conditions, the SCV algorithm was able to find several correct correspondences, as can be checked visually; the ground-truth does not exist in this case. The algorithm selected four LAF correspondences (three correct, one incorrect) out of more than 2000 tentative correspondences. Note that the correspondences shown are selected solely by the SCV algorithm, i.e. before robust model fitting.

The method works because the initial affine transformation obtained from LAF correspondence locally approximates the non-linear neighbourhood deformation in omnidirectional images. This is true for the central part of the images, while the problems occur at the boundary of spheres, where the distortion is not negligible. This is probably the reason for the incorrect correspondence which occurred close to boundary of the sphere.

6.4.5 Test on the Oxford dataset

We used a subset of the Oxford dataset[5] which has been used for performance evaluation of affine region detectors [104] and local descriptors [103]. The dataset consists of images which are distorted by various degradation: projective distortion (due to change of the camera position), image blur (from defocusing), JPEG compression artifacts and illumination changes. The ground-truth correspondences are known, since the database contains a homography mapping between the reference and distorted target images.

The input is a set of several hundreds tentative correspondences per each pair. The results for correspondence selection based on standard SIFT-ratio and on the SCV algorithm are shown in Fig. 6.14 as precision-recall curves. We can see the SCV algorithm is better in all cases. The most difficult distortion seems to be the blur, but it is deleterious for SIFT as well. The projective distortion is well captured by local affine transformation, the illumination change does not also make serious problems, since the MNCC statistic used in the growing algorithm is insensitive (however not fully invariant like NCC statistic) to affine illumination changes. Surprisingly, the SCV does not deteriorates that much with the JPEG compression. Although the images looks seriously (unnaturally) corrupted, the frequencies preserved by the JPEG compression resulted in enough correlation.

6.4.6 Image retrieval

We show the benefits of SCV algorithm in a large scale image retrieval setup, using the data set from Nistér and Stewénius benchmark [111]. It consists of 10200 images in groups of four that show the same object. Each image is retrieved from the whole data set. For each query the top N images are returned, and the score counting how many of the correct answers are in top K is computed. In the benchmark K is set to 4, giving the highest score 4, if the algorithm manages to retrieve top four images that matches

[5] `http://www.robots.ox.ac.uk/~vgg/research/affine/`

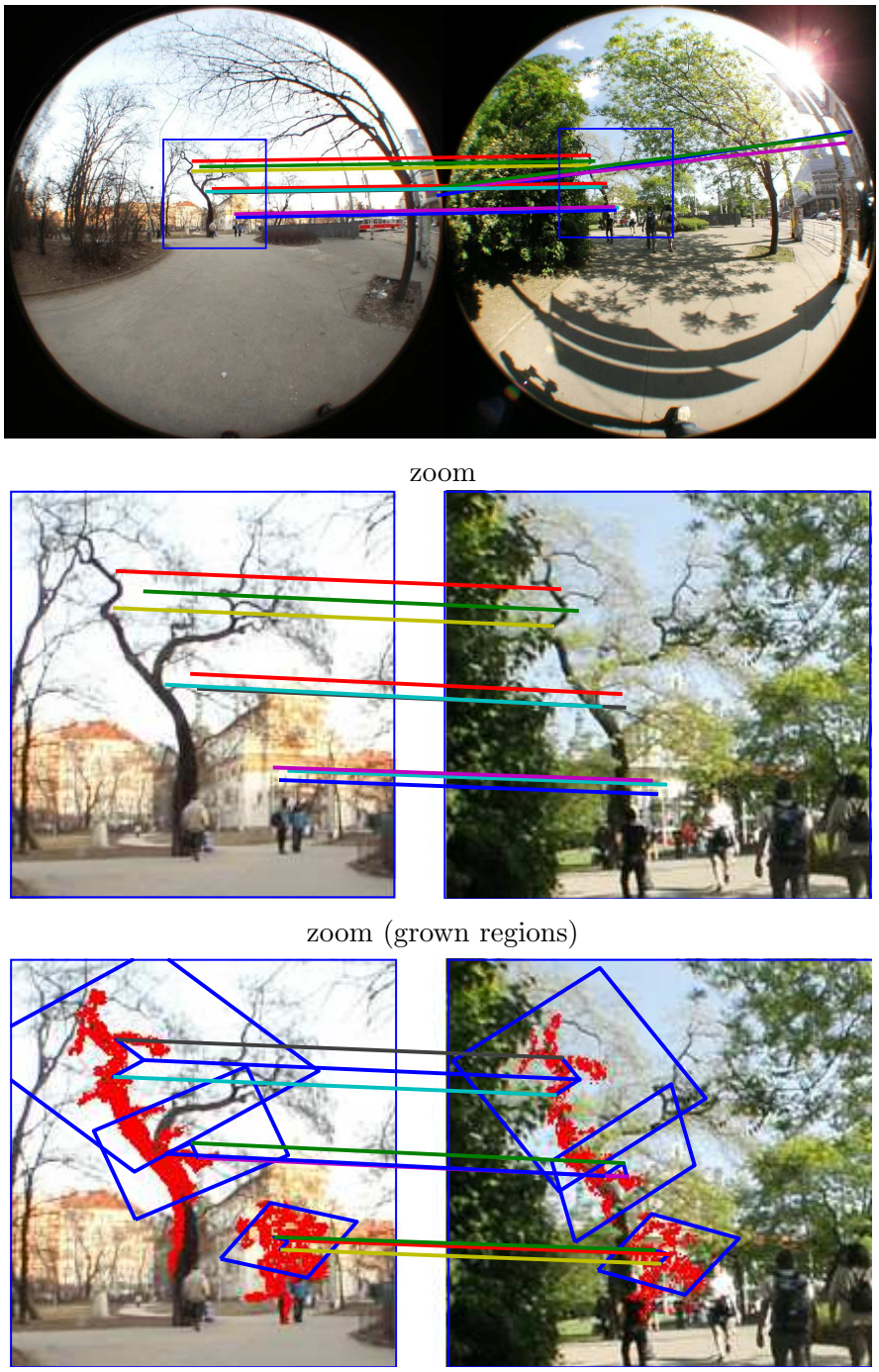


Figure 6.13: Correspondences found by the SCV algorithm in challenging omnidirectional image pair with a 'wide temporal baseline'. *Raw images by courtesy of Jan Knopp.*

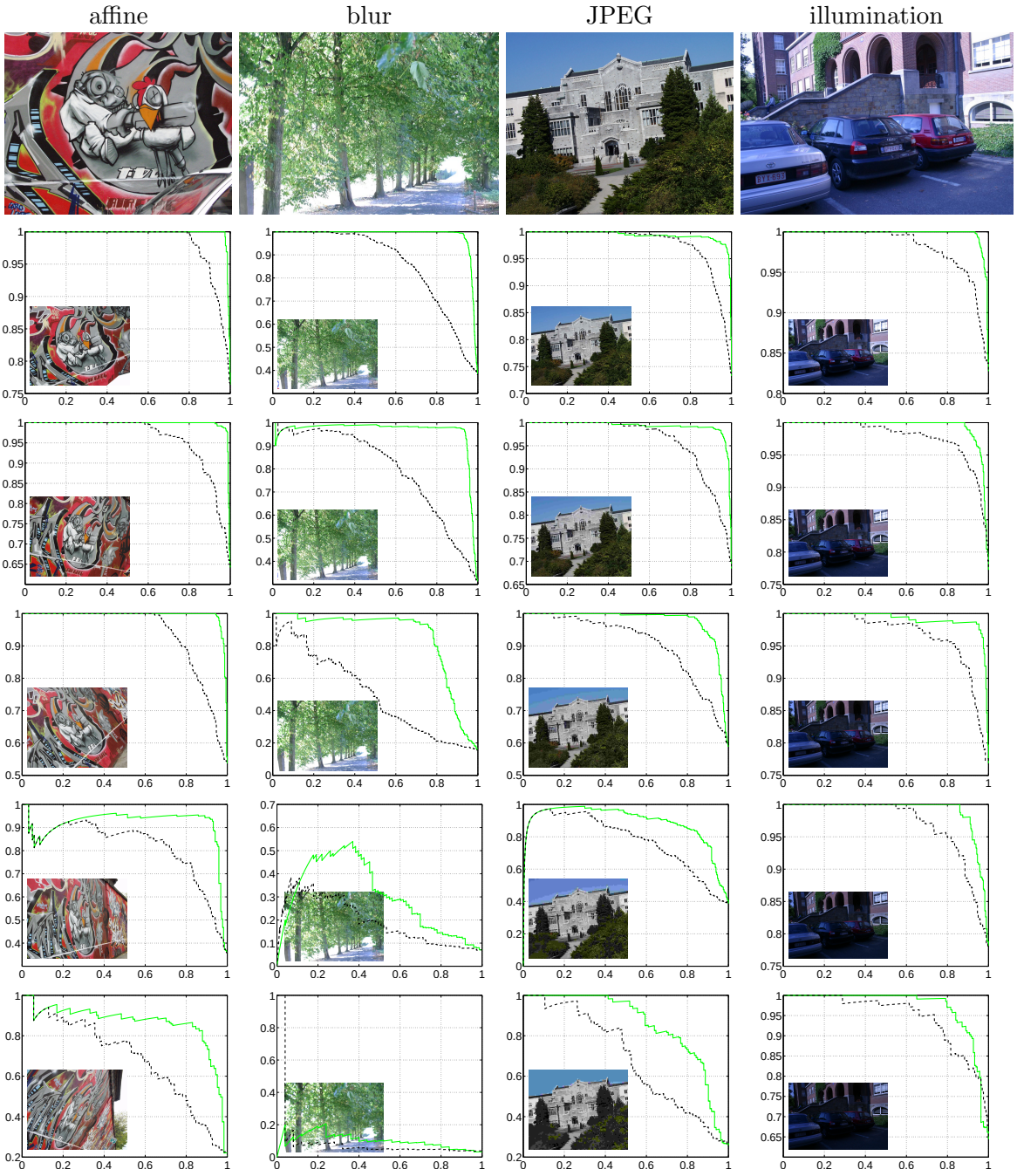


Figure 6.14: Results on the Oxford dataset. Precision-recall curves of the SCV algorithm (green) and of the SIFT ratio (black dashed), for images with an increasing degree of distortion. Reference images in the first row, target images are shown in graphs. See the electronic version to better view of the distorted images.

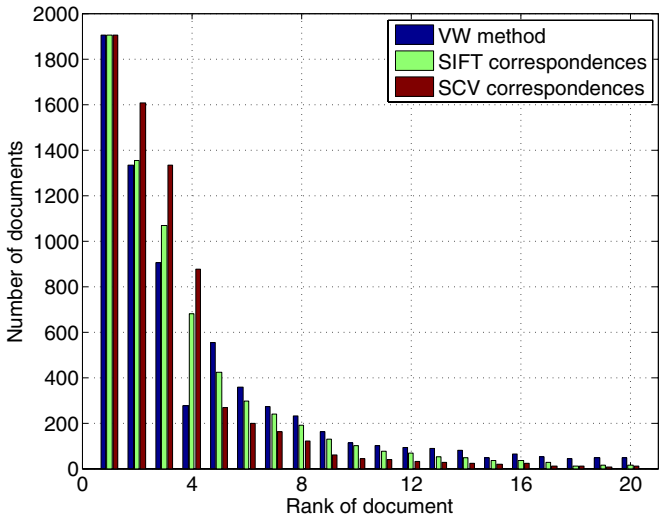


Figure 6.15: Ranking of documents based on the visual word method, the number of SIFT correspondences with distance ratio <0.8 and the number of SCV correspondences.

four instances of the object in the data set. Since the query image is also present in the data set, the worst score of algorithm returning only the query in top K is 1. The overall performance of the algorithm is computed as the average score of all 10200 queries from the data set.

A part of the solution proposed by Nistér was reimplemented. The MSERs [99] and LAFs [113] were computed on each of the images. Each of approximately 7 millions LAFs was described using SIFT descriptor [95] computed on an affine normalized patch. Then, similarly to visual words approach proposed by Šivic and Zisserman [143], we build visual words vocabulary consisting of 1 million k-means in the SIFT descriptor space and assign all the descriptors in the images to the nearest visual word. Each visual word in a given document is weighted using TFIDF (Term Frequency – Inverse Document Frequency) measure from text retrieval. The similarity of two documents is then the L_1 distance between their vectors of the visual words weights. The top K most similar documents are retrieved. The described approach is similar to the flat scoring of Nistér and Stewénius. It achieves the average score of 3.40 images retrieved per query on the whole dataset.

To evaluate the performance of the SCV algorithm we took all queries that can be improved by verifying reasonable number of images retrieved by VW method, i.e. queries where there is at least one image of the retrieved object with rank 5 to 20. There are 1904 such query images. The overall score, the average number of correct images among top 4, achieved by TFIDF visual word ranking is 2.32 for these queries. Top 20 score, i.e. the average number of correct images among top 20, is 3.54. This is the upper bound of the performance for a retrieval algorithm that resorts the top 20 retrieved images.

score	higher	same	lower
SIFT	750	926	228
SCV	1224	597	83

Table 6.2: Comparison of the scores of the VW method and rankings based on SIFT and SCV correspondences.

For comparison, tentative correspondences were computed as nearest SIFT descriptions for each of 1904 query images and its top 20 retrieved images giving altogether 38080 pairs. Tentative correspondences of each pair were then verified using the SCV ($\alpha = 0.02, \beta = 0.001$) algorithm. Finally, new ranking was established according to the number of SCV correspondences found in each pair of images. We also compared our method to the ranking based on SIFT correspondences (rank is based on the number of correspondences with SIFT distance ratio < 0.8). The performance of the SCV algorithm is compared in a histogram of ranks of the four correct images in answer to each query (see Fig. 6.15). Clearly, SCV significantly improves the ranking of the correct images bringing most of them to top 4. Its overall top 4 score on the 1904 query images is 5717 resulting in average 3.00, the average top 5 score is 3.14. The overall top 4 score for SIFT correspondences is 5004 resulting in average 2.63 and the average top 5 score is 2.85.

At last, we compared the achieved top 4 scores of both methods to the visual words method in Tab. 6.2. It shows the ranking is improved or unchanged with SCV in 95% of cases. Wrong ranking occurs typically for images of different objects with little texture (usually slightly blurred) on the same structured background. In this case, the most of, in fact correct, correspondences are found in the background which does not help retrieving a correct image.

6.5 Conclusions

We have presented a method which is able to efficiently distinguish correct and incorrect correspondences, via collecting statistics while cosegmenting gradually larger regions. We have shown this significantly benefits the matching process in challenging wide baseline scenes and improves results even in a large scale image retrieval. The process is computationally efficient and in practice requires only a fraction of a second.

7 Conclusions

The first part of the thesis was dealing with low-level processing of image signals. We proposed a complex correlation statistic which possesses the invariance to image discretization and insensitivity to affine transformation of corresponding domains. The complex correlation statistic naturally provides an estimate of its maximum position with sub-pixel precision.

The second part of thesis was more algorithmic. We designed an efficient dense matching algorithm suitable for matching high resolution images of complex 3D scenes. It avoids visiting the entire disparity space and computing huge number of correlation statistics by a sampling strategy. Only promising correspondence hypotheses are generated via growing initial correspondence seeds, which can be even random. Final decision is performed by a robust matching of competing correspondence hypotheses. The proposed algorithm keeps the robust properties (low error rate, sufficient density) of the original algorithm which computes the disparity space exhaustively, however it runs about 100 times faster.

The last part of the thesis originated from an inspiration by the success of the growing principle. This time we employed a similar growing mechanism to a problem of correspondence verification. We designed a procedure which estimates how likely a given correspondence is correct and how likely it is a mismatch. The algorithm, driven by Wald's sequential decision process, successively expands potentially corresponding regions while collecting image statistics until the decision based on learned models. We showed the resulting correspondence selection procedure is very discriminative, outperforms state-of-the-art correspondence selection based on ratio of two closest SIFT descriptors. Its running time is negligible to time spent by matching tentative correspondences, which implies that it should be always used before RANSAC. We showed benefits in challenging wide-baseline stereo problems and in image retrieval.

Contributions of the thesis are summarized in Sec. 3.3.

A reader may be disappointed that we did not show a connection of the complex correlation statistics with growing mechanisms described later. There are couple of reasons for that. Actually, the introduction of the GCS matching algorithm emerged from a necessity to have a fast stereo algorithm suitable for using the complex correlation statistic. Computing the entire disparity space of CCS was very computationally intensive, mainly because of inapplicability of sliding window-like methods as in windowed

statistics. Therefore, we were looking for an algorithm which avoids it. This way we discovered the GCS algorithm.

However, after performing several experiments with high resolution images of complex 3D scenes, using CCS and MNCC statistics, we observed that benefits of CCS are lost due to its artifacts. Using MNCC statistic was still significantly faster than CCS, due to simpler formula and simpler arithmetic. The spatial extent of CCS was too large which made problems at occlusion boundaries, stronger than those using windows of MNCC. We also observed that discretization artifacts are not so strong when using images of high-resolution. This is most probably due to the fact that camera anti-aliasing filters filter too high-frequencies out. The problem with modulo 0.5 px disparities, which manifested as contours of unassigned disparities when using MNCC, was solved alternatively by redefining the occlusion model with inhibition zone gap. The affine insensitivity is also replaced by a covariant approach used in the last chapter. The distortion of matching windows due to surface slant is captured by the local affine parameters of the correspondence seed, which can be viewed as a local fronto-parallel rectification. Sub-pixel estimates using CCS are applicable off-line, i.e. computing the CCS of correspondences according to a given integer disparity map. These estimates might be further refined by the proposed global sub-pixel disparity correction.

Although we experimentally showed that CCS keeps theoretical properties in practice, its artifacts are not negligible. The arguments above condemn the CCS to remain a theoretical concept only, without a practical applicability in matching high-resolution images of complex 3D scene. Other researchers might find this concept interesting and find a more suitable application for it.

Possible future extension of the methods We see a great potential of the growing algorithm. For instance, we can simply incorporate a discriminability enhancement by Kostková and Šára [83] aggregating correlation in disparity space.

The improvement can be in reducing the boundary artifacts. It seems straightforward to plug one of the methods mentioned in Sec. 2.1.2. The boundary artifacts are also reduced when fusing the disparity maps by Tyleček [160].

We observed in [181] that a trinocular matching significantly increases the discriminability and hereby the matching accuracy and density compared to a binocular matching. Modifying the binocular GCS algorithm to a trinocular version is not difficult and the resulting algorithm remains fast. In our 3D-reconstruction pipeline, we expect a considerable improvement when working with image triplets instead of pairs.

An important problem which occurs quite frequently in our (and apparently also in other similar) 3D reconstruction pipeline, is the inaccuracy of the calibration of cameras. We observed that even a moderate inaccuracy causes imprecise rectification which manifests by a sparsity of the matching. The reason is that the true epipolar lines are out of the image rows and the disparity has a vertical component in fact. The solution would be to relax the epipolar constraint to allow a small vertical disparity. This would give a dense set correspondences which would be usable for re-estimating or refining the camera calibration. We made first feasibility study with this, re-estimating the funda-

mental matrix of an image pair iteratively, and it works well. The rectification of several image pairs used in the thesis was refined by this method.

A possibility to further speed the algorithm up exists, in both algorithmic and implementation levels. For instance, computing a complete MNCC statistic to decide a future seed (of the best correlation) is unnecessarily expensive, since for such a decision a simpler test most probably exists. It would also make sense to implement computing neighbouring correlations à la 'sliding window'. A revision of data structures for representing the hypotheses in the disparity space is worthy. After all, a clever implementation in low-level programming language or on a modern GPU might move the algorithm closer to real-time. Current implementation is still partly in Matlab.

Another issue is a refinement of the estimated triangulated surface. We have been using an iterative mesh fairing by Kostlivá et al. [85]. This method takes only the mesh as its input. Therefore we consider incorporating a data term similarly to the proposed method of the global sub-pixel disparity correction, which should further improve an accuracy of the reconstructed 3D model.

Bibliography

[1] A. Akbarzadeh, J.-M. Frahm, P. Mordohai, B. Clipp, C. Engels, D. Gallup, P. Merrell, M. Phelps, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewenius, R. Yang, G. Welch, H. Towles, D. Nistér, and M. Pollefeys. Towards urban 3D reconstruction from video. In *Proc. 3D Data Processing, Visualization and Transmission*, 2006.

[2] L. Alvarez, R. Deriche, J. Sánchez, and J. Weickert. Dense disparity map estimation respecting image derivatives: a PDE and scale space based approach. *Journal of Visual Communication and Image Representation*, 13(1-2):3–21, 2002.

[3] B. Amberg, A. Blake, A. Fitzgibbon, S. Romdhani, and T. Vetter. Reconstructing high quality face-surfaces using model based stereo. In *Proc. ICCV*, 2007.

[4] N. Amenta, M. Bern, and D. Eppstein. The crust and the β-skeleton: Combinatorial curve reconstruction. *Graphical Models and Image Processing*, 60(2):125–135, 1988.

[5] S. Baker, T. Sim, and T. Kanade. A characterization of inherent stereo ambiguities. In *Proc. ICCV*, volume 1, pages 428–435, 2001.

[6] O. Barinova, V. Konushin, A. Yakubenko, H. Lim, and A. Konushin. Fast automatic single-view 3-d reconstruction of urban scenes. In *Proc. ECCV*, LNCS 5303, pages 100–113, 2008.

[7] R. Baxter. *Exactly Solved Models in Statistical Mechanics*. Academic Press, New York, 1990.

[8] P. N. Belhumeur. A Bayesian approach to binocular stereopsis. *IJCV*, 19(3):237–260, 1996.

[9] D. N. Bhat and S. K. Nayar. Ordinal measures for visual correspondence. In *Proc. CVPR*, pages 351–357, 1996.

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

[11] S. Birchfield and C. Tomasi. Multiway cut for stereo and motion with slanted surfaces. In *Proc. ICCV*, volume 1, pages 489–495, 1999.

[12] M. Bleyer and M. Gelautz. Graph-cut-based stereo matching using image segmentation with symmetrical treatment of occlusions. *Signal Processing: Image Communication, Special issue on three-dimensional video and television*, 22(2):127–143, 2007.

[13] M. Bleyer and M. Gelautz. Simple but effective tree structures for dynamic programming-based stereo matching. In *Proc. Internation Conference on Computer Vision Theory and Applications (VISAPP)*, pages 415–422, 2008.

[14] A. F. Bobick and S. S. Intille. Large occlusion stereo. *IJCV*, 33(3):181–200, 1999.

[15] R. C. Bolles, H. H. Baker, and M. J. Hannah. The JISCT stereo evaluation. In *Proc. DARPA Image Understanding Workshop*, pages 263–274, 1993.

[16] Y. Boykov, O. Veksler, and R. Zabih. Disparity component matching for visual correspondence. In *Proc. CVPR*, pages 470–475, 1997.

[17] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *Proc. ICCV*, volume 1, pages 377–384, 1999.

[18] D. Brewster. *The stereoscope; its history, theory, and construction, with its application to the fine and useful arts and to education.* J. Murray, London, 1856.

[19] M. Z. Brown and D. Burschka. Advances in computational stereo. *IEEE Trans. on PAMI*, 25(8):993–1008, 2003.

[20] C. Buehler, S. J. Gortler, and L. McMillan. Minimal surfaces for stereo. In *Proc. ECCV*, LNCS 2352, pages 885–899, 2002.

[21] R. Burtch. Short history of photogrammetry, 2008. Materials for lectures.

[22] J. Čech. Towards accurate stereoscopic matching. Research Report CTU–CMP–2004–05, Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic, February 2004.

[23] J. Čech, J. Matas, and M. Perd'och. Efficient sequential correspondence selection by cosegmentation. In *Proc. CVPR*, 2008.

[24] J. Čech and R. Šára. Efficient algorithms for computing correlation tables in stereoscopic vision. Research Report CTU–CMP–2004–11, Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic, October 2004.

[25] J. Čech and R. Šára. Complex correlation statistic for dense stereoscopic matching. In *Proc. SCIA*, volume LNCS 3540, pages 598–608, 2005.

[26] J. Čech and R. Šára. Efficient sampling of disparity space for fast and accurate matching. In *Proc. CVPR Workshop Towards Benchmarking Automated Calibration, Orientation, and Surface Reconstruction from Images, BenCOS 2007*. IEEE Computer Society, 2007.

[27] Q. Chen and G. Medioni. A volumetric stereo matching method: Application to image-based modeling. In *Proc. CVPR*, pages 1029–1034, 1999.

[28] O. Chum and J. Matas. Matching with PROSAC - progressive sample consensus. In *Proc. CVPR*, pages 220–226, 2005.

[29] O. Chum, J. Philbin, J. Šivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. ICCV*, 2007.

[30] M. Clerc. Wavelet-based correlation for stereopsis. In *Proc. ECCV*, LNCS 2351, pages 459–509, 2002.

[31] H. Cornelius, R. Šára, D. Martinec, T. Pajdla, O. Chum, and J. Matas. Towards complete free-form reconstruction of complex 3D scenes from an unordered set of uncalibrated images. In *Proc. ECCV Workshop Statistical Methods in Video Processing*, LNCS 3247, pages 1–12, 2004.

[32] I. J. Cox, S. L. Hingorani, and S. B. Rao. A maximum likelihood stereo algorithm. *Computer Vision and Image Understanding*, 63(3):542–567, 1996.

[33] G. D. Theory of communication. *Journal of the IEE*, 93(26):429–439, 1946.

[34] R. Deriche, C. Bouvin, and O. Faugeras. A level-set approach for stereo. In *First Annual Symposium on Enabling Technologies for Law Enforcement and Security - SPIE Conference 2942: Investigative Image Processing*, 1996.

[35] F. Devernay and O. Faugeras. Computing differential properties of 3-D shapes from stereoscopic images without 3-D models. In *Proc. CVPR*, pages 208–213, 1994.

[36] U. Dhond and J. Aggarwal. Structure from stereo—a review. *IEEE Trans. on Systems, Man, and Cybernetics*, 19(6):1489–1510, 1989.

[37] O. Faugeras and R. Keriven. Complete dense stereovision using level set methods. In *Proc. ECCV*, LNCS 1406, pages 379–393, 1998.

[38] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *IJCV*, 70(1):41–54, 2006.

[39] V. Ferrari, T. Tuytelaars, and L. van Gool. Simultaneous object recognition and segmentation from single or multiple image views. *IJCV*, 67(2):159–188, 2006.

[40] M. A. Fishler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[41] B. Flach. A diffusion algorithm for decreasing energy of max-sum labeling problem. Fakultät Informatik, Technische Universität Dresden, Germany, 1998. Unpublished manuscript.

[42] B. Flach and M. I. Schlesinger. A class of solvable consistent labeling problems. In *Proc. IAPR Workshop on Advances in Pattern Recognition*, volume LNCS 1876, pages 652–658, 2000.

[43] D. J. Fleet and A. D. Jepson. Stability of phase information. *IEEE Trans. on PAMI*, 15(12):1253–1268, 1993.

[44] D. J. Fleet, A. D. Jepson, and M. R. M. Jenkin. Phase-based disparity measurement. *CVGIP: Image Understanding*, 53(2):198–210, 1991.

[45] V. Franc and V. Hlaváč. Statistical Pattern Recognition Toolbox for Matlab, 2007.

[46] T. Fröhlinghaus and J. M. Buhmann. Regularizing phase-based stereo. In *Proc. ICPR*, volume 1, pages 451–455, 1996.

[47] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. *IEEE Trans. on PAMI*. To appear.

[48] Y. Furukawa and J. Ponce. Accurate camera calibration from multi-view stereo and bundle adjustment. In *Proc. CVPR*, 2008.

[49] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *Proc. CVPR*, 2007.

[50] P. Gargallo and P. Sturm. Bayesian 3D modeling from images using multiple depth maps. In *Proc. CVPR*, 2005.

[51] G. L. Gimel'farb, V. B. Marchenko, and V. I. Rybak. Algorithm of automatic matching of identical patches in stereopairs. *Kibernetika*, (2):118–129, 1972. In Russian.

[52] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-view stereo for community photo collections. In *Proc. ICCV*, 2007.

[53] M. Gong and Y.-H. Yang. Fast stereo matching using reliability-based dynamic programming and consistency constraints. In *Proc. ICCV*, pages 610–617, 2003.

[54] M. Gong and Y.-H. Yang. Fast unambiguous stereo matching using reliability-based dynamic programming. *IEEE Trans. on PAMI*, 27(6):998–1003, June 2005.

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

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

[57] D. Gusfield and R. W. Irving. *The Stable Marriage Problem: Structure and Algorithms*. The MIT Press, 1989.

[58] M. Habbecke and L. Kobbelt. A surface-growing approach to multi-view stereo reconstruction. In *Proc. CVPR*, 2007.

[59] R. M. Haralick and L. G. Shapiro. Image segmentation techniques. *CVGIP*, 29(1):100–132, 1985.

[60] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK, 2000.

[61] Y. Hel-Or and P. C. Teo. A common framework for steerability, motion estimation and invariant feature detection. Technical Report STAN-CS-TN-96-28, Department of Computer Science, Standford University, Stanford, California 94035, January 1996.

[62] H. Hirschmüller. Improvements in real-time correlation-based stereo vision. In *Proc. Workshop on Stereo and Multi-Baseline Vision*, pages 141–148, Kauai, Hawaii, December 2001.

[63] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. on PAMI*, 30(2):328–341, 2008.

[64] H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Trans. on PAMI*, 2009. To appear.

[65] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *Proc. CVPR*, volume 2, pages 2137–2144, 2006.

[66] J. Jia, Y. Xu, W. Liu, C. Yang, Y. Zhu, X. Zhang, and L. An. A miniature stereo vision machine for real-time dense depth mapping. In *Proc. International Conference on Computer Vision Systems*, LNCS 2626, pages 268–277, 2003.

[67] D. G. Jones and J. Malik. Computational framework for determining stereo correspondence from a set of linear spatial filters. *IVC*, 10(10):699–708, 1992.

[68] B. Julesz. *Foundations of Cyclopean Perception*. The University of Chicago Press, 1971.

[69] G. Kamberov, G. Kamberova, O. Chum, Š. Obdržálek, D. Martinec, J. Kostková, T. Pajdla, J. Matas, and R. Šára. 3D geometry from uncalibrated images. In *Proc. International Symposium on Visual Computing*, LNCS 4292, pages 802–813, 2006.

[70] S. B. Kang, R. Szeliski, and J. Chai. Handling occlusions in dense multi-view stereo. In *Proc. CVPR*, pages 103–110, 2001.

[71] J. Kannala and S. S. Brandt. Quasi-dense wide baseline matching using match propagation. In *Proc. CVPR*, 2007.

[72] J. Kannala, E. Rahtu, S. S. Brandt, and J. Heikkilä. Object recognition and segmentation by non-rigid quasi-dense matching. In *Proc. CVPR*, 2008.

[73] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proc. Eurographics Symposium on Geometry Processing*, pages 61–70, 2006.

[74] T. Kim and J. Muller. Automated urban area building extraction from high resolution stereo imagery. *IVC*, 14:115–130, 1996.

[75] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. on PAMI*, 28(10):1568–1583, 2006.

[76] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *Proc. ICCV*, pages 508–515, 2001.

[77] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *Proc. ECCV*, LNCS 2352, pages 82–96, 2002.

[78] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts. *IEEE Trans. on PAMI*, 26(2):147–159, 2004.

[79] N. Komodakis, N. Paragios, and G. Tziritaz. MRF optimization via dual decomposition: Message passing revisited. In *Proc. ICCV*, 2007.

[80] A. Koschan. What is new in computational stereo since 1989: A survey on current stereo papers. research report 93-22, Technische Universität Berlin, 1993.

[81] J. Kostková, J. Čech, and R. Šára. The CMP evaluation of stereo algorithms. Research Report CTU–CMP–2003–01, Center for Machine Perception, K333 FEE, Czech Technical University, Prague, Czech Republic, January 2003.

[82] J. Kostková, J. Čech, and R. Šára. Dense stereomatching algorithm performance for view prediction and structure reconstruction. In *Proc. SCIA*, LNCS 2749, pages 101–107, 2003.

[83] J. Kostková and R. Šára. Stratified dense matching for stereopsis in complex scenes. In *Proc. BMVC*, volume 1, pages 339–348, 2003.

[84] J. Kostlivá, J. Čech, and R. Šára. Feasibility boundary in dense and semi-dense stereo matching. In *Proc. CVPR Workshop Towards Benchmarking Automated Calibration, Orientation, and Surface Reconstruction from Images, BenCOS 2007*. IEEE Computer Society, 2007. Best Paper Award.

[85] J. Kostlivá, R. Šára, and M. Matýsková. Fairing of discrete surfaces with boundary that preserves size and qualitative shape. In *Proc. International Symposium on Visual Computing*, LNCS 5358, pages 107–118, 2008.

[86] V. A. Kovalevsky and V. K. Koval. A diffusion algorithm for decreasing energy of max-sum labeling problem. Glushkov Institute of Cybernetics, Kiev, USSR, 1975. Unpublished manuscript.

[87] J. D. Krol and W. A. van de Grind. Rehabilitation of a classical notion of Panum's fusional area. *Perception*, 11(5):615–619, 1982.

[88] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. In *Proc. ICCV*, volume 1, pages 307–314, 1999.

[89] K. N. Kutulakos and S. M. Seitz. A theory of shape by shape carving. *IJCV*, 38(3):199–218, 2000.

[90] P. Labatut, J.-P. Pons, and R. Keriven. Efficient multi-view reconstruction of large-scale scenes using interest points, Delaunay triangulation and graph cuts. In *Proc. ICCV*, 2007.

[91] Y. G. Leclerc, Q.-T. Luong, and P. Fua. Measuring the self-consistency of stereo algorithms. In D. Vernon, editor, *Proc. ECCV*, volume LNCS 1843, pages 282–298, 2000.

[92] B. Leibe, K. Mikolajczyk, and B. Schiele. Efficient clustering and matching for object class recognition. In *Proc. BMVC*, 2006.

[93] M. Lhuillier and L. Quan. Match propagation for image-based modeling and rendering. *IEEE Trans. on PAMI*, 24(8):1140–1146, 2002.

[94] M. Lhuillier and L. Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE Trans. on PAMI*, 27(3):418–433, 2005.

[95] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[96] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman and Company, New York, 1982.

[97] D. Martinec. *Robust Multiview Reconstruction*. PhD thesis, Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic, 2008.

[98] D. Martinec and T. Pajdla. Robust rotation and translation estimation. In *Proc. CVPR*, 2007.

[99] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. BMVC*, pages 384–393, 2002.

[100] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *IVC*, 22(10):761–767, 2004.

[101] M. Matoušek. *Epipolar Rectification Minimising Image Loss*. PhD thesis, Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic, 2007.

[102] Z. Megyesi, G. Kós, and D. Chetverikov. Dense 3D reconstruction from images by normal aided matching. *Machine Graphics and Vision*, 15:3–28, 2006.

[103] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on PAMI*, 27(10):1615–1630, 2005.

[104] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. van Gool. A comparison of affine region detectors. *IJCV*, 65(1/2):43–72, 2005.

[105] H. P. Moravec. Towards automatic visual obstacle avoidance. In *Proc. International Joint Conference on Artificial Intelligence*, page 584, 1977.

[106] M. Mozerov. An effective stereo matching algorithm with optimal path cost aggregation. In *Proc. DAGM*, pages 617–626, 2006.

[107] J. Mulligan and K. Daniilidis. Trinocular stereo for non-parallel configurations. In *Proc. ICPR*, volume 1, pages 567–570, 2000.

[108] J. Mulligan, V. Isler, and K. Daniilidis. Trinocular stereo: a real-time algorithm and its evaluation. In *Proc. Workshop on Stereo and Multi-Baseline Vision*, pages 10–17, Kauai, Hawaii, December 2001.

[109] D. Nehab, S. Rusinkiewicz, and J. Davis. Improved sub-pixel stereo correspondences through symmetric refinement. In *Proc. ICCV*, 2005.

[110] F. Nicodemus, J. Richmond, J. Hsia, I. Ginsberg, and T. Limperis. *Geometric consideration and nomenclature for reflectance*. National Bureau of Standards, Washington D.C., USA, 1977.

[111] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proc. CVPR*, pages 2161–2168, 2006.

[112] T. Noguchi and Y. Ohta. A simple but high-quality stereo algorithm. In *Proc. ICPR*, volume 4, pages 351–354, 2002.

[113] S. Obdržálek and J. Matas. Sub-linear indexing for large scale object recognition. In *Proc. BMVC*, 2005.

[114] A. S. Ogale and A. Yiannis. Stereo correspondence with slanted surfaces: critical implications of horizontal slant. In *Proc. CVPR*, volume 1, pages 568–573, 2004.

[115] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Trans. on PAMI*, 15(4):353–363, 1993.

[116] M. Okutomi and Y. Katayama. A simple stereo algorithm to recover precise object boundaries and smooth surfaces. In *Proc. Workshop on Stereo and Multi-Baseline Vision*, pages 158–165, Kauai, Hawaii, December 2001.

[117] M. A. O'Neill and M. I. Denos. Practical approach to the stereo matching of urban imagery. *IVC*, 10(2):89–98, 1992.

[118] G. P. Otto and T. K. W. Chau. 'Region-growing' algorithm for matching of terrain images. *IVC*, 7(2):83–94, 1989.

[119] H. Pan and J. Magarey. Multiresolution phase-based bidirectional stereo matching with provision for discontinuity and occlusion. *Digital Signal Processing*, 8(4):255–266, 1998.

[120] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, 1988.

[121] J. Philbin, O. Chum, M. Isard, J. Šivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR*, 2007.

[122] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[123] 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.

[124] J.-P. Pons, R. Keriven, O. Faugeras, and G. Hermosillo. Variational stereovision and 3D scene flow estimation with statistical similarity measures. In *Proc. ICCV*, 2003.

[125] E. Z. Psarakis and G. D. Evangelidis. An enhanced correlation-based method for stereo correspondence with sub-pixel accuracy. In *Proc. ICCV*, volume 1, pages 907–912, 2005.

[126] L. Robert and R. Deriche. Dense depth map reconstruction: A minimization and regularization approach which preserves discontinuities. In *Proc. ECCV*, LNCS 1064, pages 439–451, 1996.

[127] T. D. Sanger. Stereo disparity computation using Gabor filters. *Biological Cybernetics*, 58(6):405–418, 1988.

[128] R. Šára. Sub-pixel disparity correction. Internal working paper 98/01, Center for Machine Perception, Czech Technical University, Prague, 1998.

[129] R. Šára. Finding the largest unambiguous component of stereo matching. In *Proc. ECCV*, LNCS 2352, pages 900–914, 2002.

[130] R. Šára. Robust correspondence recognition for computer vision. In *COMPSTAT 2006 - Proceedings in Computational Statistics*, pages 119–131. Physica-Verlag, 2006.

[131] R. Šára and R. Bajcsy. On occluding contour artifacts in stereo vision. In *Proc. CVPR*, pages 852–857, 1997.

[132] R. Šára and R. Bajcsy. Fish-scales: Representing fuzzy manifolds. In *Proc. ICCV*, pages 811–817, 1998.

[133] A. Saxena, S. H. Chung, and A. Y. Ng. 3-D depth reconstruction from a single still image. *IJCV*, 76(1):53–69, 2007.

[134] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In *Proc. ECCV*, LNCS 2350, pages 414–431, 2002.

[135] D. Scharstein and C. Pal. Learning conditional random fields for stereo. In *Proc. CVPR*, 2007.

[136] D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1):7–42, 2002. `http://vision.middlebury.edu/stereo/`.

[137] M. I. Schlesinger. *Mathematical Tools of Image Processing*. Naukova Dumka, Kiev, 1989. In Russian.

[138] M. I. Schlesinger and B. Flach. Analysis of optimal labelling problems and their application to image segmentation and binocular stereovision. In *East-West-Vision: International Workshop & Project Festival on Computer Vision, Computer Graphics, New Media*, pages 55–60. Austrian Computer Society, 2002.

[139] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. CVPR*, volume 1, pages 519–526, 2006. `http://vision.middlebury.edu/mview/`.

[140] M. Shimizu and M. Okutomi. Precise sub-pixel estimation on area-based matching. In *Proc. ICCV*, pages 90–97, 2001.

[141] M. Shimizu and M. Okutomi. Significance and attributes of subpixel estimation on area-based matching. *System and Computers in Japan*, 34(12):1791–1800, 2003.

[142] M. Shimizu and M. Okutomi. Multi-parameter simultaneous estimation on area-based matching. *IJCV*, 67(3):327–342, 2006.

[143] J. Šivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, pages 1470–1477, 2003.

[144] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. In *Proc. SIGGRAPH*, pages 835–846, 2006.

[145] C. V. Stewart, C.-L. Tsai, and B. Roysam. The dual-bootstrap iterative closest point algorithm with application to retinal image registration. *Medical Imaging*, 22(11):1379–1394, 2003.

[146] C. Strecha. Multi-view stereo as an inverse inference problem. PhD thesis, KU Leuven, Belgium, May 2007.

[147] C. Strecha, T. Tuytelaars, and L. Van Gool. Dense matching of multiple wide-baseline views. In *Proc. ICCV*, 2003.

[148] C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Proc. CVPR*, 2008. `http://cvlab.epfl.ch/~strecha/multiview/denseMVS.html`.

[149] J. Sun, H.-Y. Shum, and N.-N. Zheng. Stereo matching using belief propagation. In *Proc. ECCV*, LNCS 2351, pages 510–524, 2002.

[150] R. Szeliski and D. Scharstein. Symetric sub-pixel stereo matching. In *Proc. ECCV*, LNCS 2351, pages 525–540, 2002.

[151] R. Szeliski and D. Scharstein. Sampling the disparity space image. *IEEE Trans. on PAMI*, 25(3):419–425, 2004.

[152] 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. *IEEE Trans. on PAMI*, 30(6):1068–1080, 2008.

[153] F. Tappen and W. T. Freeman. Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *Proc. ICCV*, pages 900–906, 2003.

[154] D. Terzopoulos. Multi-level reconstruction of visual surfaces: Variational principles and finite element representations. A.I. Memo 671, MIT, 1982.

[155] C. Tomasi and R. Manduchi. Stereo without search. In *Proc. ECCV*, LNCS 1065, pages 452–465, 1996.

[156] P. Torr and A. Criminisi. Dense stereo using pivoted dynamic programming. In *Proc. BMVC*, pages 414–423, 2002.

[157] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – a modern synthesis. In *Vision Algorithms: Theory and Practice*, LNCS 1883, pages 298–372, 1999.

[158] Y. Tsin and T. Kanade. A correlation based model prior for stereo. In *Proc. CVPR*, volume 1, pages 135–142, 2004.

[159] T. Tuytelaars and L. van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *Proc. BMVC*, pages 412–425, 2000.

[160] R. Tyleček. Representation of geometric objects for 3D photography. Master's thesis, Department of Cybernetics, FEE, Czech Technical University, Prague, Czech Republic, 2008.

[161] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, 1995.

[162] A. Vedaldi and S. Soatto. Local features, all grown up. In *Proc. CVPR*, pages 1753–1760, 2006.

[163] O. Veksler. Extracting dense features for visual correspondence with graph cuts. In *Proc. CVPR*, pages 689–694, 2003.

[164] O. Veksler. Fast variable window for stereo correspondence using integral images. In *Proc. CVPR*, pages 556–561, 2003.

[165] O. Veksler. Stereo correspondence by dynamic programming on a tree. In *Proc. CVPR*, volume 2, pages 384–390, 2005.

[166] G. Vogiatzis, C. Hernández, P. H. S. Torr, and R. Cipolla. Multi-view stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE Trans. on PAMI*, 29(12):2241–2246, 2007.

[167] A. Wald. *Sequential analysis*. Dover, New York, 1947.

[168] Y. Wei and L. Quan. Region-based progressive stereo matching. In *Proc. CVPR*, pages 106–113, 2004.

[169] J. Weng. A theory of image matching. In *Proc. ICCV*, pages 200–209, 1990.

[170] T. Werner. A linear programming approach to max-sum problem: A review. Technical Report CTU–CMP–2005–25, Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic, 2005. `http://cmp.felk.cvut.cz/cmp/software/maxsum/`.

[171] T. Werner. A linear programming approach to max-sum problem: A review. *IEEE Trans. on PAMI*, 29(7):1165–1179, 2007.

[172] C. Wheatstone. Contributions to the physiology of vision.part the first. on some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions*, 128:371–394, 1838.

[173] O. J. Woodford, P. H. S. Torr, I. D. Reid, and A. W. Fitzgibbon. Global stereo reconstruction under second order smoothness priors. In *Proc. CVPR*, 2008.

[174] Y. Xiong and S. A. Shafer. Moment and hypergeometric filters for high precision computation of focus, stereo and optical flow. Tech. Rep. CMU-RI-TR-94-28, The Robotics Institute, Carnegie Mellon University, Pittsburgh, 1994.

[175] Y. Xiong and S. A. Shafer. Variable window Gabor filters and their use in focus and correspondence. Technical Report CMU-RI-TR-94-06, The Robotics Institute, Carnegie Mellon University, Pittsburgh, 1994.

[176] Y. Xiong and S. A. Shafer. Hypergeometric filters for optical flow and affine matching. *IJCV*, 24(2):163–177, 1997.

[177] G. Yang, C. V. Stewart, M. Sofka, and C.-L. Tsai. Registration of challenging image pairs: Initialization, estimation, and decision. *IEEE Trans. on PAMI*, 29(11):1973–1989, 2007.

[178] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *Proc. ECCV*, LNCS 801, pages 150–158, 2004.

[179] G. Zeng, S. Paris, L. Quan, and M. Lhuillier. Surface reconstruction by propagating 3D stereo data in multiple 2D images. In *Proc. ECCV*, LNCS 3021, pages 163–174, 2004.

[180] G. Zeng, S. Paris, L. Quan, and F. Sillion. Accurate and scalable surface representation and reconstruction from images. *IEEE Trans. on PAMI*, 29(1):141–158, 2007.

[181] H. Zhang, J. Čech, R. Šára, F. Wu, and Z. Hu. A linear trinocular rectification method for accurate stereoscopic matching. In *Proc. BMVC*, volume 1, pages 281–290, 2003.

[182] Y. Zhang and C. Kambhamettu. Stereo matching with segmentation-based cooperation. In *Proc. ECCV*, LNCS 2351, pages 556–571, 2002.

[183] Z. Zhang and Y. Shan. A progressive scheme for stereo matching. In *Proc. European Workshop on 3D Structure from Multiple Images of Large-Scale Environments*, pages 68–85, 2000.

[184] T. Zickler, P. N. Belhumeur, and D. J. Kriegman. Helmholtz stereopsis: Exploiting reciprocity for surface reconstruction. In *Proc. ECCV*, LNCS 2352, pages 869–884, 2002.

[185] H. Zimmer, A. Bruhn, L. Valgaerts, M. Breuß, J. Weickert, B. Rosenhahn, and H.-P. Seidel. PDE-based anisotropic disparity-driven stereo vision. In *Vision, Modeling, and Visualization*, pages 263–272, 2008.

List of acronyms

ACM	Association for Computing Machinery
BMVC	British Machine Vision Conference
CVGIP	Computer Vision, Graphics, and Image Processing
CVPR	Computer Vision and Pattern Recognition
DAGM	Deutsche Arbeitsgemeinschaft für Mustererkennung
DARPA	Defense Advanced Research Projects Agency
ECCV	European Conference on Computer Vision
FEE	Faculty of Electrical Engineering
IAPR	International Association of Pattern Recognition
ICPR	International Conference on Pattern Recognition
IEE	Institution of Electrical Engineers
IEEE	Institute of Electrical and Electronics Engineers
ICCV	International Conference on Computer Vision
IJCV	International Journal of Computer Vision
IVC	Image and Vision Computing
LNCS	Lecture Notes in Computer Science
MIT	Massachusetts Institute of Technology
PAMI	Pattern Analysis and Machine Intelligence
SCIA	Scandinavian Conference on Image Analysis
SIGGRAPH	Special Interest Group on Graphics
SPIE	Society of Photographic Instrumentation Engineers