Deutsche Geodätische Kommission
der Bayerischen Akademie der Wissenschaften

Reihe C | Dissertationen | Heft Nr. 774

Moritz Menze

Object Scene Flow

München 2016

Verlag der Bayerischen Akademie der Wissenschaften
in Kommission beim Verlag C. H. Beck

ISSN 0065-5325 ISBN 978-3-7696-5186-7

Diese Arbeit ist gleichzeitig veröffentlicht in:
Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Leibniz Universität Hannover
ISSN 0174-1454, Nr. 323, Hannover 2016

DGK Deutsche Geodätische Kommission
der Bayerischen Akademie der Wissenschaften

Reihe C Dissertationen Heft Nr. 774

Object Scene Flow

Von der Fakultät für Bauingenieurwesen und Geodäsie
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des Grades
Doktor-Ingenieur (Dr.-Ing.)
genehmigte Dissertation

von

Dipl.-Ing. Moritz Menze

München 2016

Verlag der Bayerischen Akademie der Wissenschaften
in Kommission bei der C. H. Beck'schen Verlagsbuchhandlung München

ISSN 0065-5325 ISBN 978-3-7696-5186-7

Diese Arbeit ist gleichzeitig veröffentlicht in:
Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Leibniz Universität Hannover
ISSN 0174-1454, Nr. 323, Hannover 2016

Adresse der Deutschen Geodätischen Kommission:

Deutsche Geodätische Kommission

Alfons-Goppel-Straße 11 ● D – 80 539 München

Telefon +49 – 89 – 23 031 1113 ● Telefax +49 – 89 – 23 031 - 1283 / - 1100

e-mail post@dgk.badw.de ● http://www.dgk.badw.de

Prüfungskommission

Vorsitzender: Prof. Dr.-Ing. habil. Monika Sester
Referent: Prof. Dr.-Ing. habil. Christian Heipke
1. Korreferent: Prof. Dr. Konrad Schindler
2. Korreferent: apl. Prof. Dr.-Ing. Claus Brenner

Tag der mündlichen Prüfung: 10.05.2016

ISSN 0065-5325 ISBN 978-3-7696-5186-7

Abstract

Motion estimation is an integral part of image sequence analysis. The estimation of two- and three-dimensional displacement fields can provide valuable information about static and moving components of the environment. Motion observation and segmentation, in turn, are prerequisites for a number of highly relevant applications of computer vision like mobile robotics and autonomous driving. Today, promising performance is demonstrated in recovering shape and motion but both entities are often addressed separately. The goal of this thesis is to show how image sequence analysis can profit from the integration of reconstruction and motion estimation. Furthermore, it is demonstrated that the developed model is flexible enough to easily incorporate sophisticated prior object knowledge. While seminal approaches to both, 2D optical flow and 3D scene flow estimation, rely on continuous optimization techniques, recent trends in related work hint at promising performance of discrete inference. As important limitations of classical approaches can be overcome this line of research forms the starting point for the present work.

This thesis addresses three major questions. First, an image-based approach to two-dimensional optical flow estimation from monocular image sequences is proposed. Following successful work on stereo image matching the problem is formulated in terms of discrete optimization. Second, a novel model for three-dimensional scene flow estimation from stereoscopic image sequences is introduced. The key idea is to decompose the observed scene into a finite number of individually moving objects. Third, specific object knowledge is incorporated into the scene flow model to further regularize the ill-posed problem. All aspects are addressed using versatile random field models, a well-established tool to formalize inverse problems. For all of the three major topics suitable parametrizations and objective functions are introduced. As they involve general prior models, rigorous global optimization becomes intractable. Instead, customized approximate inference strategies are proposed yielding promising performance for each of the tasks.

Thorough experiments are conducted on public benchmark data to investigate strengths and weaknesses of the proposed approaches. On MPI Sintel and the KITTI benchmark, optical flow and scene flow estimation both perform en par with the state-of-the-art confirming the usefulness of the specified models and inference procedures. For the model-based extension of the scene flow algorithm, a prove of concept is provided on challenging sequences from training data of the KITTI 2015 scene flow benchmark. A significant reduction of the computational effort without any loss in performance is achieved by combining the results of the novel optical flow method and the scene flow approach.

Keywords optical flow, scene flow, reconstruction, random fields, active shape model, discrete optimization, image sequences

Kurzfassung

Die Schätzung von Bewegungsfeldern ist ein integraler Bestandteil der Bildsequenzanalyse. Zwei- und drei-dimensionale Bewegungsfelder liefern wertvolle Informationen, sowohl über statische als auch über individuell bewegte Elemente der Umgebung. Die zuverlässige Beobachtung und Segmentierung von Bewegungen ist eine Voraussetzung für eine Vielzahl relevanter Anwendungen der automatisierten Wahrnehmung, beispielsweise im Kontext der Robotik oder autonomer Fahrzeuge.

Für sich betrachtet sind aktuelle Methoden zur räumlichen Rekonstruktion und zur dichten Bewegungsschätzung bereits sehr leistungsfähig. Das Ziel der vorliegenden Arbeit besteht darin zu zeigen, wie die Bildsequenzanalyse von der Integration dieser Ansätze in einem gemeinsamen Verfahren profitieren kann. Darüber hinaus wird die Flexibilität der entwickelten Methodik demonstriert indem eine höhere Stufe der Interpretation in Form objektspezifischen Modellwissens integriert wird. Während die ersten Ansätze zur Berechnung des zweidimensionalen optischen Flusses und des dreidimensionalen Szenenflusses auf Grundlage klassischer kontinuierlicher Optimierungsmethoden entwickelt wurden, belegen aktuelle Forschungsergebnisse die besondere Leistungsfähigkeit der diskreten Optimierung von Zufallsfeldern auf denselben Gebieten. Da sie wichtige Grenzen der klassischen Verfahren überwindet und sich für die angestrebte flexible Modellierung eignet, bildet diese Methodik den Ausgangspunkt der vorliegenden Arbeit.

Drei wesentliche Punkte bilden den Kern dieser Dissertation. Zunächst wird ein Verfahren zur Schätzung des optischen Flusses entwickelt. Die Ergebnisse dienen als Grundlage für die weiteren Beiträge. Ein neues Conditional Random Field Modell zur Bestimmung des dreidimensionalen Szenenflusses ist der zweite Kernpunkt. Eine wesentliche Neuerung besteht in der Annahme, dass die beobachtete Szene in eine kleine Anzahl individuell bewegter Objekte zerlegt werden kann. Zur Ableitung räumlicher Informationen werden stereoskopische Bildsequenzen verarbeitet. Schließlich wird die Methode um ein Objektmodell erweitert, dass zum einen die Berechnung von Bewegung und Geometrie für Objekte regularisiert und zum anderen die Schätzung parametrisierter Rekonstruktionen ermöglicht.

In umfassenden Experimenten werden die Stärken und Schwächen der entwickelten Verfahren untersucht. Die Datengrundlage hierfür bilden öffentlich verfügbare Testdaten, die einen objektiven Vergleich zu dem aktuellen Stand der Technik erlauben. Sowohl die Ergebnisse für den optischen Fluss als auch für den Szenenfluss liegen auf dem Niveau des aktuellen Forschungsstandes und belegen die Leistungsfähigkeit der entwickelten Algorithmen. Vielversprechende Ergebnisse auf anspruchsvollen Testdaten zeigen außerdem den Nutzen der modellbasierten Erweiterung.

Schlagworte Optischer Fluss, Szenenfluss, Rekonstruktion, Conditional Random Field, Active Shape Model, Diskrete Optimierung, Bildsequenzanalyse

Nomenclature

Object Scene Flow

$\mathcal{B}$	the set of shared boundary pixels
$\mathcal{N}$	the set of neighboring image sites
$\mathcal{O}$	the set of objects
$\mathcal{R}$	a region in the image
$\mathcal{S}$	the set of superpixels
$\mathbb{R}^n$	the n-dimensional Euclidean space
$\mathbb{Z}^n$	the n-dimensional set of integers
$SO(n)$	the special matrix Lie group of rotations
$SE(n)$	the special Euclidean group describing rigid body motions
$\mathbf{K}$	the camera calibration matrix
$\mathbf{R}$	a rotation matrix
$\mathbf{X}$	a three-dimensional object point
C	the matching cost
C^{y}	the matching cost with respect to specific features, $\mathrm{y} \in \{\mathrm{dense}, \mathrm{sparse}\}$
D	the combined dissimilarity measure
D^{x}	the dissimilarity measure with respect to a specific image pair, with $\mathrm{x} \in \{\mathrm{stereo}, \mathrm{flow}, \mathrm{cross}\}$
E	an energy function
P	a probability distribution
$\mathbf{l}$	a vector of discrete labels
$\mathbf{n}$	a plane normal
$\mathbf{o}$	a random variable comprising the parameters of an object
$\mathbf{p}, \mathbf{q}$	image locations
$\mathbf{s}$	a random variable comprising the parameters of a superpixel
$\mathbf{t}$	a translation vector
$\mathbf{x}$	an image point

d	a disparity value
i, j, k	indices
l	a discrete label
t	a point in time
u, v	components of the optical flow field in x and y direction
w	a weight
α	a shape parameter for the extended Potts model
θ	a weight
κ	the shape and pose consistency term
$\pi(\mathbf{p})$	the projection of pixel $\mathbf{p}$ to a specific image
Π	the set of sparse observations
ρ	a penalty function
τ	a truncation threshold
φ	the data term
ψ	the smoothness term
ψ^{z}	the smoothness term of specific entities, $\mathrm{z} \in \{\text{depth}, \text{orientation}, \text{motion}\}$
$[\cdot]$	the Iverson bracket
$\|\cdot\|_n$	the l_n-norm

Discrete Optimization for Optical Flow

$\mathcal{K}_{\mathbf{p},\mathbf{q},l^*_{\mathbf{p}}}$	the set of labels for pixel pair $(\mathbf{p}, \mathbf{q})$ and $l^*_{\mathbf{p}}$
$\mathcal{P}$	the set of pixels to be labeled
$l^*_{\mathbf{p}}$	a fixed discrete label at pixel $\mathbf{p}$
λ	the relative weight of data and smoothness term

Joint 3D Estimation of Vehicles and Scene Flow

$\mathbf{M}$	vertex mean of the active shape model
$\mathbf{V}$	the deformed vertex positions of the active shape model
C^{y}	the shape consistency cost with respect to specific object types, with $\mathrm{y} \in \{\text{background (bg)}, \text{object (obj)}\}$
O	the occlusion penalty
S	the shape consistency term
$\mathbf{e}_i$	the i'th eigenvector
$\boldsymbol{\gamma}$	a random variable comprising shape parameters of the active shape model
$\boldsymbol{\xi}$	a random variable comprising pose parameters

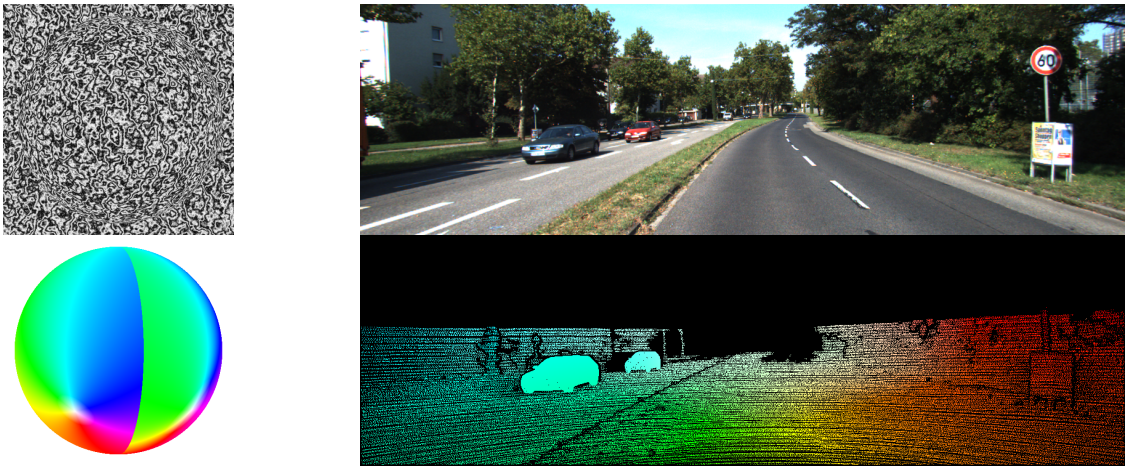
Contents

Chapter 1

Introduction

The perception of scenes in motion is a fascinating challenge for computational vision. Compared to humans, who perceive and predict motions with the greatest of ease, respective vision algorithms are still at an early stage. Decades of research on image-based motion estimation brought about a large number of publications and significantly improved models. Based on this progress the problem of comprehensive scene understanding can be addressed but is still far from being solved. Together with the semantic segmentation and interpretation of image content, robust reconstruction and displacement field estimation are key components of such systems. Realistic, adverse imaging conditions and the dimensionality reduction inherent to the perspective projection of three-dimensional scenes onto the image plane render many related vision tasks ill-posed and, therefore, especially challenging.

This thesis deals with image-based motion estimation in two different domains. On the one hand, it addresses two-dimensional optical flow. The projection of motion onto the image plane is often the only available information about a dynamic scene. While it may preserve a notion of changes in the observed scene, it looses parts of the original information in the process of image formation. On the other hand, this work investigates three-dimensional scene flow, describing a dense three-dimensional displacement field in object space. It is this spatial process, which induces the observed motion, and which is often the actual objective of motion estimation. In addition to the inherent academic interest in perceiving systems, both entities are relevant for a number of applications. While active sensors are a strong competitor in many fields, image sequences contain valuable visual information comparable to those perceived by a human. Automatic navigation of autonomous platforms requires such a detailed perception of the environment. Warning and avoidance of moving obstacles are already part of advanced driver assistance systems but are still restricted to certain types of objects, limited speed and ranges. The safe interaction of robots with their environment also requires up-to-date and precise information about their surroundings. Furthermore, motion cues are important for action and activity recognition for example in video



(a) *Sphere data set* (b) *KITTI 2015 data set*

Figure 1.1: *Two examples from different generations of optical flow benchmarks. Panel (a) depicts the reference frame of a random dot stereogram from [Huguet and Devernay, 2007] in the top row and color-coded optical flow ground truth in the bottom row. Panel (b) shows one image of the KITTI 2015 benchmark [Menze and Geiger, 2015] and ground truth optical flow including annotations of individually moving objects.*

surveillance applications but they are often considered too noisy and unreliable. All of these tasks benefit from an improved perception of surrounding shapes and motions.

Both, optical flow and scene flow estimation are closely related. Therefore, it is not surprising that the development of respective algorithms shares a common trend. Considering the early approaches to both problems, variational formulations dominate the literature (cf. Chapter 3). The increased requirements of more complex applications and more realistic benchmarks quickly unveiled the limitations of these seminal methods.

Publicly available benchmark data sets are an important tool to assess the performance of evolving algorithms and to pose new challenges to the research community. Figure 1.1 sketches the development of optical flow test data. The left panel depicts a random dot stereogram from [Huguet and Devernay, 2007] together with color-coded ground truth displacements. Similar data was used in [Horn and Schunck, 1981] to compute the first optical flow maps. The synthetic input image exhibits strong texture and small, smooth motions amenable to variational methods. However, the reference data reveals sharp motion boundaries, which require appropriate treatment. Along with the development of more robust methods, new and more challenging data sets were published. The right panel of Figure 1.1 depicts an example from the recent KITTI 2015 scene flow data set [Menze and Geiger, 2015]. It was created in the scope of this thesis, providing reference data for outdoor traffic scenes with annotations of individually moving objects. Due to real objects and imaging conditions, the depicted image contains reflections as well as saturated and textureless surfaces. In combination with large relative motions of camera and objects they pose a significant challenge to traditional methods.

Motivated by significant progress of discrete inference techniques, a growing number of approaches formulate the tasks of image matching and motion estimation in terms of discrete optimization problems. As important limitations of classical variational approaches are addressed by discrete inference in probabilistic graphical models, this line of research provides the starting point for the present work.

1.1 Problem Statement

The task of image-based motion estimation belongs to the class of inverse problems. It is also referred to as inverse optics [Poggio et al., 1985], as the goal is to infer surfaces and their motion from images. The principle of image formation comprises the dimensionality reduction from three-dimensional object space to the two-dimensional image plane. Consequently, the reconstruction of surfaces and displacements in general is not unique, rendering the problems ill-posed. To address this kind of problems, i.e. to decide for a plausible solution, it is necessary to impose constraints which are derived from model assumptions or a priori knowledge. Enforcing these restrictions on the solution is referred to as *regularization*. It can be applied in "the form of either variational principles that impose constraints on the possible solutions or as statistical properties of the solution space" [Poggio et al., 1985].

This thesis follows the second path and proposes strategies for the discrete optimization of different, complementary inference tasks on image sequences. It covers different levels of semantic interpretation of the observations and the retrieved results. On the one hand, the estimation of two-dimensional large displacement optical flow is addressed as a representative of early vision problems. Compared to stereo image matching, where discrete optimization has been investigated thoroughly and successfully, optical flow estimation poses a more complex challenge to these techniques. In particular, the cardinality of a label set encoding general two-dimensional displacements is significantly larger than the set of disparities considered in 1D stereo image matching along epipolar lines. Additionally, the labels are not ordered which renders especially efficient optimization schemes, which are well established in the stereo literature, inapplicable to the present problem.

Based on the resulting optical flow matches, a model and inference scheme is developed addressing the related problem of three-dimensional scene flow estimation. In this approach, the observations are extended to stereoscopic image sequences so that reasoning can be conducted in 3D. Although additional requirements are established concerning data acquisition, spatial phenomena are more naturally described in their original three-dimensional domain. Questions raised in this context aim at a useful parametrization and meaningful constraints on the ill-posed problem. Raising the level of interpretation slightly, we will discuss an object-based approach to scene flow estimation.

Finally, we move on to a semantic interpretation of the scene flow results to incorporate high-level object knowledge. Detailed object models allow for a parametrized reconstruction and can support regularization. They provide a powerful means to further reduce the solution space of the inverse problem at hand.

1.2 Contributions

The present work describes a holistic approach to image-based motion estimation. Starting with the observation of motion in the image plane, the underlying three-dimensional displacement field is investigated. The contributions of this thesis, addressing these issues, are:

- A novel discrete optimization approach to optical flow estimation based on a Markov random field model. To initialize the required label set of each node we efficiently sample the high-dimensional solution space inherent to the general two-dimensional image matching problem. A dedicated approximate inference technique based on dynamic programming is developed. It exploits the deliberate structure of the model to deal with large label sets as needed in optical flow estimation.

- A novel conditional random field model for scene flow estimation. The key idea is to decompose the observed scene into a finite number of individually moving objects. This minimal parametrization aims at better constraining the ill-posed problem while preserving the necessary flexibility to cope with complex scenes. The graphical model is complemented with a particle-based inference procedure that allows for the joint optimization of discrete and continuous variables.

- The scene flow model is extended so that it is able to infer pose and shape of individually moving objects. To this end, a parametrized active shape model is incorporated into the basic scene flow model. While the retrieved reconstruction is valuable information in itself, it additionally provides regularization of the sought displacement field.

1.3 Thesis Outline

The contents of this thesis are structured as follows. First, the definitions of optical flow and scene flow are reviewed in Chapter 2. It also provides the basics of continuous and discrete optimization techniques, which form the foundation for major parts of related work and all models proposed in this thesis. Next, the contents of the present work are embedded into the state-of-the-art of optical flow and scene flow estimation in Chapter 3. Research issues are identified in Section 3.4 before Chapter 4 explains the contributions of this thesis in detail. In particular, Section 4.1

elaborates on the details of discrete optimization for optical flow while Section 4.2 provides a thorough description of the scene flow approach. Section 4.3 demonstrates the flexibility of the basic scene flow model and incorporates more sophisticated object knowledge. The results of comprehensive experiments are shown in Chapter 5 to reveal strengths and weaknesses of the methods. These results and their implications are discussed in Chapter 6 before final conclusions are drawn and future directions are sketched in Chapter 7.

Chapter 2

Basics

The present thesis addresses the problems of optical flow and scene flow estimation. This chapter reviews the definition of the sought entities in Section 2.1. Subsequently, Section 2.2 provides the basic concept of continuous optimization in terms of the calculus of variations, which forms the mathematical basis of major parts of the related work and the refinement step of the presented optical flow approach. The proposed methodology is based on discrete inference in versatile random field models. An overview of the mathematical foundations of such models is provided in Section 2.3.1. More comprehensive explanations can be found in the textbooks of [Bishop, 2006], [Barber, 2012] and [Prince, 2012], for example. Section 2.3.2 addresses inference in random field models. The specific formulations developed in this thesis require specialized inference techniques whose basics are introduced in this section. In addition, Section 2.3.3 describes an efficient block coordinate descent approach to inference in random field models that is adapted to optical flow estimation later in the thesis.

2.1 Optical Flow and Scene Flow

Motion is an ubiquitous and important stimulus for most beings equipped with a visual system. It is induced by relative movement between the eye and the visible surroundings. To address its image-based estimation in terms of algorithms, a formal description of the involved phenomena needs to be established. This section distinguishes the perceived optical flow and the causing spatial motion field.

First, we will review the basic concept of optical flow. The textbook of [Horn, 1986] contains a thorough introduction to the topic. To understand the underlying principle, it is necessary to distinguish between a purely geometric approach and the actual process of image formation. Let $\mathbf{X}$ denote a three-dimensional point in object space, which is observed by a camera. The

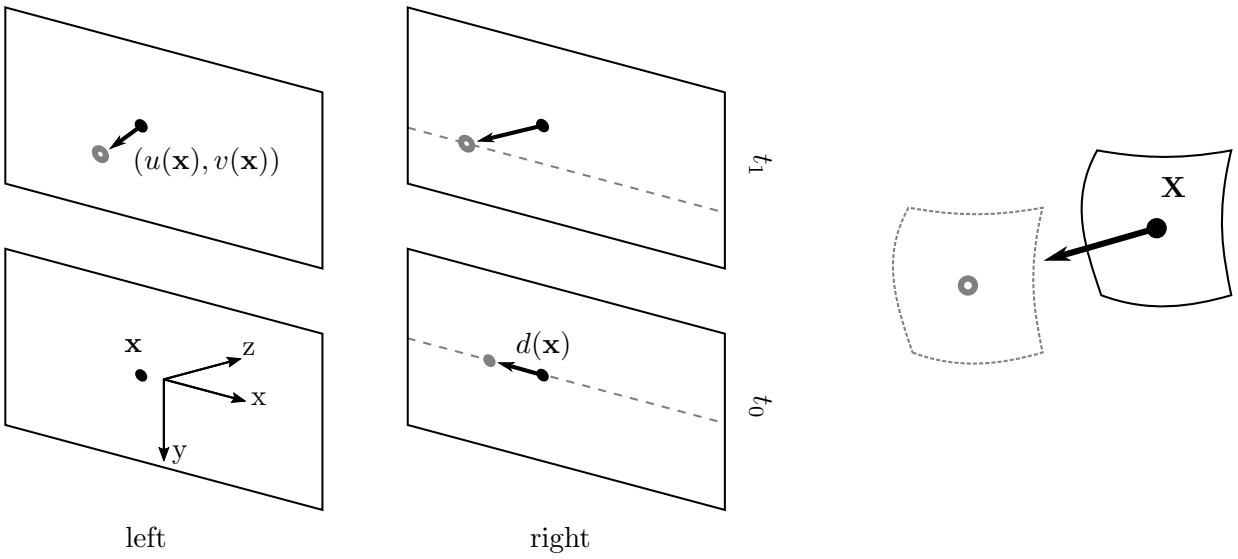


Figure 2.1: *Optical flow and scene flow. Object point* $\mathbf{X}$ *projects to image location* $\mathbf{x}$ *in the reference view (lower left), which is shown as a black dot in the other images for orientation. The projection of* $\mathbf{X}$ *to the right image at* t_0 *is displaced by disparity* d*; the epipolar line in the right images is shown as a dashed gray line. The three-dimensional scene flow vector, sketched as a black arrow in the right panel, induces the displacements in the images at* t_1*. The observed motion between the subsequent images from the left camera is referred to as optical flow* (u, v)*.*

perspective projection model can be applied to compute the two-dimensional displacement vector in the image plane that results from a relative motion between $\mathbf{X}$ and the camera. This motion can be caused by a movement of the camera, of the observed scene, or a combination of both. The two-dimensional vector field comprising the displacement vectors for all surface points visible in the image is referred to as the *motion field*. It is a theoretical entity as it follows from a purely geometric rationale.

In contrast, the term *optical flow* refers to the two-dimensional apparent motion observed by the camera. It results from the more complex process of image formation. Depending on the properties of the observed surfaces and the illumination, the optical flow does not necessarily agree with the motion field. A static object can be illuminated by a moving source of light to induce optical flow where the motion field is obviously zero. Textureless surfaces may move without causing visible displacements in their projection so that the locally estimated optical flow would be zero. Throughout this thesis, we will be interested in recovering the motion field induced by the relative movement of the camera and the observed objects. Regularization will be applied to propagate plausible motion estimates from informative image regions to ambiguous areas. The terms optical flow and two-dimensional displacement field are used in this non-local context.

The starting point for optical flow estimation, and the discussion of related work in Section 3.1, is the well-known brightness constancy assumption

$$\mathbf{I}(x, y) = \mathbf{I}'(x + u, y + v) \tag{2.1}$$

It states that the observed intensity at image location (x, y) in the reference frame equals the intensity at the corresponding position in the second frame, which is displaced by the optical flow (u, v). The images at subsequent times of exposure are denoted by $\mathbf{I}, \mathbf{I}'$. This basic relation will only hold under idealized imaging conditions. As it assumes identical intensities at corresponding pixels, it neglects the influence of noise, changes in illumination and non-Lambertian reflection.

Based on the brightness constancy assumption [Horn and Schunck, 1981] derive one constraint on the optical flow from the linearization of $\mathbf{I}'(x + u, y + v)$ using a Taylor series expansion. After simplification, the optical flow constraint is expressed in terms of all three partial derivatives of the image sequence defined by $\mathbf{I}, \mathbf{I}'$

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

where I_x, I_y, I_t denote the partial derivatives of the intensity function in the image and the temporal domain. The following integral measures the deviation from the brightness constancy assumption over the entire image domain Ω in terms of squared residuals

$$E_{\text{Data}} = \int_{\Omega} (I_x u + I_y v + I_t)^2 \, d\mathbf{x} \tag{2.3}$$

E_{Data} is referred to as a data term, as it evaluates the consistency of the estimated flow field (u, v) with the observations.

Given only one constraint on two sought components of the optical flow, a local estimation will only yield the motion in the direction of the image gradient. This phenomenon is commonly referred to as the *aperture problem*. To allow for the estimation of both components of the displacement field a second constraint is constructed exploiting another assumption. It requires the optical flow field to vary smoothly in most parts of the image. This requirement can be formalized as a smoothness constraint

$$E_{\text{Smoothness}} = \int_{\Omega} \left((u_x^2 + u_y^2) + (v_x^2 + v_y^2) \right) \, d\mathbf{x} \tag{2.4}$$

that is small if the squared sum of gradients (u_x, u_y) and (v_x, v_y) of the estimated flow field are close to zero and grows otherwise.

The weighted sum of the data term (2.3) and the smoothness term (2.4) constitutes an energy functional, which can be minimized to estimate the optical flow. The appropriate mathematical tool to address optimization problems of this kind is the calculus of variations, which will be described in Section 2.2.

Optical flow estimation corresponds to generalized image matching. It is an essential problem in computer vision, as it is relevant to a wide range of applications such as image registration, 3D reconstruction, motion estimation and motion segmentation. As described above, optical flow estimation is subject to diverse assumptions. In Section 4.1, we will develop an approach that is designed to yield accurate results under challenging conditions. However, the resulting two-dimensional displacement field does not allow for spatial analysis of the observed scene by itself. Aiming at three-dimensional scene understanding, we will also address the closely related task of scene flow estimation. Here, the goal is to recover the original three-dimensional motion field that projects to optical flow when observed by a camera.

The corresponding vision task has first been addressed by [Vedula et al., 1999] who provide a definition of the problem. In this case, the description of the sought entity is more intuitive. The term *scene flow* denotes a three-dimensional vector field describing the spatial motion of every point on every visible surface in the observed scene between two subsequent time steps. To define the origin of each three-dimensional displacement vector, scene flow estimation relies on a dense reconstruction of visible surfaces, which can be given or part of the investigations. In image-based approaches, three-dimensional reasoning is enabled by processing synchronized image sequences from stereo or multi-view camera setups.

[Vedula et al., 1999] show that the aperture problem generalizes from optical flow to scene flow. It is not possible to infer the three-dimensional motion field from multiple views using only the brightness constancy assumption. As the constraint equations for different cameras are linearly dependent, some form of regularization is required for scene flow estimation as well.

2.2 Calculus of Variations

Image-based flow field estimation is an inverse problem. Given a set of observed images, the task is to infer an optimal representation of displacements explaining the differences in the observed intensities. Mathematical approaches to solve inverse problems can be formulated in terms of an objective function E evaluating the inconsistency of a sought vector of latent parameters $\mathbf{l}$ and the observed data $\mathcal{D}$. The task of inference is then to minimize the objective function and thus the chosen inconsistency measure.

$$\hat{\mathbf{l}} = \underset{\mathbf{l}}{\operatorname{argmin}} \; E(\mathbf{l}, \mathcal{D}, \boldsymbol{\theta}) \tag{2.5}$$

The model weights $\boldsymbol{\theta}$ serve to adapt a general model to the characteristics of specific data. Their values are either deduced from the problem setup or learned in a training stage. The sought optimal configuration of parameters $\hat{\mathbf{l}}$ is the one that minimizes the objective function.

The original formulation of the optical flow problem lends itself to continuous optimization. While a detailed overview of the corresponding related work will be provided in the next chapter, this section summarizes the underlying common mathematical approach.

In the context of displacement field estimation, it is useful to introduce the concept of *functionals*. A functional is defined as a real-valued function on a vector space. As such, it maps a vector-valued function, which will typically encode a displacement field in the scope of this thesis, to a scalar, typically encoding a consistency measure between observations and displacements. In the continuous formulation of the present problem an abstract representation of the objective functional E is constructed as a weighted sum of the data term (2.3) and the smoothness term (2.4)

$$E = E_{\text{Data}} + \lambda\, E_{\text{Smoothness}} \tag{2.6}$$

where parameter λ controls the influence of the regularization. The continuous energy functional of a 2D displacement field can be specified in terms of an integral over the image domain Ω

$$E[u, v] = \int_{\Omega} F(\mathbf{x}, u(\mathbf{x}), v(\mathbf{x}), u_x(\mathbf{x}), u_y(\mathbf{x}), v_x(\mathbf{x}), v_y(\mathbf{x}))\, d\mathbf{x} \tag{2.7}$$

with u, v encoding the components of the sought vector-valued displacement field. The subscripts indicate partial derivatives. In the seminal approach of [Horn and Schunck, 1981], the derivatives of the flow field are used to formalize the smoothness constraint. The function F represents the specific mathematical model, which is a central subject of research (cf. Chapter 3.1). In particular, it comprises individual data terms and smoothness assumptions. The *calculus of variations* is the mathematical tool to determine stationary points of functionals of the form 2.7, indicating extrema of the integral. It can be applied to find the functions u, v that minimize the energy. Finding optimal u, v corresponds to solving an equation system based on the Euler-Lagrange differential equations constructed from Equation 2.7:

$$\begin{aligned} F_u - \frac{dF_{u_x}}{dx} - \frac{dF_{u_y}}{dy} &= 0 \\ F_v - \frac{dF_{v_x}}{dx} - \frac{dF_{v_y}}{dy} &= 0 \end{aligned} \tag{2.8}$$

They comprise partial derivatives of the functional and the contained functions with respect to the integration variables. The Euler-Lagrange equations provide necessary conditions for stationary

points of the functional. In general, the resulting equation system is non-linear and, consequently, linearization has to be applied. Published approaches differ in the employed strategies for linearization and numerical optimization of the equation system.

As shown in the literature, discussed in Chapter 3, variational optimization methods do provide highly accurate estimates for small displacements. However, linearization of the problem implies that the convergence radius is as small as a single pixel. To be able to estimate larger displacements, hierarchical strategies have to be applied starting from smoothed and possibly sub-sampled versions of the input images. Another option is to initialize continuous optimization with an initial displacement field that contains large displacements with pixel accuracy.

2.3 Random Field Models

As opposed to the continuous formulation, discussed in Section 2.2, the energy minimization tasks addressed in this thesis can also be viewed in terms of discrete optimization. This section reviews the basics of probabilistic graphical models and addresses aspects of inference, which are important for the models proposed in this thesis.

2.3.1 Probabilistic Graphical Models

Many tasks of computational vision are defined on regular structures, like the pixel grid or a neighborhood graph. Whenever the result at a particular image site depends on the result at neighboring image locations, graphical models lend themselves to model the respective probabilistic relationships. The basic representation of such models is a graph $\mathcal{G}$ comprising a set of nodes $\mathcal{V}$, and a set of edges $\mathcal{E}$. Nodes represent observed or hidden random variables; edges specify probabilistic relations between the nodes. The left panel of Figure 2.2 depicts a graphical representation of a Markov random field as one instance of a random field model. The blue nodes represent hidden variables, i.e. the sought parameters $\mathbf{l}$. In this model, the state of each hidden node depends on the state of up to four neighboring hidden nodes. These dependencies are visualized by the blue edges. They form a random field prior, which often implements variants of smoothness assumptions, and accounts for the regularization of the problem. Red nodes represent the image data, which are assumed to be observed throughout this thesis. The red edges indicate the dependency of the parameters on the observations. In Markov random fields, the dependency between parameters, or a label, of a hidden node and observed data is restricted to the associated image site.

Pairwise random fields implement local independence assumptions, the local *Markov property*. It states that each node is independent of any other node given all its neighbors. This property

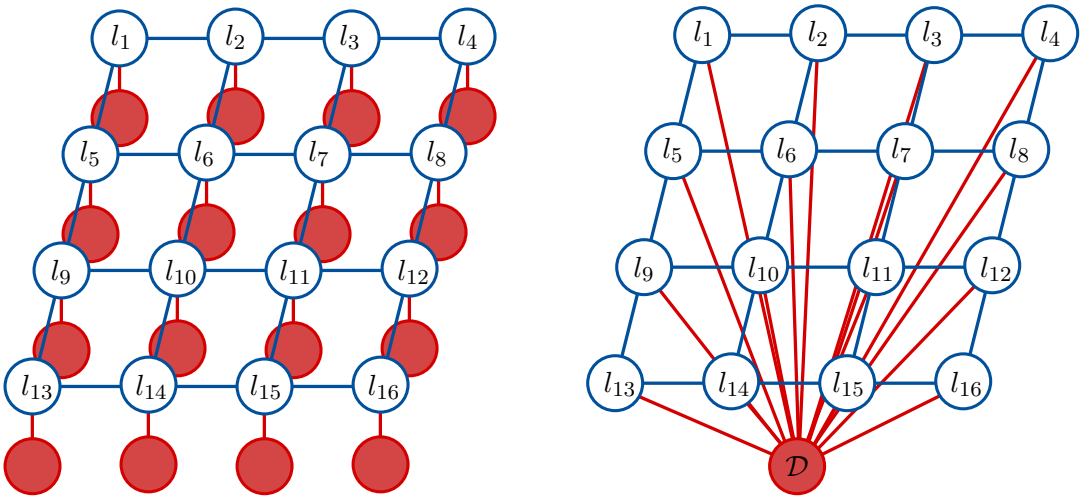


Figure 2.2: *Graphical representation of a Markov random field (left) and a conditional random field (right). Blue nodes represent hidden variables; red nodes visualize observed data. The edges model probabilistic relations between the nodes.*

heavily reduces the complexity of the probabilistic model as the probability for the assignment of a distinct label to a node only depends on the label of nodes it is connected to in a clique. A clique is defined as a fully connected sub-graph of the random field. It can be shown, that Gibbs distributions are the family of joint probability distributions satisfying the Markov property [Li, 2009] and thus describing the probabilistic characteristics of Markov random fields.

Pairwise Markov random fields are a useful representation of vision problems because the joint probability of all nodes $P(\mathbf{l})$ factorizes into a simple product of unary and pairwise terms

$$P(\mathbf{l}) = \frac{1}{Z} \prod_{\mathbf{p} \in \mathcal{V}} \Phi_{\mathbf{p}}(l_{\mathbf{p}}) \prod_{(\mathbf{p},\mathbf{q}) \in \mathcal{N}} \Psi_{\mathbf{p}}(l_{\mathbf{p}}, l_{\mathbf{q}}) \tag{2.9}$$

where Φ denotes the unary potential of node $\mathbf{p}$ given its label $l_{\mathbf{p}}$ and Ψ denotes the pairwise potential of a clique. $\mathcal{V}$ is the set of all nodes in the model and $\mathcal{N}$ is the set of all neighboring nodes, i.e. cliques, with respect to node $\mathbf{p}$. The partition function Z must be applied to ensure that the resulting probability density is valid. To this end, $\int P(\mathbf{l}) d\mathbf{l} = 1$ must hold. For small problems, Z can be computed as the sum over the un-normalized probabilities of all possible combinations of labels. Brute force computation of the normalizing constant Z becomes intractable for problems with realistic size due to the inherent combinatorial complexity ($|\mathbf{l}|^{|\mathcal{V}|}$) which is exponential in the cardinality of the hidden nodes $|\mathcal{V}|$. Thus, in general the probabilities are only known up to a scaling factor. The normalizing constant will be omitted in the following since it does not affect the optimal configuration of labels, indicated by the maximum a-posteriori probability (MAP), which is the primary interest in the scope of this thesis. However, without knowledge of the

partition function the quality of a result cannot be assessed in terms of the absolute value of its posterior probability.

For reasons of numerical stability, the joint probability is usually transformed into the negative log likelihood for optimization.

$$P(\mathbf{l}) = \frac{1}{Z} \exp\left(-E(\mathbf{l})\right) \tag{2.10}$$

With $\varphi = -\log(\Phi)$ and $\psi = -\log(\Psi)$ the task of finding optimal parameters corresponds to finding a minimum of

$$E(\mathbf{l}) = \sum_{\mathbf{p} \in \mathcal{V}} \varphi_{\mathbf{p}}(l_{\mathbf{p}}) \; + \sum_{(\mathbf{p},\mathbf{q}) \in \mathcal{N}} \psi_{\mathbf{p},\mathbf{q}}(l_{\mathbf{p}}, l_{\mathbf{q}}) \tag{2.11}$$

which is a common representation of the objective function E, also referred to as the energy, defined by a conditional random field model.

Here, $\varphi_{\mathbf{p}}$ denotes a data term with respect to all nodes of the random field $\mathcal{V}$. The pairwise term ψ is evaluated for all pairs of nodes in a defined neighborhood $\mathcal{N}$. It facilitates the incorporation of dependencies according to the local Markov property. In Equation 2.10, Z represents the partition function. It is dropped in the following, since it does not depend on the assigned labels and thus it does not affect the computation of the optimal solution [Prince, 2012]. It is important to note that, in contrast to directed graphical models, the potential functions defined for undirected edges do not necessarily need to exhibit a strict probabilistic interpretation. They may be implemented as general functions encouraging plausible configurations of labels.

Conditional Random Fields

[Kumar and Hebert, 2006] introduce discriminative random field models into computational vision. They adapt conditional random fields [Lafferty et al., 2001], which abandon the generative concept of MRF and directly model the posterior probability $P(\mathbf{l}|\mathcal{D})$. This increases the flexibility of the model. The probabilistic relations in the cliques of hidden variables are now conditioned on the given observations. As opposed to Markov random fields, the individual potentials of all cliques may depend on all observed data. A visualization of the resulting graphical model is provided in the right panel of Figure 2.2. Following [Kumar and Hebert, 2006], throughout this thesis all clique potentials with higher order than two are assumed to be zero. As for Markov random fields, the probabilistic model for CRFs is based on Gibbs distributions and the derivation of an energy function corresponds to the description in the previous section.

Today, conditional random fields are a well-established tool in Photogrammetry [Förstner, 2013] and Computer Vision [Szeliski, 2011]. They allow for flexible modeling of complex reasoning and efficient inference to tackle a broad range of vision tasks. Beyond vision research, CRF are employed in computational linguistics [Lafferty et al., 2001] and biological research, for example. The models presented in this thesis will be formulated in terms of energies similar to Equation 2.11.

The typical workflow to tackle inverse problems using probabilistic graphical models consists of three tasks: *modeling* of an appropriate objective function, *learning* of optimal model weights based on training data, and *inference* of optimal parameters. Details of the models and customized inference procedures will be described in Chapter 4 while the training of model weights will be discussed in Chapter 5.

2.3.2 Approximate Inference

The task of inference in random field models is to find an optimal configuration of labels $\hat{\mathbf{l}}$. The MAP solution of a given problem minimizes the energy function 2.11.

$$\hat{\mathbf{l}} = \underset{\mathbf{l}}{\operatorname{argmin}} \left(\sum_{\mathbf{p} \in \mathcal{V}} \varphi(l_\mathbf{p}) + \sum_{(\mathbf{p},\mathbf{q}) \in \mathcal{N}} \psi(l_\mathbf{p}, l_\mathbf{q}) \right) \tag{2.12}$$

For random field models with multiple labels per node and submodular pairwise potentials, which result from convex penalty functions, there exist graph cut methods to find the globally optimal configuration of labels [Prince, 2012]. However, in many vision applications it is important to allow for abrupt jumps of neighboring labels to preserve discontinuities like depth jumps and motion boundaries. Discontinuity preserving pairwise potentials are typically implemented as non-submodular functions, e.g. by introducing a truncation threshold or similar non-convex robust potentials. Consequently, global optimization of energies of the form (2.11) becomes intractable [Boykov et al., 2001].

These combinatorial discrete optimization problems belong to the class of NP-hard problems. In this general case, there is no guarantee to find a globally optimal solution for the original model in polynomial time. To find promising solutions instead, we resort to approximate inference techniques.

Loopy max-product belief propagation (BP) is used by [Weiss and Freeman, 2001] to approximately maximize the joint probability specified in Equation 2.9. The equivalent min-sum algorithm minimizes the energy function (2.11). The idea is to let the nodes of the random field iteratively exchange messages with their neighbors. In case of the min-sum formulation, these messages contain the minimum cumulative cost for each possible label of the target node. Equa-

tion 2.13 shows how a message m^t at iteration t is composed as a sum of data and smoothness term and accumulated information from neighboring nodes.

$$m^t_{\mathbf{p}\to\mathbf{q}}(l_\mathbf{q}) = \min_{l_\mathbf{p}} \left(\psi(l_\mathbf{p}, l_\mathbf{q}) \; + \; \varphi(l_\mathbf{p}) \; + \sum_{(\mathbf{p},\mathbf{r})\in\mathcal{N}\setminus\mathbf{q}} m^{t-1}_{\mathbf{r}\to\mathbf{p}}(l_\mathbf{p}) \right) \tag{2.13}$$

Here, we compute the message from node $\mathbf{p}$ to node $\mathbf{q}$ for one distinct label $l_\mathbf{q}$. $\mathcal{N} \setminus \mathbf{q}$ denotes the neighborhood of the source node without the target node. The message m has to be computed for all possible labels at $\mathbf{q}$. Therefore, all combinations of labels have to be evaluated in the pairwise term ψ and the computational complexity of Equation 2.13 becomes quadratic in the number of labels. After a chosen number of iterations all incoming messages are evaluated to find the belief vector, i.e., the cumulative cost for each label at each node. The algorithm converges for chains and trees and in these cases finds the globally optimal configuration of labels. [Weiss and Freeman, 2001] show that BP can also be applied to loopy graphs with arbitrary potential functions. In this case, BP will in general not find the global optimum but it may yield strong local minima. The loopy variant of BP is not guaranteed to converge.

To improve the performance of approximate inference for challenging general objective functions, [Wainwright et al., 2005] proposed max-product tree-reweighted message passing (TRW). The idea of TRW is to solve a linear programming relaxation of the original problem. To this end, the loopy graphical model is decomposed into a convex sum of spanning trees covering all nodes without any cycles. Then, max-product belief propagation can be applied to each of the trees. Since trees do not contain any loops, it takes two passes of BP to find the optimal configuration of each tree. To find a good approximation of the original energy, the solutions computed on the individual trees have to agree. As long as this is not the case, the results of individual trees are averaged and the re-parametrized trees are solved again. This step is repeated until convergence. In the original algorithm, the global energy is not guaranteed to decrease monotonically. [Kolmogorov, 2006] adapted the optimization strategy to circumvent this problem. Instead of parallel updates of all messages, a sequential update schedule is introduced. Hence, the modified version of the algorithm is referred to as *sequential tree-reweighted message passing* (TRW-S).

The scene flow models presented in Sections 4.2 and 4.3 fall into the class of general NP-hard objective functions. To approximately optimize these functions TRW-S is applied as comparative studies show it performs well on a set of diverse challenging optimization problems in the context of computational vision [Szeliski et al., 2008] [Kappes et al., 2015].

In the presence of continuous random variables, there is the additional need to discretize the solution space in order to apply discrete inference. Motivated by particle filtering approaches for inference in Markov chain models [Ihler and McAllester, 2009] propose a sample-based version of BP for general graphs, such as random fields. Adapted versions of the algorithm have

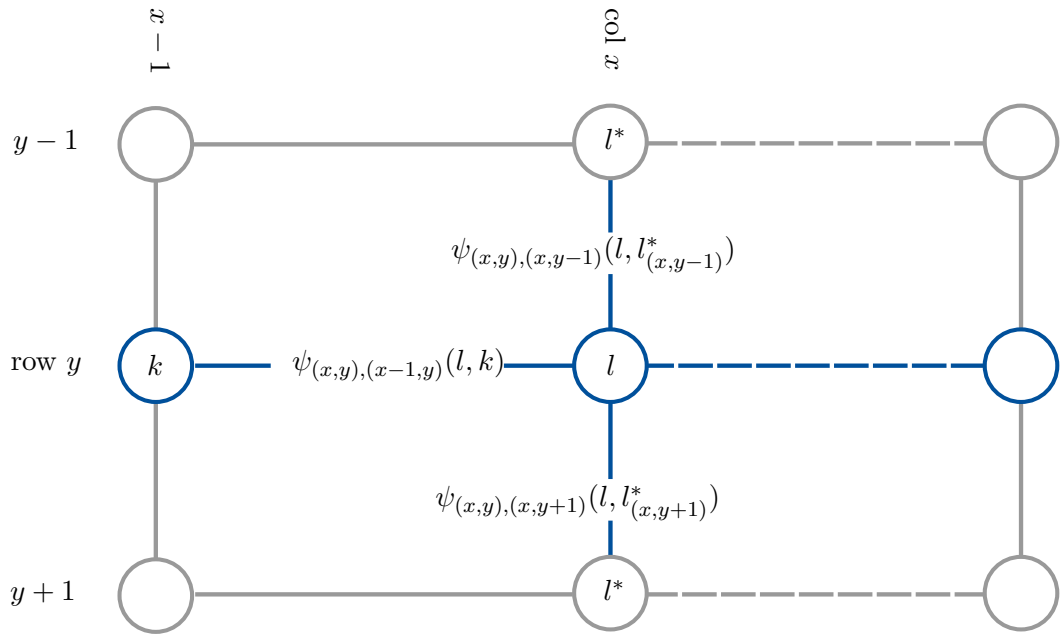


Figure 2.3: *This figure shows the idea and relevant pairwise potentials of Block Coordinate Descent optimization [Chen and Koltun, 2014] for an exemplary row y. In particular, the solid blue edges represent the interactions evaluated to compute the cumulative cost at pixel (x, y).*

successfully been applied to stereo matching [Yamaguchi et al., 2012] and optical flow estimation [Yamaguchi et al., 2013] and will be taken up in Section 4.2.5.

2.3.3 Block Coordinate Descent

In loopy graphical models, even approximate inference can be computationally demanding. To address this issue, [Chen and Koltun, 2014] propose an algorithm that aims at efficient optimization of energies of the form (2.11). Accepting a decrease in accuracy, the authors present an algorithm that is able to perform efficient approximate inference in random field models with large, ordered label sets. The iterative approach optimizes individual image rows and columns conditioned on the remaining, fixed variables.

Each of the rows and columns of a regular random field forms a chain graph. Dynamic programming can be applied to find the optimal configuration of discrete labels in such a chain. The algorithm traverses the chain and sequentially computes minimal cumulative costs for each label at each node. Following this forward pass, backtracking is applied to find the optimal combination of labels for the respective chain. The approach of [Chen and Koltun, 2014] decomposes the random field into four blocks of non-adjacent rows and columns, which are processed sequentially. The results are fixed while in turn the block of adjacent image rows is processed. Unlike loopy belief propagation, the algorithm is guaranteed to converge as each iteration will reduce the energy or leave it unchanged. The approach is computationally efficient as independent rows and

columns can be processed in parallel. This algorithm forms the basis of the inference procedure in the optical flow method proposed in Section 4.1. As it will be adapted to the specific task, the original formulation is explained in the following.

Without loss of generality, we consider the optimization of an image row y in Figure 2.3. The naïve dynamic programming algorithm recursively fills the cumulative cost matrix $\mathbf{C}$ for each pixel $\mathbf{x}$ with column index x from 1 to image width W using the update Equation 2.14. The cumulative cost for a label l at node $\mathbf{x}$ consists of three parts. The data term $\varphi_{\mathbf{x}}$ evaluates the consistency of the label and the observations at the associated image site. In the algorithm of [Chen and Koltun, 2014] it is complemented with a pairwise term ψ that penalizes inconsistencies with respect to neighboring rows. Here, $l^*_{\mathbf{p}}$ denotes the label assignment of the fixed variables, i.e., the variables *outside* row y. The most expensive part in terms of computational complexity is the evaluation of all combinations of labels to compute the pairwise smoothness potential along row y. This problem is stated in the last row of Equation 2.14. As opposed to the fixed labels of adjacent rows, the preceding node can take any of $|L|$ proposed labels.

$$\begin{aligned}\mathbf{C}(x,y,l) = \varphi_{(x,y)}(l) \\ &+ \psi_{(x,y),(x,y-1)}(l, l^*_{x,y-1}) \\ &+ \psi_{(x,y),(x,y+1)}(l, l^*_{x,y+1}) \\ &+ \min_{0 \le k < L} \left(\psi_{(x,y),(x-1,y)}(l,k) + \mathbf{C}(x-1,y,k)\right)\end{aligned} \tag{2.14}$$

Under certain conditions, it is possible to lower the complexity of the overall algorithm from $O(WL^2)$ to $O(WL)$ by computing the last term more efficiently. This complexity reduction applies, if the pairwise potential directly depends on the difference between actual labels, instead of encoded states. Furthermore, the labels have to be embedded in a grid and the pairwise penalty has to be a function of the distance in this grid. Both conditions are frequently met in early vision problems like one-dimensional stereo matching along epipolar lines, where the labels directly represent disparities, or image restoration, where they are interpretable as intensity values. [Felzenszwalb and Huttenlocher, 2006] originally showed how to speed up the computation of BP messages (2.13) by computing the lower envelope of the smoothness term ψ. In case of ordered label sets, the minimum pairwise penalty can be pre-computed in linear time for all labels k at the preceding node. The resulting look-up table provides minimal costs for any label l at the current node in constant time.

In principle, this concept generalizes to label sets of higher dimension, as encountered in two-dimensional optical flow estimation. However, this implies a dense discretization of the flow space to meet the conditions stated above.

Chapter 3

Related Work

To put the contributions of this thesis into perspective, this section provides an overview of the developments and the state-of-the-art in optical flow estimation, in Section 3.1, and scene flow estimation, in Section 3.2. Related work on model-based reconstruction is provided in Section 3.3. The final Section 3.4 discusses previous findings and summarizes open questions.

3.1 Optical Flow Estimation

Due to the challenging nature of the problem and its relevance to a wide range of applications, there exists an immense body of literature on optical flow estimation. This section is structured in three parts and touches upon a number of important publications to sketch the developments, which form the basis of this work. In particular, it starts with early variational approaches, then discusses their feature-based extensions and concludes with recent approaches in terms of discrete inference.

Variational Approaches

Seminal work on optical flow estimation dates back about 35 years to [Horn and Schunck, 1981]. The authors propose the first approach to compute optical flow as explained in Chapter 2. Observing that the two-dimensional displacement of a picture element is under-constrained by a single change of intensity, additional constraints are introduced. In their early work, Horn and Schunck restrict the problem domain to flat surfaces, uniform illumination and smoothly varying reflectance. Thus, brightness constancy and smoothness constraints can be strictly imposed on the problem, leading to a tractable formulation in terms of the calculus of variations, cf. Section 2.2. This first approach computes a displacement field, which jointly minimizes the intensity

differences at associated image locations and the gradient between neighboring flow estimates. To allow for motion discontinuities, which are inherent to more realistic input data but not considered in the pioneering formulation, subsequent work adds robustness against violations of the basic assumptions. [Black and Anandan, 1993] introduce robust statistics, familiar from the literature on geodetic adjustment theory [Koch, 1999], to account for motion discontinuities in the smoothness term and gross errors in the data term. The key idea of robust estimators is to reduce the influence of observations with large residuals by re-weighting them accordingly. In the meantime, the data term had been adapted to compare small image patches instead of individual pixels as proposed in the seminal work. Larger patches can conceptually overcome the original aperture problem and reduce the influence of image noise, but they imply new, restrictive assumptions like constant velocity inside the patch.

Although the more robust methods cope with the issue of over-smoothing at motion boundaries, they are still limited to very small displacements. This is due to the necessary linearization of the objective function during optimization. To a certain extent, the influence of this limitation can be reduced by increasing the frame rate of the input image sequences to limit the magnitude of displacements. On the other hand, restrictions with respect to memory and computational demands suggest the development of methods to infer large displacements directly. To this end, continuous methods typically rely on pyramids of spatially smoothed and sub-sampled images [Anandan, 1989]. Starting with the assumption of zero flow at the coarsest resolution, small optical flow vectors are estimated. On the next level of the pyramid, one image is warped according to the previous result to reduce the magnitude of the displacements to be estimated. [Brox et al., 2004] provide theoretical foundations for the warping and formulate a nested iteration scheme to compute an optical flow field allowing for large displacements. The authors also adapt the data term again and propose to complement the basic brightness constancy assumption with a gradient constancy assumption to cope with slight changes of illumination between the images. Many of the most successful works still follow the same variational paradigm to date. A comprehensive survey of insights and best practices is provided in [Sun et al., 2013]. The article was published together with an implementation of variational reference methods. Very successful results are reported for an extension of classical methods by a non-local smoothness term. It follows the idea of median filtering and considers the context in an image region. [Demetz et al., 2014] propose an extension of [Brox et al., 2004], which additionally solves for a parametrized representation of illumination changes. [Ranftl et al., 2014] re-formulate the smoothness constraint to further gain robustness against motion discontinuities. As a data term they propose a scale-robust variant of census matching, considering differently sized patches in the second frame.

Another starting point to improve the performance of 2D displacement field estimation is the parametrization of the underlying motion model. [Nir et al., 2008] follow this direction and propose an over-parametrization of the optical flow for improved regularization during variational optimization. At each image location, the parameters of a motion model are estimated instead

of the displacement vectors. Thus, deviations from the motion model can be penalized instead of differences between neighboring flow vectors. The idea circumvents unjustified penalties in cases where neighboring optical flow vectors differ in accordance with a consistent local motion model.

Unfortunately, the coarse-to-fine strategy, necessary to cope with large displacements, is prone to errors at small, individually moving image segments. As texture details and fine image structures vanish at small scales of the image pyramid, hierarchical approaches may fail to provide an accurate initialization for the next level of the pyramid. Under such circumstances small regions of isolated motion are absorbed by their surroundings. Consequently, their motion is likely to be lost.

Feature Correspondences

Correspondence estimation can also be based on rich feature descriptors which are able to efficiently match sparse interest points. [Stein, 2004] proposes a method based on a variant of census features [Zabih and Woodfill, 1994], [Geiger et al., 2011] describe a real-time algorithm for sparse feature matching with an extension for depth map estimation. These approaches provide high frame rates, which is desirable in the context of image sequence analysis. On the other hand, they are not able to compute dense motion fields as required in optical flow estimation. In the literature, sparse feature matches are employed to guide dense flow field computation.

[Brox and Malik, 2011] started this line of work by incorporating SIFT feature correspondences as a constraint into the variational processing pipeline. During optimization, the developing flow field is biased towards the discrete feature matches. [Braux-Zin et al., 2013] generalize the idea to sparse, sub-pixel accurate feature and segment matches. Several publications investigate more powerful techniques for sparse matching. [Weinzaepfel et al., 2013] introduce *DeepMatching*, inspired by deep convolutional nets. The algorithm is based on a deformable SIFT descriptor which is decomposed into four patches that can move individually in the target image. Thus, the descriptor becomes more suitable for non-rigid matching. The matching procedure mimics a bottom up multi-stage architecture with interleaving convolutions and max-pooling as used in convolutional nets. Accordingly, it is designed in a bottom-up fashion that first computes a similarity measure between each patch from the reference image and all possible positions in the target image. These correlation maps are aggregated to derive correlations for virtual larger patch sizes, i.e. scaled down versions of the original input image. The resulting pyramid of correlation maps is traversed from the coarsest scale to the finest level to compute the actual matches. The complete processing pipeline additionally comprises variational optical flow estimation based on the established correspondences; it is referred to as DeepFlow. [Timofte and Van Gool, 2015] replace the sparse matching of DeepFlow, i.e. DeepMatching, by a highly accurate matcher applied to sparse interest points. Their method is thus called SparseFlow. A refinement step for pre-computed sparse optical flow matches is subject of [Drayer and Brox, 2015]. To increase the accuracy and reliability of the initial set of correspondences the authors develop a refinement

algorithm that can be combined with many published methods employing feature matches. The initial matching is regularized with respect to a finite number of affine motion models in a MRF framework.

[Revaud et al., 2015] recently proposed *EpicFlow*, the name being an acronym for Edge-Preserving Interpolation of Correspondences. The proposed two-step procedure first interpolates a set of sparse input correspondences to initialize a dense motion field and then refines the results using a classical variational approach. In the first stage, optical flow vectors, as produced by DeepMatching, are interpolated between neighboring input matches. The distance metric employed to identify and weight nearest neighbor matches respects significant image edges from structured edge detection [Dollár and Zitnick, 2013]. Image edges are considered a superset of motion boundaries. Consequently, the described approach helps to align both entities. This initialization replaces the coarse-to-fine strategy for large displacement optical flow, described above, and circumvents the problem of losing information about fine image details on coarse resolution levels. In a second stage, the interpolated flow field is used as the initialization of the warping-based variational optimization proposed by [Brox et al., 2004].

The optical flow approach proposed in Section 4.1 shares similarity with these methods in the sense that it also refines an initial integer-valued flow field to sub-pixel accuracy by applying variational post-processing. However, in contrast to the above-mentioned works, the initial matches are obtained via discrete optimization with optical flow priors. This allows our algorithm to establish denser correspondences than possible with independently matched sparse feature points and, in combination with sub-pixel refinement, leads to better results.

For reasons of computational efficiency traditional approaches to feature matching, such as SIFT [Lowe, 2004], typically restrict themselves to discrete feature points. A generalization of this concept is provided by approximate nearest neighbor fields (ANNF), which yield dense correspondences. An important contribution to this field is Generalized PatchMatch [Barnes et al., 2010], which yields a number of candidate matches in a target image for each small patch in the reference frame. To this end, Generalized PatchMatch randomly initializes a correspondence field and iteratively improves the result by random search and propagation of promising displacements based on the best solution so far. In contrast to optical flow approaches there is no regularization imposed. [Revaud et al., 2015] compare a variant of PatchMatch to DeepMatches as input to EpicFlow and report consistently better results of the dedicated optical flow matcher.

[Bailer et al., 2015] follows the idea of ANNF but explicitly addresses optical flow estimation. The resulting *Flow Fields*, again, serve as input to EpicFlow post-processing. To meet the requirements of successful optical flow initialization, special care is taken to avoid grossly wrong correspondences. To this end, the authors propose a hierarchical approach performing propagation and random search based on an initialization on a coarse pyramid level.

Discrete Optimization

Alternatively, optical flow estimation can be formulated as a discrete optimization problem. Seminal work by [Black and Anandan, 1991] aims at retrieving inhomogeneous motions and explicitly allows for motion discontinuities. The objective function is composed of robust data and smoothness terms, accounting for violations of the respective underlying assumptions. Incremental minimization of the objective function is implemented as a variant of simulated annealing adaptively discretizing the solution space. This approach allows for subpixel accurate displacement estimation but still relies on a coarse-to-fine strategy to cope with large displacements.

For static scenes [Yamaguchi et al., 2013] show how to exploit the epipolar geometry between subsequent frames to constrain optical flow estimation to a 1D matching problem along the epipolar lines. To this end, the fundamental matrix between both input images is estimated and used to compensate for rotations between the views. A variant of semi global matching can then be applied to infer a semi-dense displacement field. To smooth and interpolate the results, the same paper proposes an efficient segmentation algorithm, which takes into account motion or depth discontinuities in addition to image edges. Since the approach can be seen as an extension of the, purely image-based, SLIC segmentation algorithm [Achanta et al., 2012] it is referred to as *MotionSLIC*. In a final step, a mixed discrete-continuous MRF is solved to jointly fit planar motion models to the image segments and infer occlusion boundaries. The solution space of the continuous variables is iteratively discretized as described in Section 2.3.2. Discrete optimization yields a number of compact superpixels containing dense, parametrized flow estimates.

Incorporating smoothness constraints is much more involved for discrete optical flow than for the related stereo matching problem due to the extremely large label space of the discretized 2D flow field. To avoid this difficulty, a number of approaches formulate optical flow as a segmentation problem. Based on a small set of dense flow field proposals the most likely flow field at each pixel is approximated, subject to regularization constraints. [Lempitsky et al., 2008] describe early work in this direction. By running traditional optical flow methods with different parameter settings, they generate a few hundred global displacement fields, which are combined to an intermediate result using fusion moves. After assigning each picture element to one of the proposals the optical flow is refined using continuous optimization. [Wulff and Black, 2015] share the idea that complex displacement fields can be approximated as a weighted sum of a small number of basis flow fields. To compute the basis, principal component analysis is applied to optical flow fields from hours of motion pictures. To infer the displacement field of unseen test data the weights of the most significant eigenvectors are adjusted to sparse feature matches using a robust regularized least squares approach. As these weights completely define a dense motion field, the algorithm is very fast. Because the described method conceptually blurs motion boundaries, they are refined by optimizing a MRF, which associates pixels to one of the proposal flow fields.

[Yang and Li, 2015] follow prior work and estimate piecewise homography models and a segmentation of the input image into regions of consistent motion. In this work, the data term comprises robust versions of brightness and gradient constancy as described above. A smoothness constraint is imposed on boundary pixels of adjacent segments. It encourages adjacent motion models to coincide at their boundaries. There are two groups of unknowns in this model. Discrete labels assign each pixel to one of the segments. The parameters of a global set of homographies are initialized and refined during optimization. Inference is conducted alternatingly for discrete and continuous variables. To account for local deviations from the motion model a variational refinement step is applied to the results of the original approach.

In contrast to pre-computed flow fields, a common reduced label set for all pixels is proposed by [Mozerov, 2013]. Based on a correlation measure, promising flow vectors are identified and proposed as discrete labels restricting the method to scenes with little and non-complex motion. The author reports no more than 300 distinct optical flow labels for each of the processed image pairs. The optimization of the flow field is conducted in a two-step procedure. First, an integer valued flow field is retrieved by optimizing a random field model using TRW-S. Then, the labels are exchanged with subpixel refinement proposals in several directions and the random field is iteratively optimized. Finally, bilateral filtering is applied to refine the results further. In contrast, the approach presented in this thesis pursues a more direct approach to discrete optical flow, which does not rely on proposed flow fields or a small global label set but allows for an arbitrary set of flow proposals *per pixel*.

Very recently, object recognition and feature learning have been exploited as a powerful source of information for optical flow. Notable examples are the data-driven flow transfer method of [Wei et al., 2014] and FlowNets based on deep neural networks by [Fischer et al., 2015]. In this thesis, the focus lies on a more generic model without the need for large annotated training sets. However, these ideas could be easily incorporated into the proposed model, for example, via additional unary terms, promising further gains in performance in the future.

3.2 Scene Flow Estimation

In optical flow estimation, we are interested in recovering the apparent motion in a monocular image sequence. For practical applications it is often more helpful to recover the shape and motion of the actual three-dimensional objects which induce the observed displacements. As described in Section 2.1 the corresponding task is referred to as scene flow estimation. Despite the obvious usefulness of a three-dimensional displacement field for various applications, the number of

published approaches to this problem is significantly smaller compared to publications on optical flow estimation.[1]

As in the literature on optical flow, image-based methods for scene flow estimation can be categorized into variational and discrete optimization approaches. With the advent of consumer grade active sensors like the Microsoft Kinect, depth information has also become available and is employed in a number of publications, e.g. [Herbst et al., 2013, Hornacek et al., 2014, Quiroga et al., 2014]. While active sensors typically work well for indoor scenes with limited extent, this work deals with outdoor scene flow estimation with an application to autonomous driving and therefore this section focuses on appearance-based methods.

Variational Approaches

Following the seminal approaches to optical flow [Horn and Schunck, 1981] and scene flow estimation [Vedula et al., 1999, Vedula et al., 2005], the problem of estimating a three-dimensional displacement field is traditionally formulated in a variational setting. As depth information is needed, related work proposes different ways to incorporate dense reconstruction in the variational framework. Analogous to the 2D case, optimization has to proceed in a coarse-to-fine manner to avoid local minima of the energy functional and capture large displacements. On each level of the image pyramid, an equation system is solved for small displacements. Published methods differ in the optimization strategy.

[Pons et al., 2007] alternatingly optimize the reconstruction of a surface model and the motion field. The key contribution addresses the data term. To circumvent common assumptions of similarity measures the authors propose a global prediction error evaluating the consistency of all input images, which are warped according to the reconstructed surfaces and estimated motion. The resulting algorithm appropriately handles projective distortion and partial occlusions. To regularize the results simple smoothness constraints are imposed to shape and motion, respectively. The resulting energy functional is approximately optimized in a coarse-to-fine gradient descent framework. [Huguet and Devernay, 2007] generalize the variational optical flow method of [Brox et al., 2004] to jointly infer geometry and motion. To this end, they propose a minimal representation of scene flow by four variables in the image domain. In particular, they compute the disparity at the first time step t_0, the optical flow with respect to the reference image and the disparity at the second time step t_1. Given a calibrated stereo camera, the three-dimensional scene flow can directly be computed from this representation. In the present thesis, the same parametrization is used. To jointly optimize stereo disparity and optical flow, [Huguet and Devernay, 2007] extend the respective data and smoothness terms to cover all sought entities and combine them in a unified energy functional. This formulation leads to four partial differential equations, which are

[1] In January 2016 there existed 10,837 entries tagged with 'optical flow' in the IEEE database, and 1,234 tagged with 'scene flow'.

optimized using the numerical scheme proposed by [Brox et al., 2004] for 2D optical flow. Since stereo image matching typically has to deal with large displacements, a dedicated initialization procedure is required. To this end, pre-computed disparity and optical flow maps are employed. We will pick up this strategy for the initialization of the proposed scene flow approach as described in Section 4.2.4.

An important aspect of scene flow estimation is regularization. [Basha et al., 2013] argue that smooth 3D motion fields can project to discontinuous 2D flow fields and thus propose a 3D model representing the scene as a point cloud with spatial motion vectors. This formulation enables them to apply regularization directly to the three-dimensional motion vectors and to extend their method to a multi-view set-up. [Vogel et al., 2011] replace the global total variation regularization by a piecewise rigid prior. Thus, existing sharp discontinuities in the scene flow field can be preserved more faithfully.

[Valgaerts et al., 2010] discard the common assumption of a fully calibrated stereo rig and explicitly estimate the relative orientation between the stereo heads. Consequently, the results are only retrieved up to an unknown scale factor. The energy functional becomes more complex as it comprises general stereo terms based on the unknown fundamental matrix. It is minimized in an involved coarse-to-fine optimization scheme. Furthermore, the authors decouple regularization of shape and motion, as they do not assume respective discontinuities to coincide.

For reasons of computational efficiency, [Wedel et al., 2008] completely decouple shape and motion estimation, which significantly increases the manageable frame rates but also discards valuable mutual constraints between both entities. For practical applications, subject to real-time constraints, the authors propose to consider a disparity map as given and tailor the variational optimization of [Brox et al., 2004] towards estimating motion in the image domain and the change of disparity. [Rabe et al., 2010] follow this approach and parallelize computations on a GPU. They apply Kalman filtering to each pixel individually to smooth the resulting motion vectors over longer image sequences. A comprehensive discussion of variational approaches to scene flow estimation in the context of automotive applications is provided in the textbook by [Wedel and Cremers, 2011].

Discrete Optimization

In the domains of optical flow and stereo image matching, several publications demonstrate the usefulness of slanted-plane models [Bleyer et al., 2011, Yamaguchi et al., 2013]. The visible surface of the scene or its projected motion is assumed to vary smoothly within small regions in the reference image. Extending this idea to 3D forms the basis for scene flow models in terms of discrete random fields.

An early formulation of joint stereo and motion estimation in terms of a Markov random field is described by [Isard and MacCormick, 2006]. The random field is defined on the pixel lattice with discrete random variables comprising densely discretized disparity, optical flow, change of disparity and the occlusion state for each image location. For realistically sized problems, the resulting state space becomes intractably large so the paper describes experiments on a heavily reduced image sequence of 50×40 pixels with 1536 possible labels at each node. Loopy belief propagation is applied to minimize the objective function approximately. To speed up computations the lower envelope trick [Felzenszwalb and Huttenlocher, 2006] described in Section 2.3.3 is applied.

[Yamaguchi et al., 2014] proposes a semi-dense method, which builds on the well-known semi global matching (SGM) described in [Hirschmüller, 2008]. The approach is extended to incorporate a third image from the reference camera, taken at a second time step. Based on the assumption of a static scene, this additional information increases the robustness of image matching. Together with a disparity map in the reference image, it yields an estimate of the optical flow. To smooth and extrapolate the matching results, a slanted-plane model is optimized yielding an over-segmentation of the reference view together with dense estimates of disparity and optical flow. The combined approach is referred to as *slanted plane smoothing of stereo and flow* (SPS-StFl). As in previous work of the authors [Yamaguchi et al., 2013] there is a purely stereoscopic variant of the approach (SPS-St) and one version is tailored towards optical flow estimation (SPS-Fl).

[Vogel et al., 2013b] propose a scene flow approach assuming piece-wise rigid surfaces (PRSF). It is formulated as a slanted-plane model, which decomposes the 3D scene into planar regions, each undergoing a rigid motion. The reference image is decomposed into segments accordingly and for each of the segments a parametrized representation of shape and motion is sought. Consequently, the number of unknowns is reduced compared to a pixel-wise representation. The smoothness assumption within each segment further implements a strong regularization. Inference in this model assigns each pixel to an image segment and each segment to one of several rigidly moving plane proposals in three-dimensional space, thus casting the task as a discrete labeling problem.

To initialize the plane proposals, 3D plane parameters and rigid body transformations are robustly fit to initial disparity and flow maps. These observations are evaluated with respect to an initial segmentation of the reference frame. During inference, the resulting planes are proposed for superpixels in the vicinity of the reference segment. An estimate of the ego-motion is introduced as another proposal.

The objective function optimized during inference is specified as a discrete conditional random field. Based on the plane proposals, the energy is approximately minimized using α-expansion and quadratic pseudo-boolean optimization (QPBO) [Rother et al., 2007]. First, the association of image segments to plane proposals is found and next the assignment of pixels to image segments is refined based on the initial result. Impressive performance has been demonstrated on challenging street scenes as well as on the KITTI stereo and optical flow benchmarks [Geiger et al., 2012].

In consecutive publications [Vogel et al., 2014, Vogel et al., 2015], it is also shown how to extend scene flow estimation to longer image sequences beyond the classical set-up of two subsequent stereo pairs.

The proposed scene flow method described in Section 4.2 is related to this line of work. It is based on a comparable minimal parametrization but it additionally takes advantage of the fact that many realistic scenes can be decomposed into a small number of rigidly moving objects and the background. We jointly estimate *this decomposition* as well as the motion of the objects and the plane parameters of each superpixel in the image. In contrast to [Vogel et al., 2013b, Vogel et al., 2014] where all shape and motion proposals are fixed a-priori, we optimize the continuous variables in our model jointly with the object assignments. Besides obtaining a segmentation of the objects according to their motion, the scene flow in our model is uniquely determined by only four parameters per superpixel (3 for its geometry and 1 for the object index) together with a small number of parameters for each moving object. The experimental investigations in Section 5.3 demonstrate that this minimal representation yields faithful reconstructions and is able to overcome motion ambiguities.

3.3 Model-based Reconstruction

The ill-posed problem of scene flow estimation requires strong regularization. Adequate models impose reasonably general assumptions on the observed displacement field. Considering distinct applications, it seems promising to investigate additional sources of information. Three-dimensional object models have a long history in supporting geometric reconstruction from images in a broad range of applications. On the one hand, they can provide more compact and meaningful results. On the other, they are able to express specific constraints. Compact surveys on the topic are provided in the textbooks [McGlone, 2004] and [Szeliski, 2011], for example.

Pioneering work, for example by [Braun et al., 1995] and [Debevec et al., 1996], made use of shape primitives to support photogrammetric modeling of buildings. While modeling generic objects, like buildings, is a very challenging task by itself, there are tractable approaches to formalize the geometry of objects with moderate intra-class variability. Faces and cars are relevant examples of well-defined geometry, which are frequently addressed in the literature. A widely used representation of such geometric objects is the active shape model (ASM) proposed by [Cootes et al., 1995]. This model is based on manually annotated, corresponding landmark points. Mean positions of these landmarks are computed from a set of annotated training examples. Principal component analysis of the training data yields characteristic deformations between similar shapes. Deformed versions of the model are computed as linear combinations of the mean shape and a weighted sum of the deformations. Thus, the model is flexible but it can only deform in accordance with the variability contained in the training data.

One exemplary line of work that points out the importance of a feedback-loop between early vision and high-level interpretation was published in [Leibe et al., 2006, Thomas et al., 2007]. Based on an implicit shape model the approaches are able to transfer meta information from training images to unseen object instances. This high-level object knowledge can be employed as prior information for early vision tasks. [Leibe et al., 2006] aims at the analysis of traffic scenes. Objects, ego motion and a simple facade reconstruction are jointly inferred from stereo image sequences. Here, object detection helps to mask image regions, which do not contain information on the sought entities. [Thomas et al., 2007] describe a more general approach working on single input images. In this case, depth maps and annotations of object parts are transferred from training images to unseen test images. Again, the resulting prior information can be employed to guide early vision tasks like depth reconstruction.

Recently, the integration of object models into reconstruction algorithms has attracted renewed attention. [Bao et al., 2013] support a multiview stereo approach with an object model. They compute a mean shape from laser scans of different instances of an object class along with a set of discrete anchor points. An object detector is applied to the input images to instantiate the model. Using HOG features, the mean shape is adapted to a newly observed instance of the object by registering the anchor points. [Dame et al., 2013] also use an object detector to infer the initial pose and shape parameters for an object model, which they then optimize in a variational SLAM framework. [Güney and Geiger, 2015] introduce CAD shapes to support binocular stereo matching. They make use of a semantic segmentation of the reference frame to initialize and constrain object hypotheses. So-called *displets*, object-specific disparity patches, are randomly sampled from a large set of CAD models and integrated into the estimated disparity map to fill in uncertain regions. As opposed to these methods, we will not apply an object detector but use a simple, motion-based segmentation of the scene to initialize object hypotheses.

[Zia et al., 2013a, Zia et al., 2013b, Zia et al., 2015] revisit the idea of the ASM and apply it to a set of manually annotated CAD models to derive detailed 3D geometric object class representations. While they tackle the problem of object recognition and pose estimation from single images, in this thesis, such models are used in the context of 3D scene flow estimation.

3.4 Discussion

Before the next chapter provides a detailed description of this thesis' contributions, this section briefly summarizes open questions in the related work on the estimation of two- and three-dimensional displacement fields. Based on these open points, targeted approaches will be proposed to tackle the identified research issues.

Optical Flow Estimation

Today, continuous variational approaches are well established in optical flow estimation. The underlying mathematical concept has proven its sustainable value while rigorous assumptions, necessary in early approaches, have gradually been relaxed. The data term has developed from the pixel-wise brightness constancy assumption to elaborate robust descriptors handling a significant amount of projective distortion, outliers and noise. At the same time, smoothness constraints were adapted to more realistic scenarios including motion discontinuities and piece-wise motion models.

Variational approaches are shown to provide high accuracy results for sub-pixel displacements. However, the hierarchical coarse-to-fine strategy, which is necessary to cope with large displacements in a purely variational setting, is prone to errors at fine image details. Such small structures disappear at coarse resolution levels of the image pyramid. Consequently, the motion of fine details is often not recovered correctly. In addition, appropriate robust smoothness constraints require a tradeoff between convexity and more expressive models. The first is advantageous for the optimization procedure but typically imposes very strict assumptions on the motion field. TV-L1 regularization is a common choice, which is benign for optimization but prefers piece-wise constant flow. More flexible models allow for more general motion fields with abrupt discontinuities but typically lead to non-convex energy functionals. In the non-convex case, hierarchical continuous optimization can easily get stuck in arbitrary local minima of the objective function.

To address these challenges several options are proposed in the literature. One common direction of research is to initialize the displacement field with pixel-accurate matches and refine the result using variational post-processing. Section 3.1 refers to several methods establishing or refining sparse feature matches to retrieve approximate values for large displacements. These sparse feature matches are typically only available in image regions with appropriate texture and thus do not provide guidance in ambiguous areas. Therefore, it seems promising to investigate models, which are able to compute a more dense approximation of the optical flow field. The related paper on Flow Fields initializes the continuous optimization stage with a dense approximate nearest neighbor field tailored to optical flow. The option chosen in this thesis is to re-phrase the initial matching problem in terms of a discrete probabilistic graphical model. Conceptually, the solution space per node is discretized to a finite set of promising proposal displacement vectors. Given a sophisticated initialization procedure, optical flow vectors of arbitrary length may be investigated. Inference in such models can directly be conducted on full resolution imagery, as no linearization of data and smoothness terms is required. Regularization will be needed to cope with errors in the observations and to propagate plausible motion estimates to ambiguous image regions. In random field models, regularization is transferred to a probabilistic prior on the combination of labels at adjacent image sites.

The adoption of random field models to general large-displacement optical flow estimation is challenging. Although the integration of an optical flow prior is regarded promising, closely related work [Mozerov, 2013] is restricted to a small global set of flow proposals. An interesting question in this context is how to modify the initialization to effectively reduce the cardinality of the particle set while allowing for diverse large displacements. This question will be addressed in Section 4.1, which describes a model and processing pipeline enabling discrete optimization for optical flow. Special care is taken to adapt the model design and inference strategy to the specific challenges posed by large, sparse and unordered label sets. This point will be important as established efficient inference techniques cannot directly be applied under such conditions.

Scene Flow

As in the case of two-dimensional optical flow, variational approaches to scene flow estimation are conceptually able to determine small displacements with high accuracy. These continuous optimization methods lend themselves to the analysis of slowly moving, articulated surfaces with sufficient texture. In accordance with the preceding discussion, variational approaches depend on hierarchical strategies to determine large displacements. Consequently, they are susceptible to local minima and tend to miss small objects. To provide an adequate method for the envisaged applications of mobile robotics and autonomous driving, we will develop a conditional random field model and tackle the problem using approximate discrete inference.

In a general setting, the solution space will be intractably large, even when restricting the sought displacements to a plausible discretization in image space. To effectively tackle the ill-posed scene flow problem in terms of discrete optimization, related work proposes to reduce the number of unknowns by decomposing the reference view into segments and finding a parametrized, piece-wise smooth solution. This idea is extended by additionally decomposing the observed scene into a small number of individually moving objects, which further reduces the number of motion parameters to be estimated. While the closely related approach of Vogel and colleagues [2013b] solves a labeling task based on fixed plane and motion proposals, the method introduced in Section 4.2 jointly optimizes all discrete and continuous unknowns.

Object Model

In the literature, geometric object models serve two purposes, which are relevant in the scope of this thesis. A compact representation of important results is helpful for scene understanding. In addition, a feedback mechanism can effectively utilize such results to guide early vision algorithms. The introduction of object knowledge into scene flow estimation will be investigated in Section 4.3.

The parametrized reconstruction of deformable shape models yields the desired compact representation of results. It combines semantic object information with geometric properties of the observed instance. Restricting the problem to distinct object classes, the proposed extension of the basic scene flow approach will enable the incorporation of an active shape model. The derived information provides helpful cues in the context of scene understanding.

Object models are also used in the literature to constrain image-based reconstruction. Typically, high-level object information is introduced as a regularizer on reconstructed surfaces when inadequate texture or imperfect imaging conditions affect image matching. These challenges can also be expected to compromise the performance of scene flow estimation. To address this issue, the evolving shape model will be considered in the joint optimization of shape and motion.

Chapter 4

Methodology

This chapter presents the mathematical formulation of the proposed approaches and the ideas behind them. The basis of the methodology is developed in Section 4.1, which describes a general image matching approach tailored to optical flow estimation. Section 4.2 provides the details of a novel scene flow approach. The estimation of a dense three-dimensional displacement field is constrained by a decomposition of the observed scene into a set of individually moving objects. A semantic interpretation of this object-wise decomposition is introduced in Section 4.3 to further support scene flow estimation by sophisticated object knowledge in terms of a class-specific active shape model. Core assumptions of the models will be pointed out throughout the description of the methodology. A discussion of their influence and resulting theoretical limitations is provided in Section 4.4.

4.1 Optical Flow by Discrete Optimization

This section introduces a conditional random field model for general two-dimensional image matching. The proposed processing pipeline is designed to estimate displacements between subsequent frames of a monocular image sequence. Three strategies are proposed to enable discrete optimization for optical flow based on the investigations published in [Menze et al., 2015a]. In particular, it is shown how to compute a compact yet diverse set of proposals for each hidden variable in the random field model. Dynamic programming is applied in a block coordinate descent framework for inference. Finally, the robust formulation of the smoothness term is exploited to reduce the computational complexity of the optimization procedure and allow for discrete inference in reasonable time. Following denominations in the literature, and referring to the methodological core of the approach, the developed method is referred to as *DiscreteFlow*.

4.1.1 Optical Flow Model

Following successful works on stereo image matching, optical flow estimation is formulated as a discrete inference problem in a conditional random field model. The task is to infer an optimal discrete label $l_{\mathbf{p}}$ for each image site, which encodes the optical flow. Vector $\mathbf{l}$ contains the labels assigned to all pixels $\mathbf{p}$ in the image domain $\mathcal{P}$ of the reference frame.

$$E(\mathbf{l}) = \lambda \sum_{\mathbf{p} \in \mathcal{P}} \underbrace{\varphi_{\mathbf{p}}(l_{\mathbf{p}})}_{\text{data}} + \sum_{(\mathbf{p},\mathbf{q}) \in \mathcal{N}} \underbrace{\psi_{\mathbf{p},\mathbf{q}}(l_{\mathbf{p}}, l_{\mathbf{q}})}_{\text{smoothness}} \tag{4.1}$$

The objective function $E(\mathbf{l})$ serves to evaluate the consistency of a label assignment $\mathbf{l}$ with the formalized model assumptions. It is composed of a unary data term φ, and a pairwise smoothness term ψ. The former evaluates the similarity of image locations in subsequent frames associated by a label $l_{\mathbf{p}}$. The latter encourages locally smooth flow fields. $\mathcal{N}$ is the set of neighboring pixels. Here, it is defined on a 4-connected grid so the maximum cliques in the graph are of order two by design. The weight λ is chosen to balance the influence of both terms.

The data term φ returns the unary cost at pixel $\mathbf{p}$ given its label $l_{\mathbf{p}}$:

$$\varphi_{\mathbf{p}}(l_{\mathbf{p}}) = \min\left(\left\|\mathbf{d}_{\mathbf{p}} - \mathbf{d}'_{\mathbf{p}}(l_{\mathbf{p}})\right\|_1, \tau_{\varphi}\right) \tag{4.2}$$

For robustness, the data term evaluates the truncated l_1 difference of dense DAISY descriptors [Tola et al., 2010] with truncation threshold τ_{φ}. $\mathbf{d}_{\mathbf{p}}$ denotes the descriptor at pixel $\mathbf{p}$ in the reference image and $\mathbf{d}'_{\mathbf{p}}(l_{\mathbf{p}})$ is the descriptor in the target image, associated by label $l_{\mathbf{p}}$. Using a robust data term is important to ensure a sufficiently large influence of the smoothness prior. In cases where the descriptors do not support a correct match, e.g. due to strong scale change or at image points close to motion boundaries, the unary potential will not induce higher penalty than τ_{φ} and the smoothness term will be able to influence the label assignment.

The smoothness term ψ implements the assumption that neighboring image locations $(\mathbf{p}, \mathbf{q})$ will typically exhibit similar optical flow values. In particular, ψ evaluates the l_1 difference of neighboring flow vectors truncated at threshold τ_{ψ}.

$$\psi_{\mathbf{p},\mathbf{q}}(l_{\mathbf{p}}, l_{\mathbf{q}}) = w_{\mathbf{p},\mathbf{q}} \ \min\left(\left\|\mathbf{f}_{\mathbf{p}}(l_{\mathbf{p}}) - \mathbf{f}_{\mathbf{q}}(l_{\mathbf{q}})\right\|_1, \tau_{\psi}\right) \tag{4.3}$$

Here, $\mathbf{f}_{\mathbf{p}}(l_{\mathbf{p}})$ returns the two-dimensional displacement vector at pixel $\mathbf{p}$ induced by a distinct label $l_{\mathbf{p}}$. A robust smoothness prior allows for motion discontinuities as it circumvents the problem of over-smoothing at motion boundaries. In the presented approach, the truncated pairwise penalty

is also crucial for efficient inference. Section 4.1.2 describes how the truncation is exploited to reduce the computational complexity during inference.

Motion discontinuities can be encouraged to coincide with significant image edges. To this end, the pairwise potential is weighted by the factor $w_{\mathbf{p},\mathbf{q}}$ that considers the local strength of image edges. It reduces the influence of the smoothness prior between neighboring image locations on opposite sides of an image edge.

$$w_{\mathbf{p},\mathbf{q}} = \exp\left(-\alpha\, \kappa^2_{\mathbf{p},\mathbf{q}}\right) \tag{4.4}$$

Here, $\kappa_{\mathbf{p},\mathbf{q}} \in [0, 1]$ measures the strength of the edge between neighboring image sites $\mathbf{p}$ and $\mathbf{q}$. The parameter α controls the shape of the weighting function.

4.1.2 Efficient Inference

The model proposed in Section 4.1.1 specifies a loopy graph, as depicted in Figure 2.2, with general potential functions. Consequently, global optimization of the objective function is NP-hard. For large label sets, as conceptually encountered in optical flow estimation, even approximate inference is computationally intractable. In this section, three strategies are described that allow for approximate inference in reasonable time. Figure 4.1 provides an overview of the proposed methodology and visualizes the basic ideas.

Diverse Flow Proposals

One major challenge, limiting the applicability of discrete optimization, is the cardinality of the required label set. While in typical labeling tasks, like semantic segmentation, the number of possible labels ranges from two to a low double-digit amount, this problem is especially pronounced in general, two-dimensional image matching. Naïve discretization of the 2D flow space would return all pixels within a regular search window bounded by the maximum admissible flow amplitude or the size of the target image, respectively. To reduce this potentially very large set of candidate matches, approximate nearest-neighbor search is applied in descriptor space. The goal is to identify the most promising match hypotheses within a potentially large search window. While a variety of different image descriptors are available, in this work, the DAISY descriptor by [Tola et al., 2010] is applied. It can be computed efficiently for each pixel of an input image and closely resembles the successful SIFT descriptor [Lowe, 2004]. In the present approach, DAISY descriptors are computed for all pixels in both images.

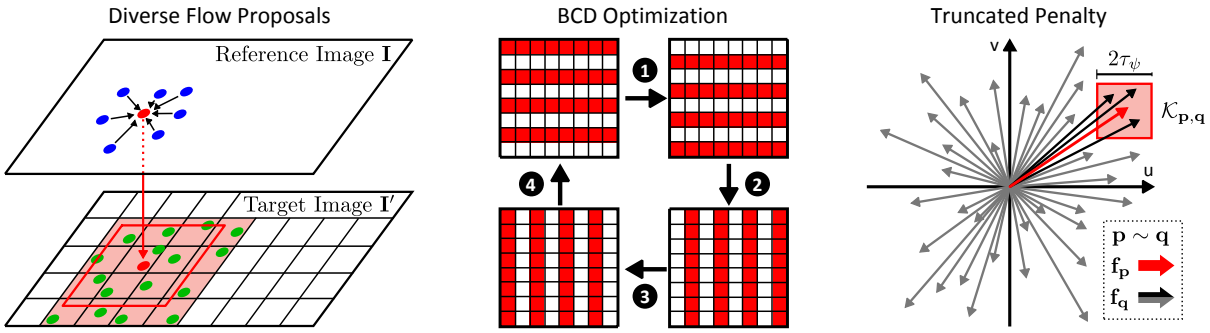


Figure 4.1: *Strategies enabling efficient discrete optimization for optical flow. Left: A large set of diverse flow proposals is computed per pixel (red node) by combining nearest neighbors in feature space from a set of grid cells (green nodes) with winner-takes-all solutions from neighboring pixels (blue nodes). The red square indicates the search region. Center: Dynamic programming is applied in a block coordinate descent framework, iteratively optimizing all image rows and columns (red) conditioned on neighboring blocks (white). Right: Taking advantage of robust penalties, we reduce pairwise computation costs by pre-computing the set of non-truncated ($< \tau_\psi$) neighboring flow proposals (black) for each flow vector (red).*

Nearest neighbor search in the high-dimensional descriptor space is challenging for two reasons. On the one hand, exhaustive search over the complete set of possible matches is prohibitively slow. On the other, nearest-neighbor matches tend to concentrate on a few promising image regions. To provide an efficient search structure, the target image is divided into a number of regularly sized cells. For each of the cells a randomized kd-tree is built, comprising the descriptors of all contained pixels. Kd-trees are one efficient approach to approximate nearest neighbor search in high-dimensional spaces and are used as a representative of this class of algorithms. A relaxed search range of the maximum expected displacement is applied to determine which cells to query for each reference pixel. Ensuring a uniform distribution of match candidates over all queried cells approximates non-maximum suppression and increases the diversity of the derived proposal set. A visualization of the proposal generation is provided in the left panel of Figure 4.1.

As for the smoothness potential, we make use of the observation that neighboring image locations tend to exhibit similar optical flow values. Hence, the set of nearest-neighbor match hypotheses is complemented by a set of proposals derived from neighboring pixels in the reference image. To this end, a number of image sites are randomly sampled in the vicinity of the reference pixel. The respective nearest-neighbor matches of the sampled image locations are added to enrich the proposal set. Care is taken not to duplicate proposals, which are already contained in the particle set. Whenever one of the randomly chosen proposals already exists in the label set of the reference pixel, the next best match hypothesis is added.

Block Coordinate Descent Optimization

The set of diverse flow proposals is significantly smaller than the complete set of all possible matches. However, it is still large enough to be computationally demanding for existing approximate inference techniques. Following [Chen and Koltun, 2014], this section generalizes an efficient optimization strategy, originally proposed for stereo image matching, to two-dimensional optical flow estimation.

As described in Section 2.3.3, the basic idea of block coordinate descent is to iteratively optimize individual rows and columns of the regular random field model conditioned on the remaining, fixed variables. The general optimization approach can directly be transferred to the proposed grid-structured optical flow model. After generating the proposal set all random variables are initialized based on one iteration of dynamic programming on the rows and columns, respectively. In general, the resulting labels from optimizing the two groups of blocks will differ. To find the initial assignment, one of the resulting labels at each node is chosen randomly. This procedure will typically not yield a local optimum of the energy function. It is designed to provide reasonable approximate values for the subsequent iterative optimization. During the main inference procedure, each row and column is assigned a flag reporting whether or not the last iteration changed the label of one of the contained nodes. As soon as no more changes occur, the labels in the respective chain are fixed and removed from the optimization. The inference procedure is stopped, when no change of labels reduces the overall energy.

Due to the more complex structure of the pairwise penalty for two-dimensional displacements, the original approach is modified concerning the handling of pairwise terms. In the present formulation, labels implicitly encode displacement vectors in the sparse set of proposals from the preceding section. This is in contrast to one-dimensional image matching along epipolar lines or dense discretization of two-dimensional displacements where the labels directly represent the sought offset. In the presence of such ordered label sets the efficient distance transform proposed by [Felzenszwalb and Huttenlocher, 2006] can be applied to exploit the computational advantages explained in Section 2.3.3. In our case, this technique is not applicable and we have to resort to an alternative strategy to reduce the computational burden.

Exploitation of the Truncated Penalty

Instead of imposing ordering constraints on the label set, the truncated form of the pairwise potential in Equation 4.3 is exploited to accelerate dynamic programming. Given the truncation threshold τ_ψ in the smoothness term, a reduced set of *relevant* neighboring proposals $\mathcal{K}_{\mathbf{p},\mathbf{q},l}$ can be pre-computed. This reduced set depends on the image sites $\mathbf{p}$ and $\mathbf{q}$ forming the clique, and on the label l of the reference node.

$$\mathcal{K}_{\mathbf{p},\mathbf{q},l} = \{k \in \{1,\ldots,L\} \,|\, \|\mathbf{f}_{\mathbf{p}}(l) - \mathbf{f}_{\mathbf{q}}(k)\|_1 < \tau_\psi\} \tag{4.5}$$

The set of relevant proposals $\mathcal{K}$ contains all labels k at pixel $\mathbf{q}$ for which the flow $\mathbf{f}_{\mathbf{q}}(k)$ is within τ_ψ from the flow $\mathbf{f}_{\mathbf{p}}(l)$ associated with label l at pixel $\mathbf{p}$. The rightmost panel of Figure 4.1 illustrates $\mathcal{K}_{\mathbf{p},\mathbf{q}}$ for a single flow vector at pixel $\mathbf{p}$ which is shown in red. A set of flow vectors at pixel $\mathbf{q}$ is depicted with black arrows indicating relevant proposals and gray arrows exceeding τ_ψ. Based on the set $\mathcal{K}$, the computation of the actual pairwise penalty is restricted to those neighboring flow vectors that fall below the truncation threshold. All remaining proposals yield a fixed pairwise penalty, namely τ_ψ.

Given the definition in Equation 4.5, the computation of the most costly part of dynamic programming can be accelerated. Without further modification the last term of Equation 2.14 induces quadratic complexity in the number of labels. Using $\mathcal{K}$ it can be re-formulated as

$$\min\left(\min_{k \in \mathcal{K}_{(x,y),(x-1,y),l}} \left(\psi_{(x,y),(x-1,y)}(l,k) + \mathbf{C}(x-1,k)\right),\, c \right) \tag{4.6}$$

where the first argument represents the unavoidable computation of pairwise penalties with respect to the reduced label set. The constant c takes the value of the minimal combination of cumulative cost $\mathbf{C}(x-1,k)$ at the preceding pixel and the maximum pairwise penalty for an arbitrary change of displacements. It is given by

$$c = \min_{0 \le k < L} \left(w_{(x,y),(x-1,y)} \cdot \tau_\psi + \mathbf{C}(x-1,k)\right) \tag{4.7}$$

Importantly, due to the fixed, maximum pairwise penalty, c does not depend on the label l of the reference pixel. Thus, it is valid for all l and can be computed in $O(L)$. Based on this consideration, Equation 4.6 can be evaluated (for all l) in $O(\sum_l |\mathcal{K}_{\mathbf{p},\mathbf{q},l}|)$ instead of $O(L^2)$, where $|\mathcal{K}_{\mathbf{p},\mathbf{q},l}| \ll L$ in practice due to the diversity of the proposal set as evidenced by the experimental evaluation in Section 5.2.2. It is also important to note that the sets $\mathcal{K}_{\mathbf{p},\mathbf{q},l}$ can be pre-computed before starting block coordinate descent. They are reused during all iterations.

Outlier Removal

An important step in image matching approaches is the validation of established correspondences. In stereo approaches, it is common to verify associations from the left image to the right one using the redundant results of matching the images in reversed order, i.e. right to left. Only associations, which are similarly established in both passes up to some outlier threshold, are

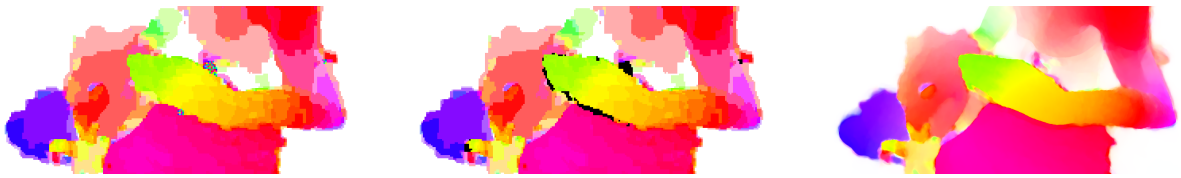


Figure 4.2: *Outlier removal and post-processing. All images show color-coded optical flow maps. The left panel shows the result of the forward pass of discrete optimization. Black regions in the center image visualize segments, which are removed by outlier rejection. The right panel provides the final result after EpicFlow post-processing.*

accepted as valid matches. The same validation strategy is commonly applied to optical flow estimation, as for example in the related approach of [Mozerov, 2013]. The proposed processing pipeline implements a similar forward-backward check. To this end, correspondences are also computed using the second frame as the reference view. Both resulting flow fields are compared to identify and remove inconsistent matches.

The forward-backward check is complemented by a second strategy to identify unreliable matches. Often, small isolated segments result from noise or occluded regions. To identify implausible segments, connected component analysis is applied to the resulting optical flow map. A simple region-growing algorithm returns segments of similar displacements, which do not contain discontinuities greater than a chosen consistency threshold. The area of the retrieved segments is compared to a chosen threshold on the desired minimum size. Smaller segments are removed from the result. The effect of outlier removal is depicted in the center of Figure 4.2.

Variational Refinement

The approach presented so far yields an integer-valued optical flow field as shown in the center of Figure 4.2. The displacement is undefined at image locations which did not pass the outlier removal procedure. To interpolate missing values and to refine the result to sub-pixel accuracy, the respective post-processing steps of EpicFlow, as discussed in Section 3.1, are applied. To make this section self-contained, the major steps of the refinement procedure are briefly reviewed.

The chosen approach consists of two stages. First, the results of sparse feature matching are interpolated to retrieve an initial dense displacement field. Interpolating the displacement vector at a distinct pixel involves the computation of a weighted average of neighboring correspondences. As the simple Euclidean distance in the image plane does not respect motion boundaries it is replaced by a geodesic distance taking into account image edges. Assuming that motion boundaries coincide with a subset of all image edges this approach circumvents interpolation between sparse matches on differently moving objects. In the original publication EpicFlow interpolates the results of DeepMatching. In the present work, these correspondences are replaced with the result of the discrete optimization procedure described in the preceding sections.

The interpolated dense displacement field does not incorporate additional image information apart from the edges. It serves to initialize the variational refinement described in [Brox et al., 2004] to exploit the high accuracy potential of continuous optimization without applying the error-prone hierarchical strategy to compute approximate values for large displacements. The data term of the functional is composed of variants of the classical color and gradient constancy assumptions [Zimmer et al., 2011]. For regularization, the smoothness constraint from Equation 2.4 is applied, which is down-weighted at image edges to avoid smoothing of the preserved motion boundaries. The right panel of Figure 4.2 depicts the estimated dense optical flow field after interpolation and variational sub-pixel refinement. The missing regions are filled in and obvious discretization artifacts are smoothed. A detailed evaluation of the described processing steps is provided in Section 5.2.

4.2 Object Scene Flow

The preceding section addresses motion field estimation in the image domain as a typical instance of early vision. The resulting optical flow field is an essential observation for three-dimensional displacement field estimation. In this section, the focus is on the classical scene flow setting where the input data consists of two consecutive stereo image pairs. The goal of the investigations is to determine a dense reconstruction of the scene together with the three-dimensional displacement of each reconstructed point between subsequent time steps. Section 4.2.1 describes the basic scene flow model based on the theoretical foundations of random fields provided in Section 2.3.1. Next, the choice of data and smoothness terms is discussed in Sections 4.2.2 and 4.2.3. The last two sections provide details of the initialization (4.2.4) and the inference procedure (4.2.5).

4.2.1 Scene Flow Model

Joint estimation of geometry and three-dimensional motion of an observed scene are enabled by processing stereoscopic image sequences. We make the following general assumptions about the available imagery. Throughout this thesis, the interior orientation of the cameras and the relative orientation of the stereo heads is assumed to be known. Based on this information the images are rectified so that epipolar lines are projected to corresponding image rows and stereo image matching reduces to a one-dimensional association problem. Furthermore, the synchronization of the stereo cameras is regarded as sufficiently accurate to neglect influences induced by offset exposure.

Following prior art, summarized in Section 3.2, we employ a slanted-plane model to capture geometry and motion. In particular, we assume that the variable three-dimensional structure of the scene can be approximated by a set of piecewise planar surface elements, each undergoing

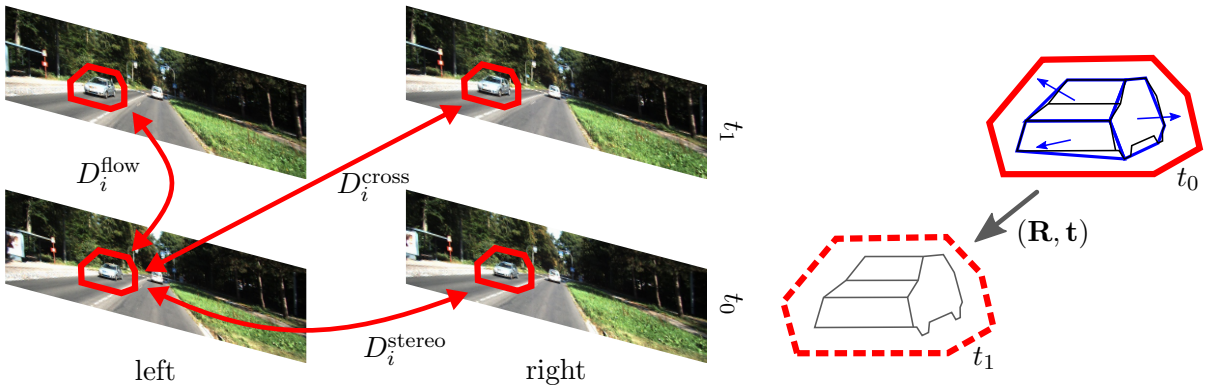


Figure 4.3: *Scene flow data terms and parametrization. Each superpixel in the reference view is associated with a plane in three-dimensional space, shown in blue on the right, and an object label. The label decides, which object motion is inherited by the segment. One individually moving object is marked in red. To compute the data terms D the superpixels are projected into each of the remaining input images where a similarity measure is evaluated.*

a rigid body transformation. These surface elements are associated with image segments, which completely cover the image domain of a reference view. The use of image segments helps to constrain the problems at hand, as the estimation of plane and motion parameters becomes over-determined. It is important to allow for sufficient flexibility of the surface and motion model. To capture all significant discontinuities in the sought entities, a significant over-segmentation of the image is required.

The major novelty of the presented model is the assumption that the observed scene decomposes into a finite number of *rigidly moving objects*. To emphasize this feature the proposed method is referred to as *Object Scene Flow*. As will be discussed in Section 4.4, this fundamental assumption holds true for a number of relevant applications. It is important to note that the static elements in the scene, which will be referred to as the *background*, can be easily handled as one of the objects in the proposed formulation. Like individually moving foreground objects, these parts of the scene are assumed to move rigidly with respect to the observer. For moving cameras in static environments, this single background object is able to capture the complete observed motion. Based on the decomposition into objects, motion estimation simplifies to the optimization of a small number of transformation parameters. Technically, the resulting random field model contains two different types of hidden variables representing the superpixels and the objects.

To formalize the model in terms of a conditional random field, let $\mathcal{S}$ denote the set of superpixels and $\mathcal{O}$ denote the set of objects. Each individual superpixel $\mathbf{s}_i \in \mathcal{S}$ is associated with a region $\mathcal{R}_i$ in the reference image and a random variable $(\mathbf{n}_i, l_i)$. In particular, $\mathbf{n}_i \in \mathbb{R}^3$ describes a plane in 3D by its normal, scaled with the shortest distance from the origin. Thus, $\mathbf{n}_i^T \mathbf{X} = 1$ for points $\mathbf{X} \in \mathbb{R}^3$ on the plane. Further details of the parametrization are provided in Appendix A. The discrete label $l_i \in \{1, \ldots, |\mathcal{O}|\}$ assigns each superpixel to one of the objects. Each object $\mathbf{o}_k \in \mathcal{O}$ is associated with a random variable $(\mathbf{R}_k, \mathbf{t}_k) \in SE(3)$, which contains a rotation matrix and a

translation vector describing its rigid body motion in 3D. Note that each superpixel associated with object $\mathbf{o}_k$ inherits its rigid motion parameters. In combination with the plane normal $\mathbf{n}_i$, this fully determines the three-dimensional scene flow of the surface element. The parametrization of one object in terms of surface elements is sketched in the right panel of Figure 4.3. It depicts a schematic approximation of the visible surfaces of a car at t_0 by three planes in three-dimensional space. The outlines of the planes are drawn in blue with blue arrows representing their normals $\mathbf{n}_i$. In the sketch, all three planes are assigned to a single object, which is visualized by the solid red polygon. Consequently, a common rigid body transformation, represented by the black arrow, defines their three-dimensional motion between the considered time steps t_0 and t_1. The transformed object is shown in the dashed red polygon.

The model assumptions specify a conditional random field in terms of the following energy function

$$E(\mathbf{s}, \mathbf{o}) = \sum_{i \in \mathcal{S}} \underbrace{\varphi_i(\mathbf{s}_i, \mathbf{o})}_{\text{data}} + \sum_{(i,j) \in \mathcal{N}} \underbrace{\psi_{ij}(\mathbf{s}_i, \mathbf{s}_j)}_{\text{smoothness}} \tag{4.8}$$

where $\mathbf{s} = \{\mathbf{s}_i | 1 \leq i \leq |\mathcal{S}|\}$ and $\mathbf{o} = \{\mathbf{o}_k | 1 \leq k \leq |\mathcal{O}|\}$. $\mathcal{N}$ denotes the set of adjacent superpixels in $\mathcal{S}$. As opposed to the pre-defined neighborhood relations in a grid-structured model, the number of neighboring superpixels is not fixed. For each image site, it corresponds to the number of image segments sharing boundary pixels. The respective adjacency matrix is computed based on the segmentation of the reference frame. A relative weight between data and smoothness term is included in the individual components, as described in the following sections. This allows for a fine-grained adjustment of the components' influence on the result. Given the objective function 4.8, our goal is to jointly infer the geometry of each segment $\mathbf{n}_i$, the association l_i of superpixels to objects, and the rigid body motion $(\mathbf{R}_k, \mathbf{t}_k)$ of each object $\mathbf{o}_k$.

To constrain the problem we make use of different observations in the data term, all of which will be explained in detail in the next section. The smoothness term implements the model assumptions that depth and motion will vary smoothly between neighboring image segments except for the case of abrupt, significant changes of the respective entities. All pairwise penalties, forming the smoothness term, will be discussed in Section 4.2.3.

4.2.2 Data Term

The data term of the random field model (4.8) serves to evaluate the compatibility of proposed parameters and the observed images. In the present approach, it implements the assumption that corresponding image locations should be similar in appearance across the four input images. This similarity assumption is enforced by penalizing the dissimilarity between a segment in the reference view and its projection to all three remaining images. The necessary transformations are

established by the combination of shape parameters and the rigid body transformation, inherited from the assigned object. As the data term relies on complementary information from the different types of hidden variables, it defines a pairwise potential between segments and objects

$$\varphi_i(\mathbf{s}_i, \mathbf{o}) = \sum_{k \in \mathcal{O}} [l_i = k] \cdot D_i(\mathbf{n}_i, \mathbf{o}_k) \tag{4.9}$$

where $[\cdot]$ denotes the Iverson bracket. It returns 1 if the argument is true and 0 otherwise. Consequently, φ_i is only evaluated with respect to the currently assigned object. The function $D_i(\mathbf{n}_i, \mathbf{o}_k)$ returns a dissimilarity measure for superpixel $\mathbf{s}_i$, which depends on plane parameters $\mathbf{n}_i$ and the rigid body motion of the assigned object $\mathbf{o}_k$. To gather information from all images, the dissimilarity measure is composed of a stereo, a flow and a cross term. The three terms are computed between a reference view and the other three images, as illustrated in the left part of Figure 4.3. Without loss of generality, the left image at t_0 is defined as the reference view throughout this thesis. The complete dissimilarity measure in the data term reads as follows:

$$D_i(\mathbf{n}, \mathbf{o}) = D_i^{\text{stereo}}(\mathbf{n}, \mathbf{o}) + D_i^{\text{flow}}(\mathbf{n}, \mathbf{o}) + D_i^{\text{cross}}(\mathbf{n}, \mathbf{o}) \tag{4.10}$$

Again, the relative weighting is postponed to the following detailed description to allow for fine-grained control. Each of the constituting terms is defined in Equation 4.11 as the sum of matching costs C over all pixels $\mathbf{p}$ inside superpixel s_i. Matching costs are computed by warping each pixel according to a homography induced by the associated geometry and motion. The comparison of image sites around the reference pixel and the transformed target pixel can be expressed as:

$$D_i^{\text{x}}(\mathbf{n}, \mathbf{o}) = \sum_{\mathbf{p} \in \mathcal{R}_i} C_{\text{x}}\big(\, \mathbf{p}, \underbrace{\mathbf{K}\left(\mathbf{R}_{\text{x}}(\mathbf{o}) - \mathbf{t}_{\text{x}}(\mathbf{o}) \cdot \mathbf{n}^T\right)\mathbf{K}^{-1}}_{3\times3 \text{ homography}} \mathbf{p}\,\big) \tag{4.11}$$

Here, x refers to one of the different matching modalities specified above: $\text{x} \in \{\text{stereo}, \text{flow}, \text{cross}\}$. $\mathbf{K} \in \mathbb{R}^{3\times3}$ denotes the camera calibration matrix containing the elements of the interior orientation. For clarity of presentation, the interior orientation of the left and right camera is assumed to be equal. The transformation parameters $(\mathbf{R}_{\text{x}}(\mathbf{o}), \mathbf{t}_{\text{x}}(\mathbf{o}))$ are applied to map a 3D point in reference coordinates to a 3D point in another camera coordinate system according to the relative camera orientation and the rigid body motion of $\mathbf{o}_k$. To directly transform homogeneous image coordinates from one view to another a two-dimensional projective transformation can be applied. The corresponding homography matrix is composed of two planar projections. First, the reference pixel is transformed to a three-dimensional object point in the plane of its superpixel. Next, it is mapped to the respective partner image. For the stereo term the original plane parameters are used while for projections to images at t_1 the plane normal is transformed according to the object

motion. Consequently, $\mathbf{R}_\mathrm{x}$ and $\mathbf{t}_\mathrm{x}$ depend on the matching modality x and are augmented with the parameters of the relative camera orientation where necessary.

The matching cost $C_\mathrm{x}(\mathbf{p}, \mathbf{q})$ returns a dissimilarity measure between a pixel at location $\mathbf{p} \in \mathbb{R}^2$ in the reference image and a pixel at location $\mathbf{q} \in \mathbb{R}^2$ in the target image. In the proposed model, we take advantage of dense correspondences as well as sparsely matched image features. Matching costs $C_\mathrm{x}(\mathbf{p}, \mathbf{q})$ are defined as a weighted sum of these two groups of observations with individual weights θ:

$$C_\mathrm{x}(\mathbf{p}, \mathbf{q}) = \theta_{1,\mathrm{x}}\, C_\mathrm{x}^\mathrm{dense}(\mathbf{p}, \mathbf{q}) + \theta_{2,\mathrm{x}}\, C_\mathrm{x}^\mathrm{sparse}(\mathbf{p}, \mathbf{q}) \tag{4.12}$$

The dense matching cost $C_\mathrm{x}^\mathrm{dense}(\mathbf{p}, \mathbf{q})$ is computed as the Hamming distance $\|\cdot\|_\mathrm{H}$ of the respective 5×5 census descriptors, truncated at C_max. This patch-based similarity measure was introduced by [Zabih and Woodfill, 1994] and found to work well in the context of optical flow estimation [Vogel et al., 2013a]. It efficiently builds a descriptor of a small image region by concatenating the binary results of intensity value comparisons. Consequently, it is robust against additive changes in illumination. We assign an outlier value of C_out to projected points $\mathbf{q}$ leaving the image domain Ω of the left frame at t_1. The same parameter values C_max and C_out are used for the densely computed stereo, flow and cross terms.

$$C_\mathrm{x}^\mathrm{dense}(\mathbf{p}, \mathbf{q}) = \begin{cases} \rho_{C_\mathrm{max}}\left(\|\mathbf{p} - \mathbf{q}\|_\mathrm{H}\right) & \text{if } \mathbf{p} \in \Omega \\ C_\mathrm{out} & \text{otherwise} \end{cases} \tag{4.13}$$

While the employed census descriptor accounts for small deviations from the similarity assumption, it cannot cope with systematic influences like occlusions or perspective distortion. To limit the effect of such grossly wrong observations, e.g. next to depth discontinuities, a robust penalty function is $\rho_\tau(x)$ is applied to compute the matching cost. It truncates the distance x at threshold τ, which yields the penalty function $\rho_\tau(x) = \min(x, \tau)$.

In addition to the dense matching term, a second type of observations is used. It evaluates the consistency of displacements induced by the estimated parameters and those computed by specialized, semi-dense matching approaches, which will be explained in more detail in Section 4.2.4. As these additional observations are only available at a subset of image locations, the following case discrimination is conducted:

$$C_\mathrm{x}^\mathrm{sparse}(\mathbf{p}, \mathbf{q}) = \begin{cases} \rho_{\tau_1}\left(\|\pi_\mathrm{x}(\mathbf{p}) - \mathbf{q}\|_2\right) & \text{if } \mathbf{p} \in \Pi_\mathrm{x} \\ 0 & \text{otherwise} \end{cases} \tag{4.14}$$

Here, $\pi_{\mathrm{x}}(\mathbf{p})$ denotes the warping of pixel $\mathbf{p}$ according to the set of sparse feature correspondences, $\mathbf{q}$ is the result of warping the reference pixel according to the estimated parameters as before. Π_{x} is the set of pixels in the reference image for which correspondences have been established. As before, x refers to the matching modality and the truncation threshold τ_1 limits the influence of outliers in the observations.

4.2.3 Smoothness Term

The task of the smoothness term in Equation 4.8 is to encourage smooth transitions between adjacent superpixels. To regularize the shape of the surface and to prefer compact objects, the following smoothness term is defined on the CRF. Weights θ control the influence of the three constituting terms:

$$\psi_{ij}(\mathbf{s}_i, \mathbf{s}_j) = \theta_3\, \psi_{ij}^{\mathrm{depth}}(\mathbf{n}_i, \mathbf{n}_j) + \theta_4\, \psi_{ij}^{\mathrm{orientation}}(\mathbf{n}_i, \mathbf{n}_j) + \theta_5\, \psi_{ij}^{\mathrm{motion}}(\mathbf{s}_i, \mathbf{s}_j) \tag{4.15}$$

First, regularization of depth is achieved by penalizing different disparity values d at shared boundary pixels $\mathcal{B}_{ij}$. The relevant image locations are pre-computed together with the adjacency matrix of the superpixels. This well established constraint is derived from the observation that surfaces of distinct objects typically exhibit only gradual changes of geometry. As the presented approach relies on an over-segmentation of the reference image, which decomposes consistent surfaces into fragments, it is necessary to impose regularization on the surface geometry. To allow for depth discontinuities, as typically encountered at object boundaries, the robust penalty function ρ_{τ_2} is applied limiting the penalty to τ_2:

$$\psi_{ij}^{\mathrm{depth}}(\mathbf{n}_i, \mathbf{n}_j) = \sum_{\mathbf{p} \in \mathcal{B}_{ij}} \rho_{\tau_2} \left(\| d(\mathbf{n}_i, \mathbf{p}) - d(\mathbf{n}_j, \mathbf{p}) \|_1 \right) \tag{4.16}$$

Here, the function $d(\mathbf{n}, \mathbf{p})$ returns the disparity at pixel $\mathbf{p}$ induced by the plane normal $\mathbf{n}$ of the respective segment.

Second, the orientation of neighboring planes is encouraged to be similar. This is a necessary extension of the preceding pairwise term to fully formalize the aforementioned smoothness assumption. While the requirement of consistent disparity at boundary pixels attaches neighboring segments, it does not penalize implausible folds in the reconstructed surface. To this end, the similarity of plane normals $\mathbf{n}$ is evaluated using the cosine similarity. Again, a threshold is applied to allow for sudden changes of surface orientation where needed.

$$\psi_{ij}^{\mathrm{orientation}}(\mathbf{n}_i, \mathbf{n}_j) = \rho_{\tau_3} \left(1 - \frac{|\mathbf{n}_i^T \mathbf{n}_j|}{(\|\mathbf{n}_i\| \|\mathbf{n}_j\|)} \right) \tag{4.17}$$

Finally, coherence of the assigned object indices is enforced by an orientation-sensitive Potts model in $\psi_{ij}^{\text{motion}}$. The task of this term is to penalize fragmented objects, as typically objects in the observed scene project to connected image regions. The Iverson bracket activates this penalty wherever neighboring superpixels are assigned to different objects. This setting, in general, introduces a motion boundary in the result and should only occur where it is supported by profound image evidence.

$$\psi_{ij}^{\text{motion}}(\mathbf{s}_i, \mathbf{s}_j) = w(\mathbf{n}_i, \mathbf{n}_j) \cdot [l_i \neq l_j] \tag{4.18}$$

The weight $w(\cdot, \cdot)$ in the coherence term is defined as

$$w(\mathbf{n}_i, \mathbf{n}_j) = \frac{|\mathbf{n}_i^T \mathbf{n}_j|}{(\|\mathbf{n}_i\| \|\mathbf{n}_j\|)} \cdot \exp\left(-\frac{\alpha}{|\mathcal{B}_{ij}|} \sum_{\mathbf{p} \in \mathcal{B}_{ij}} (d(\mathbf{n}_i, \mathbf{p}) - d(\mathbf{n}_j, \mathbf{p}))^2 \right) \tag{4.19}$$

and prefers motion boundaries that coincide with folds in the reconstructed surface. Here, α is the shape parameter of the penalty function which is normalized by the number of shared boundary pixels $|\mathcal{B}_{ij}|$. The penalty induced at motion discontinuities increases in situations where the surface orientation of the compared superpixels is similar.

4.2.4 Initialization

The model specified in the previous sections contains a mixture of discrete and continuous variables. While the label l_i corresponds to a discrete object index, plane normals $\mathbf{n}_i$ and motion parameters $(\mathbf{R}_k, \mathbf{t}_k)$ live in continuous domains. Before the discrete inference procedure, described in the next section, can be applied, some of the observations are pre-computed and all variables need to be initialized. Due to the non-convex objective function, proper approximate values are required.

Image Segments and Initial Correspondences

Section 4.2.2 specifies a data term, which is composed of dense census features and additional sparse observations. The latter consist of pre-computed disparity maps for both stereo pairs and semi-dense optical flow matches between both images of the left camera.

In particular, SPS-Stereo [Yamaguchi et al., 2014], as described in Section 3.2, is applied as a state-of-the-art stereo matching technique. Because the underlying assumption of a static scene holds only for simultaneously captured images, the dedicated stereo variant of the algorithm is

Figure 4.4: *Typical over-segmentation of a car from SPS-Stereo [Yamaguchi et al., 2014]. The figure shows a magnified part of a representative input image.*

employed. In addition, SPS-Stereo computes an over-segmentation of the reference image into a parametrized number of compact superpixels. Throughout this thesis, the parameter setting from the original publication is employed as it is trained on the KITTI data set, which will be used in the experiments in Chapter 5. Typical images from this data set are decomposed into approximately one thousand segments, which is shown to yield best performance in the paper. The segmentation relies on color cues as well as the disparity map and provides the superpixels and initial plane parameters for Object Scene Flow. Where necessary, the disparity plane normals can be re-parametrized to an equivalent 3D representation given the parameters of the camera orientation. The mathematical details are provided in Appendix A.

One typical over-segmentation of a car is depicted in Figure 4.4. While most of the outline is faithfully recovered, shadows lead to segmentation errors around the rear wheel. The rear window also contains a spurious articulated segment. In contrast to the related approach of [Vogel et al., 2013b], the segmentation is not refined during inference to limit the computational burden. Thus, errors in the segmentation will directly transfer to bleeding artifacts at object boundaries. To counter this effect, superpixels are chosen to be sufficiently small to faithfully capture scene geometry and motion.

To guide the matching process between subsequent frames of the reference camera, semi-dense optical flow matches are introduced as additional observations. In Section 4.1, we discussed DiscreteFlow, which is designed to address this task. At this point, we do not need a dense correspondence field. Consequently, we discard EpicFlow post-processing and make use of the discrete matches after BCD optimization and outlier removal. Finally, both types of initial observations are combined to establish correspondences connecting the reference view to the right frame at t_1. In particular, optical flow vectors in the left image are combined with disparities in the left view at t_1 to predict the complete displacement vector for the cross term. These observations can only be evaluated where both, disparity and optical flow, are available. In Figure 4.3 they are shown as D_i^{cross}.

Figure 4.5: *Depth- and motion-based visibility mask. Image regions visible in the first frame but not in the second frame are indicated by the darker areas.*

Object Hypotheses

In addition to the segmentation of the reference view, the proposed method relies on the fundamental concept of objects. This section describes how they are instantiated based on the sparse observations described above. At this point, the application of object detectors is discarded for the sake of generality. As we are ultimately interested in motion estimation, we will employ motion segmentation to detect individually moving objects.

The combination of initial disparity and optical flow estimates establishes a set of scene flow correspondences. This sparse approximation of the sought three-dimensional displacement field is examined to reveal consistent motion patterns. First, the dominant motion is retrieved from the set of correspondences. It typically describes the relative motion of background and camera. An adequate approach to robust rigid motion estimation from two stereo image pairs is described in [Geiger et al., 2011]. Based on a uniformly distributed subset of N feature matches, the parameters of the dominant rigid body transformation are estimated. To this end, feature points from the reference frame are triangulated exploiting stereo correspondences and the known relative orientation of the stereo camera pair. The observation model is derived by projecting the triangulated three-dimensional object points to the images at t_1 according to the sought transformation. The respective projection equation is given by

$$\pi(\mathbf{X}, \mathbf{R}, \mathbf{t}) = \mathbf{K} \cdot ([\mathbf{R}|\mathbf{t}] \cdot [X, Y, Z, 1]^T - [s, 0, 0]^T) \tag{4.20}$$

with camera matrix $\mathbf{K}$, Rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$ as before. To map an object point to the right image at t_1, the stereo base s of the camera rig has to be considered, which is set to zero to map points to the left camera at t_1. Local optimization is applied to estimate transformation parameters, which minimize the combined re-projection error in both frames at t_1

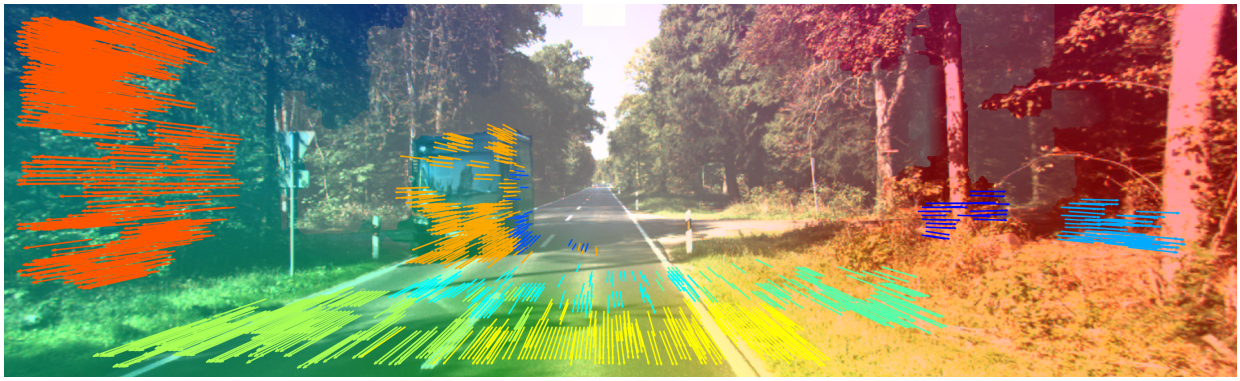


Figure 4.6: *Object hypotheses generation. This figure depicts the reference view superimposed by the color-coded dense optical flow field induced by the camera motion. The groups of similarly colored vectors correspond to the individually moving object hypotheses generated by motion segmentation. In addition to plausible clusters on the van there are erroneous hypotheses caused by shadows and vegetation.*

$$e = \sum_{i=1}^{N} \|\mathbf{x}_i^{(left)} - \pi^{(left)}(\mathbf{X}, \mathbf{R}, \mathbf{t})\|^2 + \|\mathbf{x}_i^{(right)} - \pi^{(right)}(\mathbf{X}, \mathbf{R}, \mathbf{t})\|^2 \tag{4.21}$$

In a least squares adjustment, the re-projection error is minimized to find optimal parameter values. The resulting transformation parameters define our first motion hypothesis. In addition, this transformation is used for the motion-based segmentation of the sparse scene flow observations.

Before further objects are extracted, the number of outliers in the correspondence set is reduced. To this end, image regions in the reference frame, which probably leave the image domains of the remaining views, are excluded. More specifically, all pixels with valid disparities in the first image of the left camera are triangulated and projected into the second view, based on the background motion. Static points falling outside the image domain cannot be matched correctly. They are removed from the set of sparse correspondences Π. One exemplary visibility mask is illustrated in Figure 4.5. This procedure potentially also removes some feature points lying on dynamic objects close to the image boundary. Such points could be matched correctly, but empirical tests indicate that the benefits this simple outlier removal strategy prevail.

In a last step, foreground object hypotheses are retrieved. Individually moving objects are defined as consistent clusters of scene flow correspondences, which do not agree with the background motion. To identify such outliers, a threshold of 5 pixels is applied to the endpoint error of motion vectors induced by the background motion and the sparse scene flow correspondences. Object hypotheses are retrieved using a simple clustering approach. From the set of outliers, we randomly sample 50 initial correspondences throughout the image. A three-dimensional rigid motion transformation is robustly fit to all correspondences within a ball of radius 2.5 meters around the initial matches using the 3-point RANSAC algorithm. The maximum admissible

Algorithm 1 Particle-based approximate energy minimization

Set number of particles N_s, N_o
Initialize superpixels $\mathbf{s}^0$ and objects $\mathbf{o}^0$
Initialize standard deviations σ_s^0, σ_o^0
for $j = 1$ to num_iterations **do**
    Sample N_o particles for each $\mathbf{o}$ from $\mathcal{N}(\mathbf{o}, \sigma_o)$
    Sample $0.5 \times N_s$ particles for each $\mathbf{s}$ from $\mathcal{N}(\mathbf{s}, \sigma_s)$
    Propagate $0.5 \times N_s$ particles to each $\mathbf{s}$
    $(\mathbf{s}^j, \mathbf{o}^j) \leftarrow$ Solve the discretized MRF using TRW-S
    Update $\sigma_i^j = \sigma_i^j \exp(-j/10)$
end for
return superpixels $\mathbf{s}^j$ and objects $\mathbf{o}^j$

number of object hypothesis is a parameter of the proposed approach. Section 5.3.3 discusses the choice of an appropriate value and its influence on the results. To retrieve the required number of object hypotheses the available rigid motion proposals are sorted with respect to the number of inliers. While choosing the top $|\mathcal{O}| - 1$ hypotheses, we apply non-maximum suppression to avoid multiple overlapping proposals on the same object. Whenever the center of an additional proposal is too close to an existing one, the smaller hypothesis is pruned. A result of this process is illustrated in Figure 4.6, where 9 motion hypotheses have been recovered. The color-coded clusters of consistent motion vectors are not only found on actual objects. They can also be caused by shadows or groups of outliers on vegetation for example. The proposed approach allows for a number of false positive object hypotheses. Usually they will not be assigned any image segments during inference. A few remaining ghost objects will also not affect the result as long as the estimated motion is close to that of the background.

4.2.5 Approximate Inference

The goal of inference is to jointly adjust the unknown discrete and continuous parameters subject to the model assumptions formalized in the objective function 4.8. Inspired by the work of [Yamaguchi et al., 2012], an iterative sampling-based algorithm is applied to solve a sequence of discrete random fields by approximately minimizing the objective function (4.8) in each iteration. To facilitate discrete inference, the respective solution spaces of the continuous hidden variables are discretized.

Meaningful approximate values are derived from the initialization procedure described in the preceding section. To refine the initial values a set of hypotheses is sampled around the current solution of each parameter. The final proposal set also contains the best solution so far as to ensure that the objective function decreases or remains constant during optimization. Before each iteration, the continuous random variables are discretized anew; before the first iteration, they are sampled in the vicinity of their initial values and subsequently around the current MAP

estimate. At each step, proposals are drawn from normal distributions with the current solution as the mean value. The variance of the distributions is reduced after each iteration to refine the discretization and encourage convergence. Due to the highly non-convex objective function this strategy is likely to encounter similar problems as variational optimization approaches and get stuck in poor local minima. This issue is addressed by complementing the proposal set of each superpixel with a number of MAP solutions from neighboring image sites. Thus, promising shape proposals are propagated to nearby image segments.

Optimization of the loopy CRF specified in Equation 4.8 with respect to all discrete and discretized superpixel and object parameters is an NP-hard combinatorial problem. As explained in the basics in Chapter 2, the incorporation of robust non-convex pairwise potentials renders global optimization inapplicable. Instead, each of the successive discrete random fields is approximately optimized based on the current proposal set. To this end, sequential tree-reweighted message passing (TRW-S) is applied. The complete particle-based approximate energy minimization approach is summarized in Algorithm 1. It bears similarities with smoothing-based optimization, which was proposed by [Leordeanu and Hebert, 2008]. In contrast to that approach, the standard deviation is heuristically reduced by a chosen factor in each iteration. While some local optima may be missed due to the strict, deterministic decrease of the search space, the parameter values should converge more quickly. As the approach provides no guarantee of convergence, the number of iterations is empirically chosen in Section 5.3.3.

4.3 Joint 3D Estimation of Vehicles and Scene Flow

The scene flow model introduced in the previous section is based on a decomposition of the observed scene into rigidly moving objects. So far, rigidity is the major assumption made with respect to the objects. In this section, the flexibility of the developed scene flow model is demonstrated. Adding a semantic interpretation to the motion-based segmentation, more sophisticated object knowledge can be incorporated into the model. Focusing on dedicated applications, we are able to further constrain scene flow estimation and to derive a parametrized reconstruction of objects. In particular, the method is specialized towards the perception of traffic scenes as encountered in the context of advanced driver assistance systems and autonomous driving. To this end, a geometric model of cars is introduced as described in a preliminary version in [Menze et al., 2015b]. A versatile three-dimensional representation of deformable geometric models is explained in Section 4.3.1 and the conditional random field from the preceding section is extended to allow for the estimation of shape and pose parameters of individual object models in Section 4.3.2. The core of the extension is a dedicated shape and pose penalty, which is described in detail in Section 4.3.3. Section 4.3.4 extends the initialization procedure from the preceding section to the new parametrization and Section 4.3.5 adapts the particle-based optimization procedure to jointly infer reconstructions and scene flow.

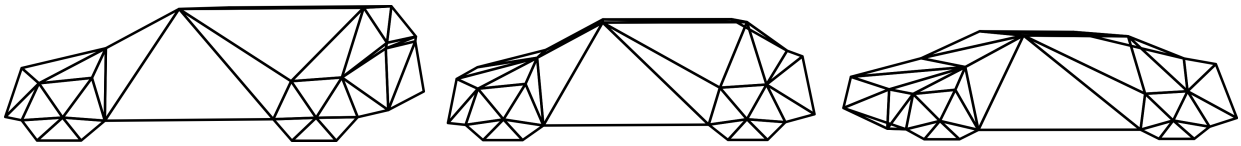


Figure 4.7: *Instances of the three-dimensional active shape model. The mean shape is shown in the center with* $\boldsymbol{\gamma} = (0, 0)$*. The left and right panels contain instances illustrating the range of possible deformations with shape parameters* $\boldsymbol{\gamma}_{left} = (-1.0, -0.8)$ *and* $\boldsymbol{\gamma}_{right} = (1.0, 0.8)$*.*

4.3.1 3D Object Model

Active shape models (ASM) are a well-known representation of deformable geometric models. In the literature, they are used in the context of reconstruction as well as object detection and tracking in images or in three-dimensional object space, cf. Section 3.3. In this work, an ASM is used to encode prior knowledge and to further specify the objects $\mathbf{o}_k$, which have been introduced as a key component of the scene flow approach in Section 4.2.1. In order to restrict the high-dimensional space of possible object shapes, we follow [Zia et al., 2013b] and use their three-dimensional active shape model. A training set of 38 manually annotated CAD models of passenger cars forms the basis for this geometric representation. Thanks to the corresponding annotations of key points on the training models, they can be aligned and rescaled so that mean vertex positions can be computed. Principal component analysis is applied to the set of key points to retrieve the directions of the most dominant deformations between the samples in the training set.

Based on the resulting active shape model, novel object instances can be generated within the range of deformations in the training set. Shape parameters $\boldsymbol{\gamma}$ can be chosen to compute deformed vertex positions $\mathbf{V}$ from a linear sub-space model

$$\mathbf{V}(\boldsymbol{\gamma}_k) = \mathbf{M} + \sum_{i=\{1,2\}} \boldsymbol{\gamma}_k^{(i)} \sigma_i \, \mathbf{e}_i \tag{4.22}$$

where $\mathbf{M}$ is the vertex mean and $\mathbf{e}_i$ denotes the i'th eigenvector from the PCA, weighted by the standard deviation σ_i of the corresponding eigenvalue. As indicated in the sum, the ASM is limited to the first two principal components as a tradeoff between model complexity and the quality of the approximation. Figure 4.7 depicts the mean shape in the center and deformed versions of the model on the left and right, illustrating the range of different layouts covered by the first two principal components. While the first component accounts mostly for the size of the object, the second determines its general shape. As shown in the figure, a broad range of passenger cars are covered by this simple model.

To incorporate shape and pose parameters into the scene flow approach, the basic object model $\mathbf{o}$, introduced in Section 4.2.1, is extended by two additional vectors. The shape parameters $\boldsymbol{\gamma}$ control the influence of the individual deformations of the ASM. Vector $\boldsymbol{\xi}$ comprises the pose parameters of the extended object model. The two-dimensional position on the ground plane and a heading angle provide a compact representation of the position and orientation of the model relative to the reference camera. Altogether, the extended representation of objects $(\boldsymbol{\gamma}_k, \boldsymbol{\xi}_k, \mathbf{R}_k, \mathbf{t}_k)$ comprises a total of 11 parameters. Since the number of objects composing a scene is assumed to be small, this extension increases the complexity of the overall model only slightly.

In the extended conditional random field model, the shape parameters $\boldsymbol{\gamma}_k$ of object $\mathbf{o}_k$ are optimized for consistency with the jointly estimated superpixel parameters. We define a triangular mesh connecting the vertices $\mathbf{V}(\boldsymbol{\gamma}_k)$ as depicted in Figure 4.7. It allows to generate virtual disparity maps by rendering a depth buffer of the meshed model using standard computer graphics. Based on the orientation parameters of the stereo cameras the resulting depth map can be converted to an equivalent representation in disparity space.

4.3.2 Extension of the Scene Flow Model

The flexible design of random field models allows for a straight-forward extension of the scene flow energy function 4.8. To constrain the additional parameters of the revised object model an additional shape and pose consistency term is incorporated. We re-define the scene flow problem as a conditional random field in terms of the energy function

$$E(\mathbf{s}, \mathbf{o}) = \sum_{i \in \mathcal{S}} \Big(\underbrace{\varphi_i(\mathbf{s}_i, \mathbf{o})}_{\text{data}} + \underbrace{\kappa_i(\mathbf{s}_i, \mathbf{o})}_{\text{shape\&pose}} \Big) + \sum_{(i,j) \in \mathcal{N}} \underbrace{\psi_{ij}(\mathbf{s}_i, \mathbf{s}_j)}_{\text{smoothness}} \tag{4.23}$$

Here, $\mathbf{o}$ is the extended object representation introduced in the previous section. $\mathbf{s}$ represents the same planar superpixels as before, and $\mathcal{N}$ denotes the set of adjacent superpixels. Data and smoothness terms $\varphi(\cdot)$, $\psi(\cdot)$ remain the same as in the basic scene flow model. The novel shape and pose consistency term $\kappa(\cdot)$ is explained in the following section. To avoid computationally demanding higher-order potentials, we do not model interdependencies between objects explicitly. The Potts model between adjacent image segments discourages overlapping objects.

4.3.3 Shape and Pose Consistency Term

The novel shape and pose consistency term encourages object models to agree with the planes of the associated superpixels. In accordance with the employed parametrization of scene flow, the evaluation of a consistency measure will be conducted in disparity space. Like the original data

Figure 4.8: *Visualization of the shape and pose consistency term. The reference image is superimposed with a color-coded representation of the induced penalty. Yellow and red colors indicate high penalties while blue colors encode low costs. The left panel shows the example after initialization, the right panel shows the penalty after optimization.*

term in Equation 4.9, the additional term κ decomposes into computationally tractable pairwise potentials between superpixels and the assigned objects:

$$\kappa_i(\mathbf{s}_i, \mathbf{o}) = \sum_{k \in \mathcal{O}} \left([l_i = k] \cdot S_i(\mathbf{n}_i, \mathbf{o}_k) + [l_i \neq k \wedge k > 1] \cdot O_{ik}(\mathbf{o}_k) \right) \tag{4.24}$$

Here, κ_i defines the cost function for a distinct image segment $\mathbf{s}_i$. It is composed of the shape consistency term S and an occlusion penalty O which will be defined in the following.

$S_i(\mathbf{n}_i, \mathbf{o}_k)$ enforces consistency between the shape of object $\mathbf{o}_k$ and the assigned planes described by $\mathbf{n}_i$. In analogy with the data term, shape consistency is evaluated with respect to the object associated with the superpixel via l_i. The Iverson bracket activates the respective combinations. The penalty function S_i considers two cases

$$S_i(\mathbf{n}, \mathbf{o}) = \begin{cases} C^{\text{bg}} & \text{if } \mathbf{o} \text{ is background} \\ C_i^{\text{obj}}(\mathbf{n}, \mathbf{o}) & \text{otherwise} \end{cases} \tag{4.25}$$

C^{bg} denotes a constant penalty for superpixels associated with the background object. It is imposed to avoid a bias towards purely static scenes. $C_i^{\text{obj}}(\mathbf{n}, \mathbf{o})$ compares two disparity maps to evaluate the shape consistency of foreground objects. To this end, the meshed 3D model of object $\mathbf{o}_k$ is projected to a disparity map according to $\boldsymbol{\xi}_k$ and $\boldsymbol{\gamma}_k$. The penalty C_i^{obj} is computed as the sum of the truncated absolute differences between the virtual disparity map and the disparities induced by the plane $\mathbf{n}_i$. Differences are computed for all pixels inside $\mathcal{R}_i$, which coincide with the projection of $\mathbf{o}_k$. Pixels remaining uncovered by the projected model are penalized with a multiple of C^{bg}. This encourages the projected model to approximately align with superpixel boundaries. Note that in contrast to the data term D_i, which encourages consistency between estimated 3D plane parameters and image observations, this term evaluates the consistency between the deformed shape model and the reconstructed superpixels.

The second part of Equation 4.24 is the occlusion penalty O_{ik}. It penalizes a possible overlap between parts of a foreground model and superpixels that are assigned to a different object via the arguments of the leading Iverson bracket. The overlap penalty itself is chosen to be proportional to the overlap of the projected model of object $\mathbf{o}_k$ with the background superpixel $\mathbf{s}_i$. This term is crucial to prevent object models from exceeding the true object boundaries. Figure 4.8 provides a visualization of the shape and pose consistency term. The reference image is superimposed with a color-coded representation of the induced penalty. Yellow and red colors indicate high penalties while blue colors encode low costs. The left panel shows the example after initialization; the right panel shows the penalty after optimization.

4.3.4 Initialization

As before, the extended scene flow model requires meaningful initial values. The motion-based segmentation from Section 4.2.4 is adapted to the more complex parametrization. An appropriate option concerning the additional shape parameters is offered by the trained active shape model. Each object hypothesis is initialized with the mean shape of the model by setting its shape parameters $\boldsymbol{\gamma}$ to zero.

A slightly more sophisticated initialization is required for the pose parameters, which reside in a much larger solution space. Consequently, meaningful optimization results depend on good approximate values. Section 4.2.4 describes a motion-based segmentation, which is employed to generate object hypotheses for the basic scene flow approach. The mean positions, reduced to the ground plane, and the moving directions of the best hypotheses are used as initial values for the object pose parameters $\boldsymbol{\xi}$. As before, this initialization procedure will lead to some spurious object hypotheses. During inference, such false positives are pruned if no superpixels are assigned to them.

4.3.5 Approximate Inference

Like the employed graphical model, the optimization strategy proposed for the basic scene flow approach is flexible enough to include the additional shape and pose parameters. The outline of the optimization procedure is similar to Algorithm 1 in Section 4.2.5. To approximately minimize the energy function, we iteratively and adaptively discretize the domains of the continuous variables in the outer loop of the particle-based inference framework.

The dimensionality of the shape parameters is efficiently reduced to the two most important components of the ASM. However, sampling the wide range of object instances covered by these parameters requires a large number of particles. They are additionally augmented with the sought pose parameters, which are also covered by particle sampling. Consequently, the naïve combination

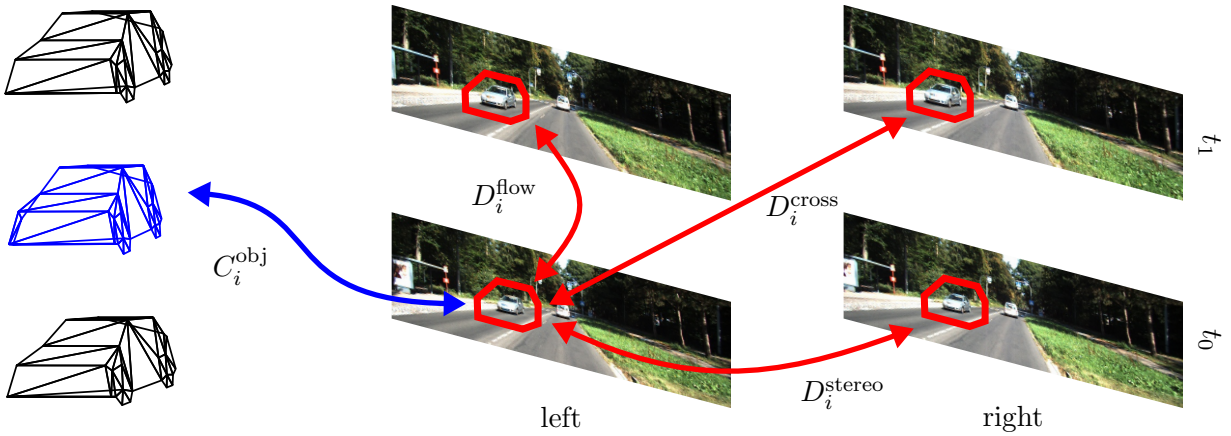


Figure 4.9: *Visualization of the optimization pipeline. For each object the best shape and pose proposal (depicted in blue) is selected before its motion parameters are jointly optimized with the superpixel parameters (red arrows).*

of all object particles, i.e. motion, shape and pose, as suggested by the original algorithm, would lead to a significant increase in the computational complexity of the problem. Note that this is not caused by the moderate number of additional parameters, but by the large number of particles needed to explore the associated search space.

To keep the computational burden tractable, we perform informed sampling of pose and shape parameters based on the respective data term. In each iteration of the outer loop of Algorithm 1, we draw 50 particles, which jointly sample pose and shape parameters from normal distributions centered at the preceding MAP solution. As before, the respective standard deviations are iteratively reduced. To prune the proposals, the shape and pose consistency term (4.24) is evaluated for each particle with respect to the disparity map induced by the current MAP solution of the superpixels. Since no interdependencies between the object models were introduced, this pruning step only requires the evaluation of the shape and pose data term. Thus, it efficiently pre-selects the particle that induces the lowest energy. Only the best particle of each object is accepted and introduced into the optimization of Equation 4.23 by TRW-S. This pruning step is visualized for one exemplary object in the left panel of Figure 4.9. During the optimization of the remaining parameters, the shape and pose consistency term remains active. Here, it is evaluated with respect to the injected best object particle to guide the optimization of the remaining variables.

4.4 Discussion

Based on the preceding detailed descriptions, in this section we discuss the model assumptions and derive the advantages and limitations they impose on the individual methods.

DiscreteFlow

In contrast to classical variational methods, the proposed optical flow approach is designed to directly infer a good combination of large displacement vectors subject to regularization by a random field prior. As naïve discretization of the solution space is not feasible given realistic ranges of displacement, promising flow proposals are computed based on a feature descriptor. Thus, the achievable performance critically depends on the quality of the proposal set. It is possible that the approximate nearest neighbor search is affected by adverse imaging conditions or strong scale changes, for example. Section 5.2 describes an oracle experiment, which serves to evaluate the completeness of the proposed particle set.

The key advantage of DiscreteFlow with respect to existing approaches is its ability to effectively produce a dense grid of, potentially very large, two-dimensional displacements. The building blocks allowing for efficient approximate discrete inference are twofold. The diverse proposal set and an appropriate strategy for the computation of pairwise penalties collaborate seamlessly to explore a very large search space with reasonable computational complexity. The result of the presented discrete inference approach is an integer-valued flow field with undefined regions where inconsistent matches have been removed. This consolidated, semi-dense displacement field serves as input to the scene flow algorithms, which constitute the further contributions of the present thesis.

To retrieve an accurate, dense displacement field, which can be compared to the results of related optical flow methods, undefined regions need to be interpolated and the integer-valued flow field has to be refined to sub-pixel accuracy. Following related work, the respective post-processing steps of EpicFlow are applied as part of the proposed processing pipeline. In the original publication the flow field is initialized using DeepMatching. As described in Section 3.1, this sparse image matching approach addresses wide-baseline non-rigid matching. In contrast, DiscreteFlow relies on the invariance of the DAISY descriptor against moderate changes of view-point and focusses on establishing a dense gird of correspondences. Expressive optical flow priors are imposed on neighboring displacement vectors to tailor the presented matching procedure towards dense motion field estimation.

It will be informative to extend the proposed method to purely discrete inference. One framework that would allow for an iterative refinement of the optical flow field is the particle-based inference scheme discussed in Section 4.2.5. The related method of [Mozerov, 2013] applies a similar refinement step. As the foundation for the other contributions is successfully built, and a meaningful evaluation of the optical flow results can be performed based on the described approach, this extension is left to future work.

Object Scene Flow

The proposed Object Scene Flow approach is based on two foundations. One is provided by the novel motion observations discussed above and the other is the object-based formulation of the problem.

As pointed out in the beginning of Section 4.2, the key assumption of the presented model is that the observed scene decomposes into a finite set of rigidly moving objects. This assumption will not hold for articulated objects depicted at large scale, where rigid transformation parameters are not valid within larger image regions. However, it will start to be beneficial in scenarios depicting articulated objects at moderate scale. In such cases, it will help to determine the dominant motion, even for articulated parts of the scene, like bicyclists and pedestrians for example. Consequently, major parts of the observed scene will be modeled correctly and highly relevant information can be retrieved. In a number of relevant applications, like autonomous navigation in structured, man-made environments or highway scenarios in automotive vision tasks, the assumption is valid for the entire scene. The qualitative evaluation in Section 5.3.4 shows that the proposed method is able to provide promising results in such situations.

Sparse and dense observations have been combined in the data term to exploit the advantages of both approaches. While sparse features provide a meaningful initialization and reliable observations in appropriately textured image regions, a dense similarity measure is important to guide the refinement of the slanted plane parameters. Adverse effects of the employed oversegmentation of the reference frame are compensated by dedicated smoothness terms tailored to scene flow estimation and allowing for discontinuities where indicated by image evidence. Related work advocates the estimation of a consistent segmentation with respect to all four input images [Vogel et al., 2015]. While this approach requires solving the association problem for all views, and thus increases the complexity of the optimization problem, it allows for a consistent handling of all image observations and occlusions. A view-consistent formulation could improve the performance of Object Scene Flow and would provide the basis for an extension to longer image sequences.

The non-convex objective function specified in Equation 4.8 suggests that the result of the scene flow approach will depend on the quality of the initialization. This challenge is tackled with a combination of a state-of-the-art stereo matcher and optical flow observations from DiscreteFlow, providing accurate approximate values. To keep the initialization of objects as general as possible it is addressed by a simple motion-based segmentation of scene flow matches. For the sake of generality, we renounce on a class-specific object detector, which would be helpful to prune hypotheses that do not correspond to the objects of interest. The presented approach copes with erroneous object hypotheses by either not assigning any superpixels to them or adjusting their motion pa-

rameters to values close to the background motion. As we are interested in motion estimation, it deliberately does not detect objets, which are currently static like waiting pedestrians.

Due to the complex structure of the energy function, inference cannot guarantee to retrieve its global optimum. However, the proposed particle-based optimization scheme allows to jointly infer the segmentation of the scene into objects and refine the continuous variables representing geometry and motion.

Joint Estimation of Vehicles and Scene Flow

Obviously, the introduction of class-specific prior knowledge limits the generality of the scene flow approach. As stated in Section 4.3.1 the employed ASM covers a broad range of passenger cars, which account for a significant part of motorized traffic, but excludes other types of vehicles and pedestrians. On the other hand, this extension provides a pragmatic solution that allows for the application of the presented approach while models that are more general are developed. The detailed object knowledge provides strong guidance during inference. In fact, the reconstruction in terms of a parametrized object model is the goal of perception in some applications.

To fully exploit the potential of class-specific object knowledge it seems promising to incorporate dedicated object detectors in future work. This extension would help to prune false positive object hypotheses before starting the optimization and, at the same time, allow to discriminate different object classes. As a consequence, it would be possible to inject models of different object types and to adapt the model to non-rigid motions where necessary, e.g. for observed pedestrians.

By design, the proposed model is limited in handling occlusions between object models. While related work on holistic scene understanding [Geiger and Wang, 2015] has to deal with significant mutual occlusions between a number of inferred objects, the strategy to address this issue typically involves complex higher-order potentials in the graphical model, which are beyond the scope of this thesis.

An extension of the approach to longer image sequences, as described in [Leibe et al., 2006] and [Vogel et al., 2015], would be promising as well. On the one hand, it allows to impose temporal constraints on the trajectories of the camera and observed objects. On the other, a temporally consistent reconstruction of the environment would provide strong prior information for object detection. Spurious hypotheses next to the road could also be pruned based on such geometric constraints.

Chapter 5

Experiments and Results

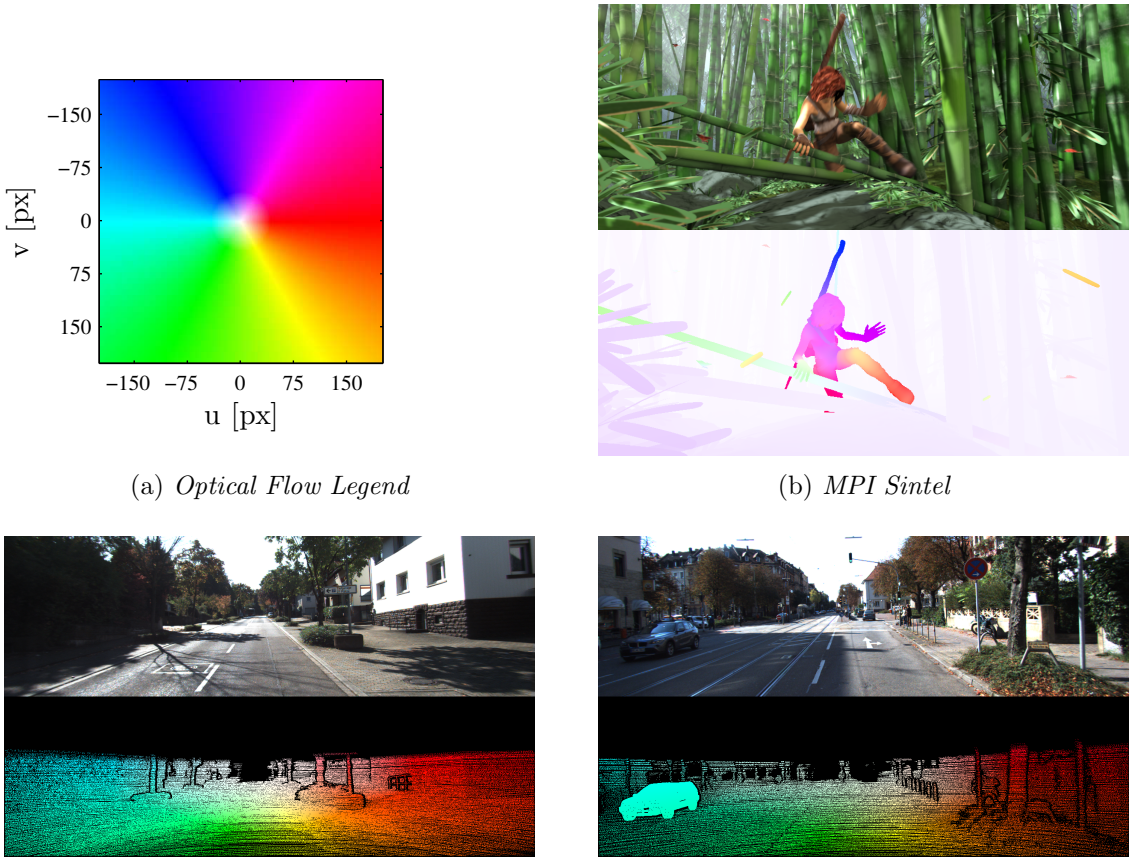
In the preceding chapter, three related algorithms for motion estimation have been proposed. The goal of this chapter is to investigate the validity of underlying assumptions and the performance of the novel approaches. It explains the design and provides the results of the corresponding experimental evaluation. Section 5.1 introduces the data sets used for the investigations. In the following, the developed methods are investigated in turn. Section 5.2 addresses optical flow estimation, the basic scene flow approach is evaluated in Section 5.3 and Section 5.4 deals with its model-based extension. A thorough discussion of the presented results and their implications will be provided in Chapter 6.

5.1 Data

The empirical evaluation of the proposed methods is conducted on three different public data sets. Figure 5.1 provides exemplary data from all employed benchmarks showing that the contained imagery and reference data differ significantly. To point out their usefulness the characteristics of the individual data sets will be discussed in the following.

[Butler et al., 2012] proposed *MPI Sintel*[1], a very challenging optical flow benchmark based on synthetic data from an independently produced short movie. Although realism of the imagery is limited due to the completely virtual production, some of the common influences of imaging devices and outdoor scenes are added to the image data. More importantly, the synthetic production provides accurate dense reference data for very complex scenes and motions. A comprehensive set of training images is published together with ground truth optical flow maps for validation purposes. The ground truth for the test data is held back to allow for a fair comparison to other methods.

[1] http://sintel.is.tue.mpg.de/



(a) *Optical Flow Legend*

(b) *MPI Sintel*

(c) *KITTI 2012*

(d) *KITTI 2015*

Figure 5.1: *Examples from the employed benchmark data sets. Panel (a) shows the optical flow legend used throughout this thesis. The top row of (b), (c) and (d) shows the reference input image, the second row provides reference optical flow maps. For MPI Sintel, the flow values are multiplied by a factor of 5 to enhance the contrast.*

The training and test data of MPI Sintel consist of two different versions of the same images. The *clean* portion of the data set contains shaded renderings of the virtual scenes. Surface properties and the direction of illumination are considered to produce realistic appearance. The imagery of the *final* portion is compromised with simulations of imaging artifacts, such as motion and defocus blur. Figure 5.2 compares an exemplary frame from both parts of the data set. The input images are provided in the left column, the result of DiscreteFlow is depicted in the second column with error maps on the right. The endpoint error at each pixel is normalized and shown as an intensity value. In this example, an overexposure effect in the sky region leads to increased errors in the top left area of the final version, depicted in the second row. Note however, that the endpoint error decreases on the birds in this region and the motion of these small salient regions is erroneously propagated into the untextured surrounding. This effect is less pronounced in the clean version where the erroneous motion is bounded by the birds and the background motion is correctly retrieved in the surrounding.

Figure 5.2: *Different passes of MPI Sintel. The first column contains the reference image. The clean pass includes smooth shading and specular reflections; the final pass contains full rendering with all effects including blur due to camera depth of field and motion. Color-coded results of DiscreteFlow are provided in the center. The endpoint error at each pixel is encoded as an intensity value in the right column.*

For real-world imagery, the generation of reference data for motion fields is more challenging. One reason for this is that no sensor exists which is capable of capturing optical flow ground truth in complex environments. Hitherto, this lead to a shortage of appropriate reference data, especially for scene flow evaluation. As an alternative, synthetic renderings of spheres [Huguet and Devernay, 2007, Valgaerts et al., 2010], as shown in Figure 1.1, other geometric primitives [Vogel et al., 2011, Cech et al., 2011, Basha et al., 2013] or simple street scenes [Wedel et al., 2008, Rabe et al., 2010] were typically employed to measure the quantitative performance of respective algorithms. More realistic image sequences form the basis of the KITTI benchmark suite [Geiger et al., 2012]. It provides a number of diverse challenges with a focus on automotive applications. The provided stereoscopic image sequences were captured from a car driving in regular traffic on public roads. Three-dimensional reference data is captured by a 360° laser scanner mounted on top of the car. A similar approach is followed by [Kondermann et al., 2015] who register stereo imagery with existing scans of an urban environment and add error bars to the resulting reference values. Although the images are much more realistic, compared to synthetic renderings, both data sets provide reference data for static scenes only. In this thesis, several experiments are conducted on the stereo and optical flow data from the original KITTI benchmark, which is referred to as *KITTI'12*, according to the date of publication. Since stereoscopic image sequences are available, KITTI'12 can be processed with scene flow approaches but the reference data does not cover full 3D displacements.

The KITTI benchmark suite was complemented with a scene flow extension within the scope of the present thesis [Menze and Geiger, 2015]. This extension will be referred to as *KITTI'15*. It comprises annotations of individually moving objects, making it the first realistic scene flow data set. Starting from the raw data, which was collected for the KITTI project, the annotation process consists of two major steps. First, the static background of the scene is recovered from laser range measurements by removing all dynamic objects and compensating for the vehicle's ego-motion.

Second, dynamic objects are re-inserted by fitting detailed CAD models to the individually moving parts of the point clouds in each frame. Although the annotation process allows for a higher density of reference values on the individually moving foreground objects, about 85% of the annotations lie on the static background. The data set and an online evaluation on test data with held back reference are available as part of the KITTI benchmark suite[2]. In addition to the evaluation of scene flow estimates, it allows for the individual evaluation of results for the stereo matching and optical flow sub-problems.

Several experiments in this section require the availability of reference data, which is published for the training sets of the employed benchmarks. KITTI'12 and KITTI'15 provide moderate numbers of 194 and 200 training samples, respectively. To enable unbiased comparison and to decrease the computational burden during training, the annotated training data is split into two sets of equal size. One half is used for parameter training and the other serves validation purposes. MPI Sintel comprises 1064 partly redundant training images. Here, 20% of the data are used for training, yielding a total of 213 images. The remaining images provide the basis for validation. The characteristics of the available reference data are provided in Table 5.1. It shows maximum displacement vectors, which are comparable between the instances of KITTI and significantly larger for Sintel. All three data sets require the ability to estimate large displacements and thus lend themselves to the evaluation of the proposed methodology. More detailed statistics are provided in the referenced publications.

	u [px]		v [px]		d [px]	
	min	max	min	max	min	max
MPI Sintel	-360	370	-233	313	-	-
KITTI'12	-179	226	-49	73	4	228
KITTI'15	-191	242	-52	86	5	230

Table 5.1: *Maximum displacements contained in the benchmark data. The numbers in this table refer to non-occluded pixels.*

5.2 Optical Flow by Discrete Optimization

This section starts with a description of the training procedure, which is employed to find appropriate parameter values (5.2.1). Next, the proposed optical flow approach is evaluated on all three data sets described in the previous section. The investigations start with three pilot studies (5.2.2), which experimentally validate the assumptions made in the description of the model in Section 4.1. A quantitative comparison of DiscreteFlow to the state-of-the-art follows in Section 5.2.3. Finally, a number of qualitative results are presented in Section 5.2.4 giving a visual impression of the advantages and shortcomings of the approach.

[2]KITTI 2015: http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php

Parameter	Symbol	Sintel	KITTI
Relative weight of unary and pairwise terms	λ	0.050	0.232
Truncation threshold of the data term	τ_ϕ	2.5	2.5
Truncation threshold of the smoothness term	τ_ψ	15	15
Size of the label set	L	500	500
Number of neighbors to be sampled	N	200	200
Stride for discrete CRF in pixels		4	4
Standard deviation for sampling neighbors		5	5
Similarity threshold (fw-/bw-check)		10	10
Minimum segment size (outlier removal)		100	100
Consistency threshold (outlier removal)		10	10

Table 5.2: *Parameters of DiscreteFlow trained on MPI Sintel and KITTI'15.*

The general formulation presented in Section 4.1 would allow for a dense pixel-wise computation of labels. Because a final interpolation and continuous refinement step is conducted, discrete optimization is performed on a subset of pixels in the reference frame. This further reduces computational complexity and meets the demands of the variational post-processing, which is reported to work best on semi-dense input matches [Bailer et al., 2015]. The parameter controlling the distance between processed image sites is referred to as *stride*. As the trainig of the remaining parameters depends on the choice of the stride value, it is fixed to 4 for all experiments.

5.2.1 Parameter Training and Sensitivity

Training the parameters of the optical flow random field model (4.1) analytically is a complex task, as it requires the computation of the gradient of the objective function. The objective function for parameter learning is not to be confused with the energy (4.1) but consists of an error metric with respect to available reference data. In general, a gradient measure can be expected to be noisy and related methods are susceptible to local minima in the vicinity of the initial values. To circumvent these problems and to find parameter values yielding a strong local minimum of the error measure, derivative-free direct search is applied as described in [Hooke and Jeeves, 1961]. This approach iteratively samples the range of plausible parameter values and directly compares the results.

To keep the complexity tractable, individual parameters are investigated in turn. Each parameter is initialized with a plausible initial value and search range. To approximate a local minimum of the estimation error for each of the variables, the search ranges of the parameters are sampled at a fixed number of points. The estimation error is evaluated based on the available reference data. After each parameter is investigated, the search range is reduced to 60% of its original size and the procedure is repeated. Table 5.2 reports the values learned using direct search and used for the experimental evaluation in the following sections.

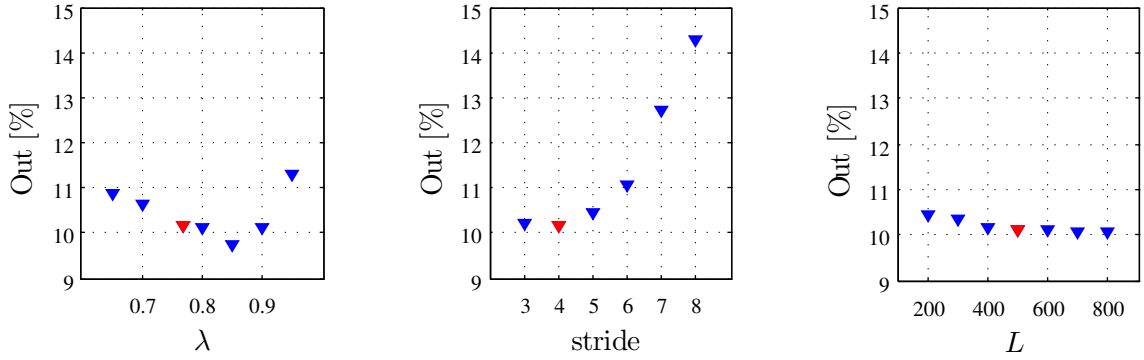


Figure 5.3: *Sensitivity analysis of DiscreteFlow parameters on the validation data of KITTI'15. λ is the relative weight of data and smoothness term, stride is the gap between neighboring image sites in pixels. L is the cardinality of the label set; the ratio of randomly sampled proposals from neighboring reference pixels is fixed to $0.4\,L$. Parameter values learned on the training data are marked in red.*

The first experiment serves to validate the choice of parameters described above. To this end, the most important model parameters are varied around the best value found in training and the resulting error rates are compared. To allow for an unbiased comparison, the evaluation is conducted on the non-occluded parts of the validation portion of KITTI'15. The results are visualized in Figure 5.3; they are consistent with the numerical values reported in panel (c) of Table 5.3. The parameter values learned on the training set are marked in red. In the random field model, smoothness and data term are balanced by the relative weight λ. The left panel of Figure 5.3 shows a non-convex relationship between the outlier percentage and the chosen parameter value. On the training data, we find a different but close-by minimum compared to the depicted validation data. For the stride length, the chosen parameter yields the minimum error on the validation data. The results show the expectable degradation with growing stride. Concerning the cardinality of the label set L, we chose a trade-off between accuracy and computational complexity. The results do not improve significantly for proposal sets larger than 500.

A second insight gained from this experiment concerns the sensitivity of the model to the choice of its parameters. While the results for λ and L differ in the range of two percentage points over the investigated parameter range, significant degradation of up to four percentage points can be observed for the stride length. Note that in this experiment only one parameter is varied while the remaining values remain constant. Especially in the case of the stride parameter a decrease in performance is to be expected due to non-adjusted parameter combinations.

5.2.2 Pilot Studies

The first pilot study investigates the quality of the proposal set computed by DiscreteFlow. To initialize the required discrete labels for each node of the random field, and to allow for efficient

	EPE [px]	Out [%]
DM+DeepFlow [Weinzaepfel et al., 2013]	3.09	9.97
DM+EpicFlow [Revaud et al., 2015]	**2.53**	8.55
Proposed without refinement	4.00	10.82
Proposed+DeepFlow	2.95	8.99
Proposed+EpicFlow	2.65	**7.84**
Proposed+Oracle	0.88	3.88

(a) **MPI Sintel**

	EPE [px]	Out [%]
DM+DeepFlow [Weinzaepfel et al., 2013]	1.49	7.28
DM+EpicFlow [Revaud et al., 2015]	1.47	7.45
Proposed without refinement	2.69	11.12
Proposed+DeepFlow	1.23	5.74
Proposed+EpicFlow	**1.19**	**5.56**
Proposed+Oracle	0.58	1.03

(b) **KITTI 2012**

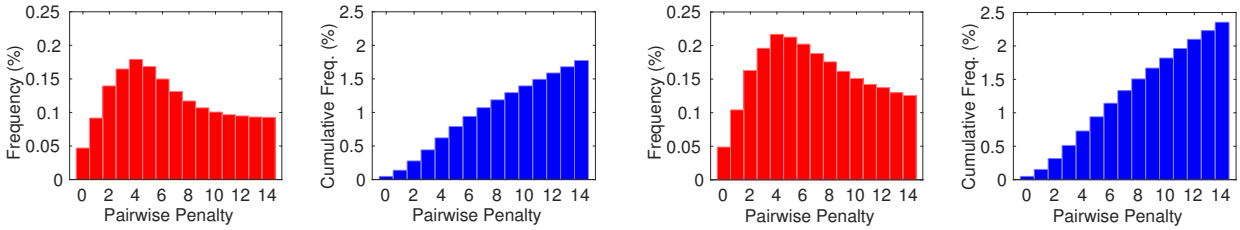
	EPE [px]	Out [%]
DM+DeepFlow [Weinzaepfel et al., 2013]	5.18	18.01
DM+EpicFlow [Revaud et al., 2015]	4.11	16.81
Proposed without refinement	5.07	17.94
Proposed+DeepFlow	3.04	11.84
Proposed+EpicFlow	**2.52**	**10.18**
Proposed+Oracle	0.78	2.07

(c) **KITTI 2015**

Table 5.3: *Pilot study on the validation portion of MPI Sintel and KITTI training sets.*

inference in the optical flow model, a diverse set of flow proposals is computed. The quality of this proposal set determines the upper bound of the performance achievable with the discrete optimization approach. As discrete inference can only assign optical flow vectors contained in the label set, its completeness is crucial in the sense that it should contain a high-quality proposal at each image location.

Table 5.3 shows numeric results on the validation portions of MPI Sintel and the KITTI training sets. From top-to-bottom, the tables provide the results of DeepFlow [Weinzaepfel et al., 2013], EpicFlow [Revaud et al., 2015], the proposed method in combination with the refinement stages of DeepFlow and EpicFlow, as well as an *Oracle* result. The latter is the core of this experiment. It refers to the optical flow map obtained by selecting the flow vector from the proposal set, which exhibits the smallest endpoint error compared to the available ground truth. As described in Section 4.1.2, the presented approach uses only the refinement stages of DeepFlow and EpicFlow. It replaces the DeepMatches (DM), which form the foundation for the original approaches. In contrast to DeepMatching, the proposed method is able to compute an integer-valued, dense



(a) **MPI Sintel** (b) **KITTI**

Figure 5.4: *This figure shows the frequency of neighboring flow proposal vectors with respect to their endpoint distance on MPI Sintel (a) and KITTI (b) in red. The blue plots depict the cumulative frequency corresponding to* $|\mathcal{K}_{\mathbf{p},\mathbf{q},l}|/|L|$ *over the distance threshold* τ_ψ*, respectively.*

displacement field. To show the influence of the refinement step, the evaluation of this initial flow field is provided in the third row of each table. Simple nearest neighbor interpolation is applied to fill in regions that did not pass the consistency check.

Error rates are provided in terms of the average endpoint error (EPE) and the outlier percentage with respect to a threshold of 3 pixels. This corresponds to the standard evaluation metrics of most optical flow benchmarks. As the focus of this experiment is on correspondences, which can correctly be established, evaluation is conducted on the non-occluded parts of the reference data. Pixels becoming occluded or leaving the target image are excluded. To allow for a straightforward evaluation and comparison, the semi-dense oracle result is dilated to cover all pixels using nearest neighbor interpolation.

As evidenced by this experiment, the proposed method obtains state-of-the-art performance on the validation sets, outperforming the baseline approaches in most of the error metrics. While discrete inference provides reasonable results on its own, the refinement step significantly reduces the error on all data sets. As can be expected, the endpoint error decreases much more than the outlier percentage. This reflects the high quality of the initial flow field, which yields promising outlier percentage, and the importance of upgrading the results to sub-pixel accuracy.

The quality of the proposal set is evaluated by the oracle experiment. Choosing the proposal closest to the reference data reduces the outlier percentage significantly by 50% compared to the best result reported in Table 5.3 (a). The effect is even larger considering the average endpoint error, which is reduced by 65% of the best result computed on MPI Sintel. Table 5.3 (b) and (c) confirm this trend on the evaluation portion of both KITTI data sets, respectively. These results indicate that our proposal set contains flow vectors close to the ground truth for most of the pixels while being diverse enough to avoid local minima and reduce the computational complexity of the pairwise terms as will be seen from the next experiment.

The second pilot study addresses a different property of the proposals. As explained in Section 4.1.2, the characteristics of proposal sets at neighboring image sites influence the computa-

	EPE [px]	Out [%]
DeepMatches [Weinzaepfel et al., 2013]	2.12	9.24
DiscreteFlow stride 8	1.35	5.13
DiscreteFlow stride 4	**1.12**	**4.78**

Table 5.4: *Comparison of DeepMatches and the discrete flow field as computed by the proposed approach on the validation set of MPI Sintel. In all results, outliers have been removed by a consistency check.*

tional complexity of the pairwise potential in the optical flow model. Figure 5.4 shows histograms of the frequency of neighboring flow proposals with respect to the distance of their endpoints for label sets of size $L = 500$. While the pairwise term in Equation 4.3 can take $L^2 = 250,000$ states in total, less than 2.5% of them fall below the chosen truncation value of 15 pixels and need to be evaluated in Equation 4.6. Thus, pre-calculating the set of relevant neighboring proposals $\mathcal{K}_{\mathbf{p},\mathbf{q},l}$ potentially saves almost two orders of magnitude in computation time assuming sequential processing.

Compared to EpicFlow, the contributions of the optical flow approach concern the set of input matches. In the presented approach, they are the result of discrete inference and outlier removal, and they replace DeepMatches, which are used in the original publication. The performance of the matching approaches is compared in Table 5.4. Following standard conventions, it is evaluated in terms of average endpoint error and outlier percentage. On the same validation subset of Sintel training images as in the previous pilot studies the error metrics are computed as the sum of outliers and endpoint errors normalized by the respective total number of matches. In all results, outliers have been removed. As DeepMatches are defined on a semi-dense grid of stride 8, the results of the novel approach are provided with the same stride length and with stride 4 as used in the evaluation of overall performance. The error rates decrease significantly when compared using the same stride. For the final stride length of 4, both, the average endpoint error as well as the outlier percentage decrease by more than 45% of the original error rate.

As DiscreteFlow conceptually solves the computationally more demanding dense image matching problem, its runtime cannot compete with that of DeepMatching, which takes 15 seconds to establish sparse correspondences in a typical image pair from MPI Sintel. However, the three strategies described in Section 4.1 allow to apply discrete inference to two-dimensional image matching of realistic imagery in reasonable 180 seconds. About 15% of the runtime is spent for descriptor extraction, 45% for proposal generation, and 40% for discrete optimization.

5.2.3 Comparison to Related Methods

To evaluate the performance of the complete proposed processing pipeline on diverse challenging benchmarks, the results are submitted to the three public evaluation portals described in Sec-

tion 5.1. Reference data for the published test images is held back, allowing for a fair comparison with state-of-the-art methods. For the same reason, we can only discuss quantitative results as provided by the evaluation on the respective websites. For consistency, the format of the numerical values follows the respective online evaluation portals. However, not all post decimal positions reveal significant differences in performance.

All tables reporting results on test data contain duplicate entries of the proposed methods. One is marked with the respective citation. It provides the results published in the original articles. The other entry, in italic typeface, shows the results of the final version of the algorithm as described in this thesis. Differences will be discussed in the corresponding sections of this chapter. As different ranking criteria apply to the different benchmarks, the decisive error metric is marked in bold typeface in the column headings.

Table 5.5 comprises the results on the MPI Sintel optical flow benchmark. This evaluation contains challenging scenes from an animated movie with more diverse and complex motions compared to KITTI. The evaluation protocol requires submitting results for both versions of the data set. In both categories, the presented approach performs slightly better than EpicFlow in terms of the ranking criterion. The average endpoint error is reduced by 8% on the *final* portion and by 16% on the *clean* data. In contrast to the preliminary version of DiscreteFlow published in [Menze et al., 2015a], the final approach does not contain the edge-sensitive weighting of the smoothness term defined in Equation 4.4. It was able to reduce the error on Sintel training data slightly but this finding was not confirmed on the KITTI benchmarks. The final implementation without this weight improved the results on Sintel slightly.

Deep+R is the result of DeepFlow with refined initial matches [Drayer and Brox, 2015]. The refinement consists of the optimization of a MRF over the initial matches and thus bears similarities with DiscreteFlow. The EPE is reduced significantly with respect to the original approach but neither reaches the performance of EpicFlow nor DiscreteFlow. The best results on the final portion are achieved by Flow Fields, which initializes EpicFlow with a dedicated approximate nearest neighbor field.

The overall results on the KITTI 2012 optical flow data set are provided in Table 5.6. This table contains only general optical flow methods, as additional constraints, such as depth information from stereo matching, cannot be imposed in general. Methods using additional information are contained in Table 5.11b where they are compared to the result of Object Scene Flow. Here, the results are ranked according to the outlier percentage of non-occluded pixels with respect to a threshold of 3 px. Compared to EpicFlow, this metric is significantly reduced by 27%. Notably, the purely variational approaches of [Ranftl et al., 2014] and [Demetz et al., 2014] perform very well on this data set. The latter achieves the smallest average endpoint error on all annotated image locations.

	EPE [px]		
	All	Noc	Occ
DiscreteFlow	**3.438**	**0.989**	**23.428**
DiscreteFlow [Menze et al., 2015a]	3.567	1.108	23.626
Flow Fields [Bailer et al., 2015]	3.748	1.056	25.700
DM+EpicFlow [Revaud et al., 2015]	4.115	1.360	26.595
PH-Flow [Yang and Li, 2015]	4.388	1.714	26.202
AggregFlow [Fortun et al., 2014]	4.754	1.694	29.685
TF+OFM [Kennedy and Taylor, 2014]	4.917	1.874	29.735
Deep+R [Drayer and Brox, 2015]	5.041	1.481	34.047
SPM-BP [Li et al., 2015]	5.202	1.815	32.839
SparseFlowFused [Timofte and Van Gool, 2015]	5.257	1.627	34.834
DM+DeepFlow [Weinzaepfel et al., 2013]	5.377	1.771	34.751

(a) **Clean**

	EPE [px]		
	All	Noc	Occ
Flow Fields [Bailer et al., 2015]	**5.810**	**2.621**	31.799
DiscreteFlow	6.061	2.877	32.025
DiscreteFlow [Menze et al., 2015a]	6.077	2.937	**31.685**
DM+EpicFlow [Revaud et al., 2015]	6.285	3.060	32.564
TF+OFM [Kennedy and Taylor, 2014]	6.727	3.388	33.929
Deep+R [Drayer and Brox, 2015]	6.769	2.996	37.494
SparseFlowFused [Timofte and Van Gool, 2015]	7.189	3.286	38.977
DM+DeepFlow [Weinzaepfel et al., 2013]	7.212	3.336	38.781
FlowNetS+ft+v [Fischer et al., 2015]	7.218	3.752	35.445
SPM-BP [Li et al., 2015]	7.325	3.493	38.561
AggregFlow [Fortun et al., 2014]	7.329	3.696	36.929

(b) **Final**

Table 5.5: *Evaluation on MPI Sintel test set. This table lists the top 10 ranked methods on the MPI Sintel flow benchmark in terms of endpoint error (EPE) in all, matched (Noc) and unmatched (Occ) regions. The results on the clean images are shown in panel (a). The results on the final set are shown in panel (b).*

As opposed to the 2012 benchmark, the KITTI 2015 data set contains annotations for individually moving objects. Consequently, the optical flow fields become inhomogeneous with more pronounced motion boundaries. At the time of writing (in spring 2016), the online table does not contain many dedicated optical flow approaches. Thus, the results on this benchmark are presented together with the results of scene flow approaches in Table 5.7. The additional information used by individual methods is marked in the second column. Due to more complex motions, the overall outlier percentage increases from around 5% to above 8%. Note that these numbers are not directly comparable as the error metric of KITTI'15 differs slightly from the previous version. Details on the computation of the outlier percentage are provided in Section 5.3. Compared to

	Outliers [%]		EPE [px]	
	Noc	All	Noc	All
DiscreteFlow	**5.74**	14.29	**1.2**	3.1
PH-Flow [Yang and Li, 2015]	5.76	**10.57**	1.3	2.9
Flow Fields [Bailer et al., 2015]	5.77	14.01	1.4	3.5
NLTGV-SC [Ranftl et al., 2014]	5.93	11.96	1.6	3.8
DDS-DF [Wei et al., 2014]	6.03	13.08	1.6	4.2
TGV2ADCSIFT [Braux-Zin et al., 2013]	6.20	15.15	1.5	4.5
DiscreteFlow [Menze et al., 2015a]	6.23	16.63	1.3	3.6
BTF-ILLUM [Demetz et al., 2014]	6.52	11.03	1.5	**2.8**
Data-Flow [Vogel et al., 2013a]	7.11	14.57	1.9	5.5
DM+DeepFlow [Weinzaepfel et al., 2013]	7.22	17.79	1.5	5.8
DM+EpicFlow [Revaud et al., 2015]	7.88	17.08	1.5	3.8

Table 5.6: *Evaluation on KITTI 2012 test set. This table shows the performance of the top 10 ranked optical flow methods in terms of outliers and endpoint error (EPE) in non-occluded and all annotated regions. For comparability, only pure optical flow methods are shown, excluding motion stereo methods and techniques which use stereo information or more than two frames as input.*

EpicFlow, the error rate for non-occluded regions achieved with DiscreteFlow decreases, in this case by 34%. Classical variational approaches [Brox et al., 2004, Sun et al., 2013] perform poorly on this data set.

5.2.4 Qualitative Results

The results of individual stages of the proposed processing pipeline are provided in this section. The contained figures visualize the contribution of each step and compare the final result to that of EpicFlow on training data from Sintel and KITTI'15.

Figure 5.5 provides qualitative results on the training portion of MPI Sintel. The top row depicts the ground truth optical flow map next to the reference input image. In this sequence, the characters are engaged in a fight yielding large and heterogeneous displacements. The second row shows the result of the oracle experiment described in Section 5.2.2. To give an impression of the quality of the proposed particle set, the ground truth map is employed to select the best proposal for each pixel. Next to the oracle result, there is a color-coded error map. Each pixel encodes the endpoint error at the respective image location. Blue colors indicate errors below the 3 px threshold while yellow and red colors indicate outliers. The logarithmic color-code is provided below the figure. The oracle result visually confirms the completeness of the particle set, as there is a valid proposal for almost each pixel in the reference frame. The circular structures in this visualization result from discretization artifacts due to the integer-valued flow proposals. Consequently, they remain present in the result of the forward pass of discrete optimization (3rd row) and after the backward pass and outlier removal (4th row). This effect vanishes after sub-pixel refinement as

	Setting	Outliers [%]			
		All	All-bg	All-fg	Noc
OSF	st	**8.06**	**5.38**	**21.50**	**6.56**
OSF [Menze and Geiger, 2015]	st	8.37	5.62	22.17	6.83
PRSF [Vogel et al., 2013b]	st	14.39	11.73	27.73	9.97
DiscreteFlow		20.35	18.82	28.02	11.50
SGM+SF [Hornacek et al., 2014]	st	22.24	20.91	28.90	15.51
DiscreteFlow [Menze et al., 2015a]		22.38	21.53	26.68	12.18
EpicFlow [Revaud et al., 2015]		27.10	25.81	33.56	17.61
DeepFlow [Weinzaepfel et al., 2013]		29.18	27.96	35.28	19.15
SGM+C+NL [Sun et al., 2013]	st	36.10	34.24	45.40	26.46
SGM+LDOF [Brox and Malik, 2011]	st	39.91	40.81	35.42	30.58
Horn Schunck [Sun et al., 2013]		42.18	39.90	53.59	34.13
GCSF [Cech et al., 2011]	st	47.00	47.38	45.08	38.74
DB-TV-L1 [Zach et al., 2007]		47.97	47.52	50.23	40.19
VSF [Huguet and Devernay, 2007]	st	49.64	50.06	47.57	41.70
HAOF [Brox et al., 2004]		50.29	49.89	52.28	42.93
PolyExpand [Farnebäck, 2003]		53.32	52.00	59.94	46.23
Pyramid-LK [Bouguet, 2000]		72.91	71.84	78.32	68.57

Table 5.7: *Evaluation on KITTI 2015 optical flow test set. The results of all ranked methods are provided, which were published by the end of 2015. Approaches using information from stereo matching are marked in the second column. On KITTI 2015 the ranking criterion is the outlier percentage of all pixels with respect to the scene flow error metric.*

depicted in the fifth row of Figure 5.5. In this example, variational post-processing converges to negligible errors in the well-textured left hand part of the image. In the weakly textured regions on the right, the error increases but stays below the three pixel threshold.

In the example shown in Figure 5.5 the proposed method performs better than EpicFlow with DeepMatches towards the end of the weapon and on the leg of the man on the right. Both segments move in the opposite direction of their surroundings. Since the post-processing is identical, DiscreteFlow provides a better initialization compared to DeepMatching in this situation.

Figure 5.6 contains a failure case on very challenging imagery. Here, the strong, articulated motion of the dragon induces heavy motion blur. Consequently, there are only a few correct proposals on the legs and the wing of the character. In addition, a strong shadow beneath it leads to black, saturated image regions, which also cause proposal generation to fail. The forward pass of the discrete optimization yields erroneous matches in these areas. During the outlier rejection stage, implausible flow estimates are correctly removed leaving relatively large undefined areas. Removed matches are shown in black in the fourth row of the figure. Due to scarce image information, EpicFlow is not able to extrapolate the reliable results correctly. This holds for the initialization with DeepMatches as well. The outlier percentage of the original implementation

exceeds that of DiscreteFlow by 2 percentage points, as the proposed method works better in image regions beneath the wing and head of the dragon.

The performance of the proposed approach on more realistic imagery is demonstrated in Figure 5.7 on a traffic scene captured from a vehicle. A static background and one individually moving foreground object lead to a significantly reduced complexity of the observed motion compared to the preceding examples. However, a shadow on the depicted car poses a challenge to appearance-based motion estimation. The particle set is not affected by this issue, as evidenced by the flawless error map in panel (b). The solution retrieved from discrete inference depends on the image information through the data term. On the white body of the car, the shadow covers identical image regions in both frames. Consequently, the optimization result in panel (c) estimates zero motion in this region. Misleading image cues are consistent between the forward and backward pass so the respective correspondences overcome the outlier rejection step. The latter works well in image regions, which become occluded by the moving vehicle in the second frame. Here, forward and backward pass yield inconsistent results and the corresponding matches are pruned. Based on the erroneous initialization around the shadow, the continuous refinement consolidates the erroneous zero motion estimate on the car. The result of EpicFlow in the last panel shows that in this case DeepMatching is not able to guide the motion estimation correctly on major parts of the vehicle.

The described effects are confirmed by the very complex example shown in Figure 5.8. Small errors on several cars are successfully removed and interpolated while the reflection of a lane marking on the visible front door of the leftmost vehicle causes a similar error as described above. Panel (b) shows that the proposal set contains valid hypotheses for the entire leftmost car, which are not completely recovered during inference. This is due to misleading image evidence on reflecting and transparent surfaces. A detailed discussion of this sequence is provided in Section 5.4.2 where the results of all proposed methods are compared.

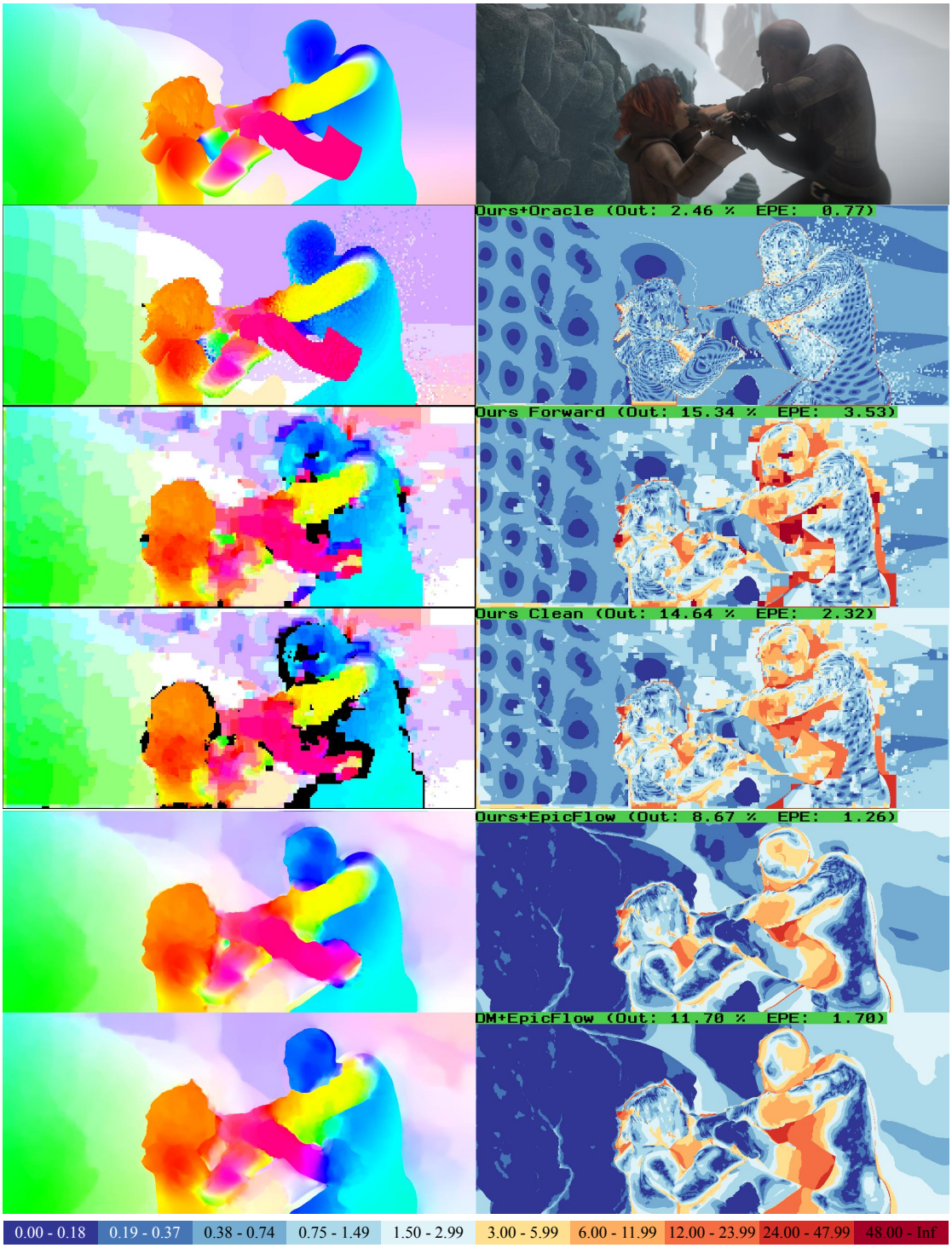


Figure 5.5: *Qualitiative results of DiscreteFlow on MPI Sintel training data. From top-to-bottom: Reference data and input image, optical flow result and error map of the oracle solution, our result without refinement, our result with outlier removal, our method with EpicFlow refinement, the result of EpicFlow [Revaud et al., 2015]. The error maps encode the endpoint error using the logarithmic scale provided in the last row.*

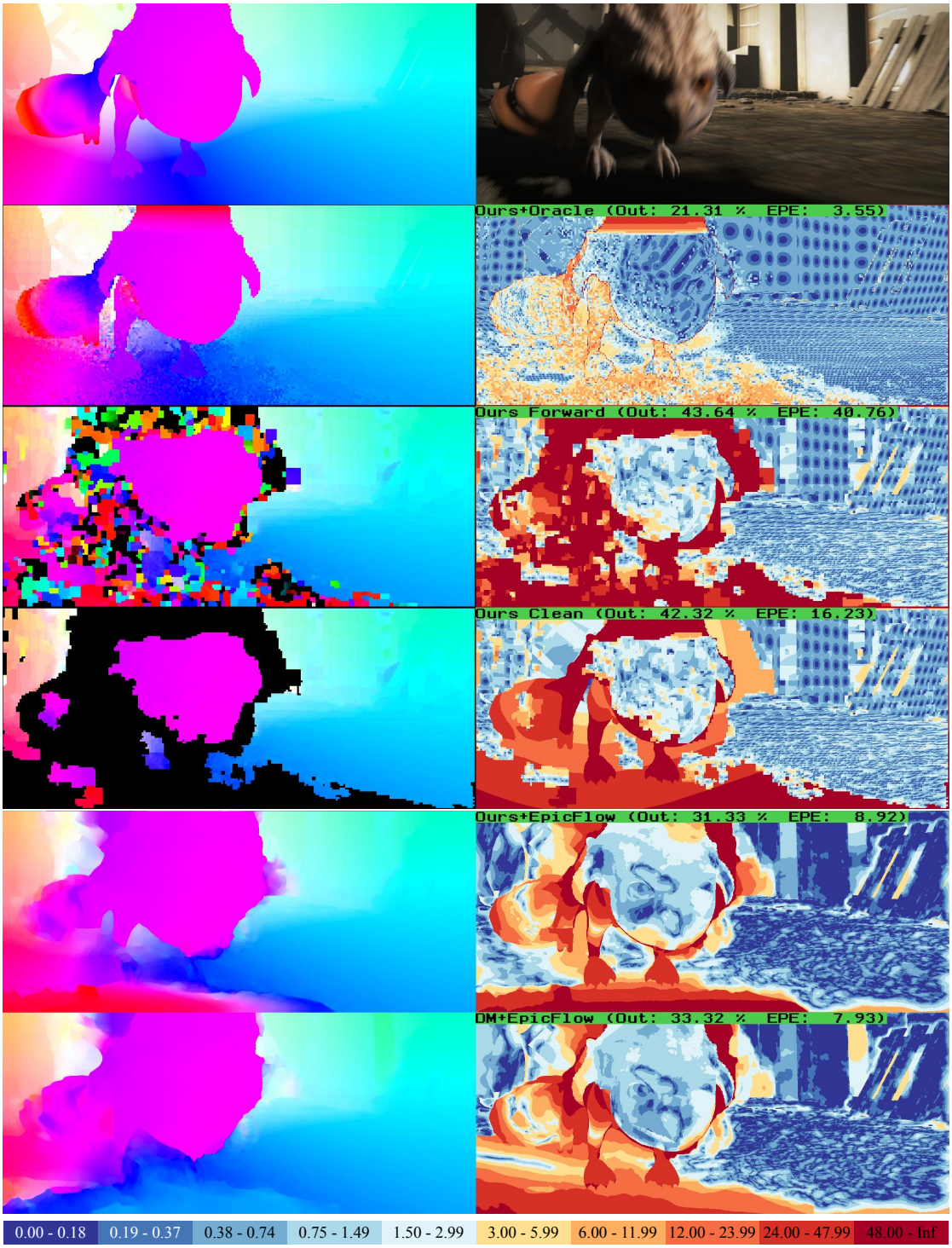
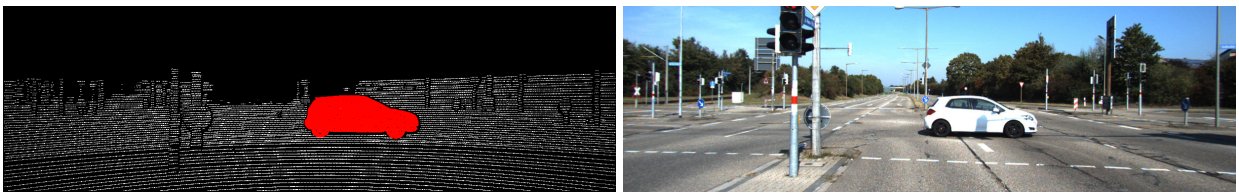
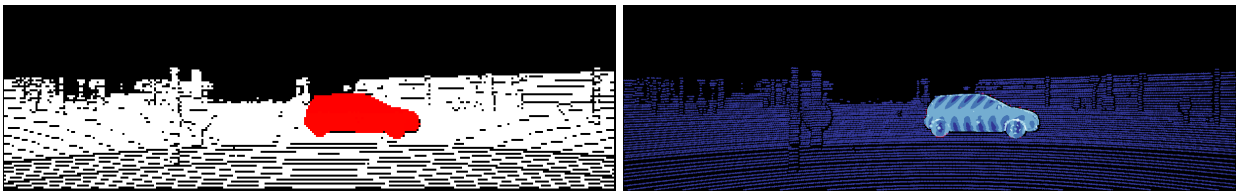


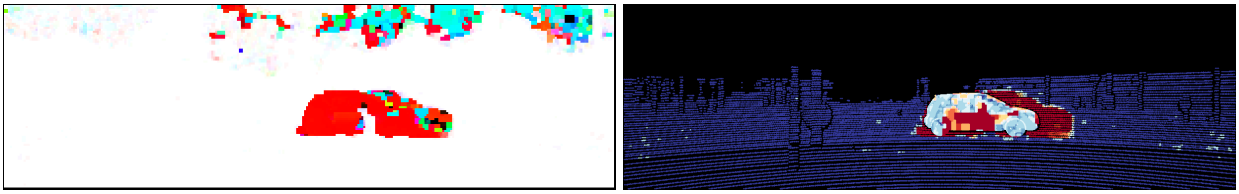
Figure 5.6: *Qualitiative results of DiscreteFlow on MPI Sintel training data. From top-to-bottom: Reference data and input image, optical flow result and error map of the oracle solution, our result without refinement, our result with outlier removal, our method with EpicFlow refinement, the result of EpicFlow [Revaud et al., 2015]. The error maps encode the endpoint error using the logarithmic scale provided in the last row.*



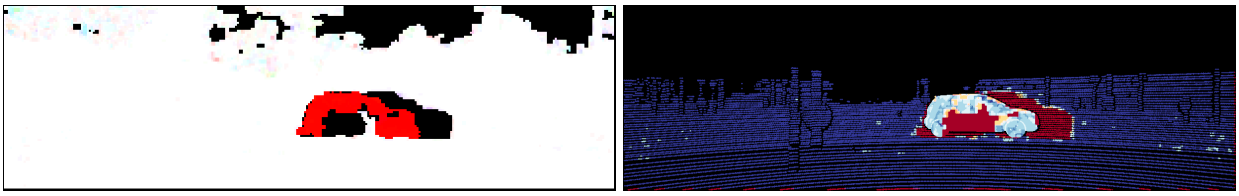
(a) *Reference data*



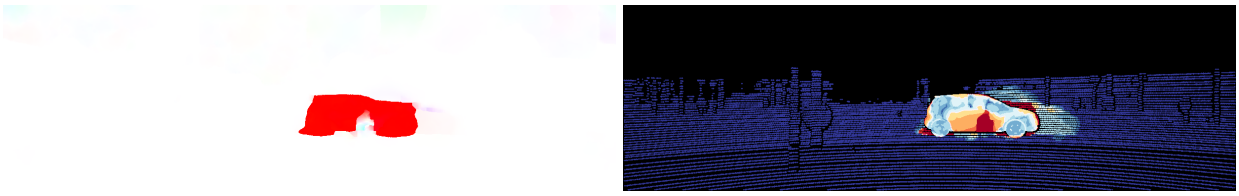
(b) *Oracle result (Out: 0.11%)*



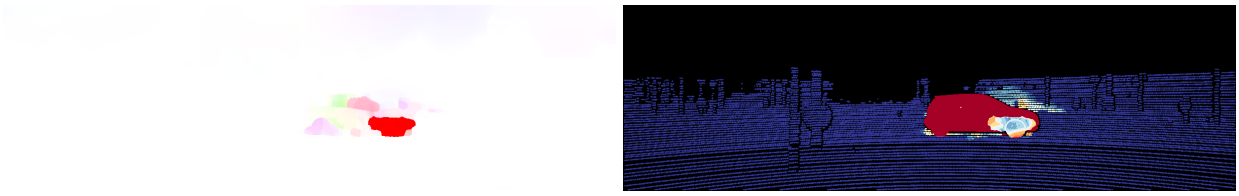
(c) *DiscreteFlow, forward pass (Out: 14.16%)*



(d) *DiscreteFlow, outliers removed (Out: 15.20%)*



(e) *Result after EpicFlow post-processing (Out: 10.58%)*



(f) *Result of original EpicFlow (Out: 23.00%)*

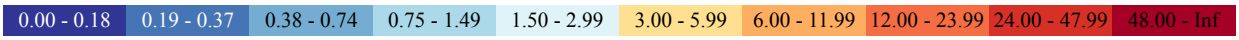
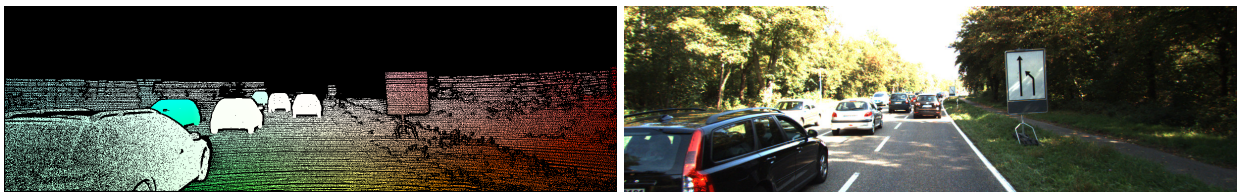
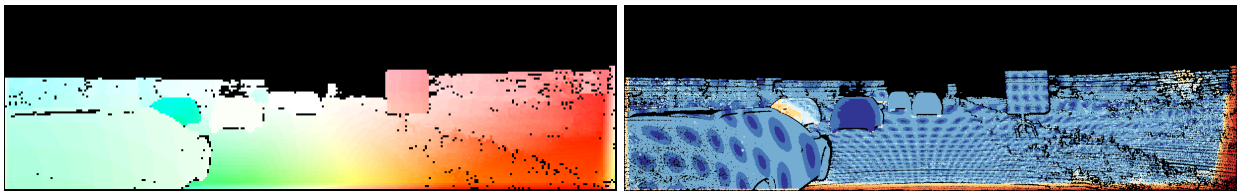


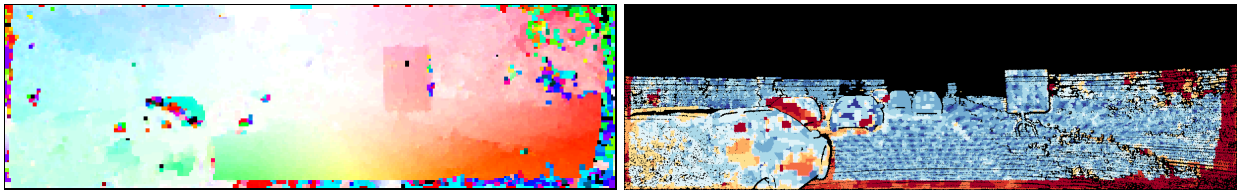
Figure 5.7: *Qualitiative results of DiscreteFlow on KITTI'15 training data.*



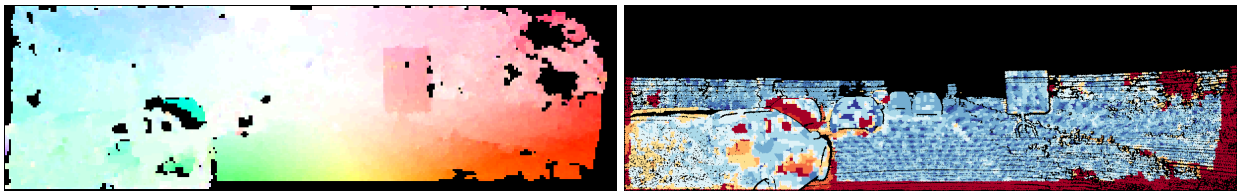
(a) *Reference data*



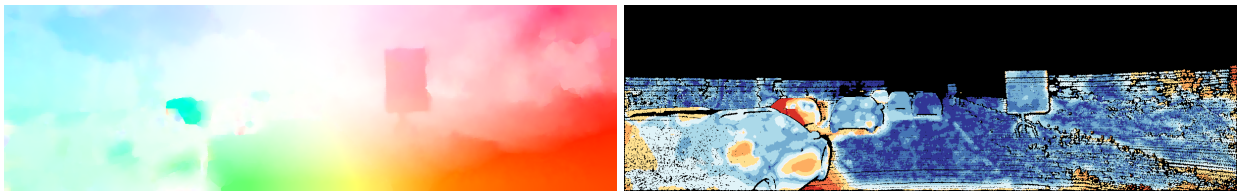
(b) *Oracle result (Out: 5.88%)*



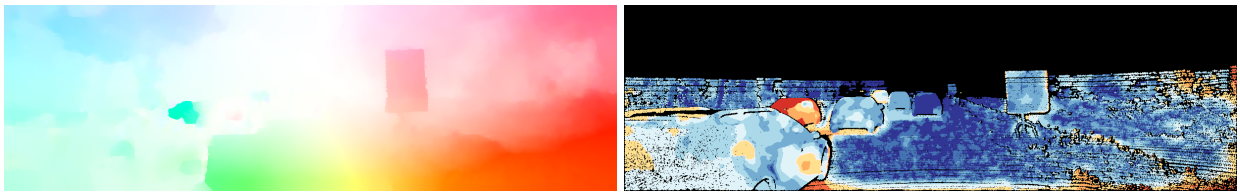
(c) *DiscreteFlow, forward pass (Out: 21.06%)*



(d) *DiscreteFlow, outliers removed (Out: 21.72%)*



(e) *Result after EpicFlow post-processing (Out: 12.38%)*



(f) *Result of original EpicFlow (Out: 10.79%)*

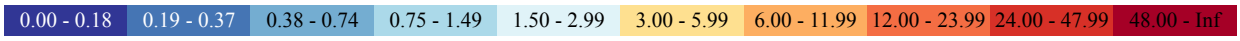


Figure 5.8: *Qualitiative results of DiscreteFlow on KITTI'15 training data.*

5.3 Object Scene Flow

In this section, the experimental evaluation of the scene flow approach is provided. After a brief introduction of an appropriate 3D error metric in Section 5.3.1, training and influence of the model parameters are described in Section 5.3.2. Section 5.3.3 gives a number of quantitative results to investigate the influence of individual parts of the model and compares the performance of the proposed method to the state-of-the-art. Qualitative results on the training data of the KITTI'15 scene flow benchmark are shown in Section 5.3.4.

5.3.1 Evaluation Protocol

To allow for a comprehensive comparison with related methods the results of the scene flow approach are reported for both KITTI data sets. According to the available reference data, the evaluation metrics differ. For the evaluation on the KITTI 2012 benchmark, the standard evaluation protocol is applied. It provides stereo and optical flow outliers separately using an error threshold of 3 pixels. As described in Section 5.1 the evaluation considers the disparity map in the first frame only. The novel KITTI 2015 scene flow data set contains appropriate reference data for the second frame. Thus at each pixel in the reference view, four values are evaluated, which uniquely determine the 3D scene flow: Two disparity values, one in the first and one in the second frame, and the displacement in x and y direction of the reference image coordinate system. This also allows one to evaluate the combination of all three measures in a single scene flow metric, which considers only pixels with correct disparities *and* flow as correct results. All entities are evaluated at each valid ground truth pixel in the reference view. A result is counted as an error if the disparity in one frame or the optical flow vector exceeds a distance of 3 pixels *and* 5% of its true value. This combination of error metrics ensures an evaluation, which is faithful with respect to the uncertainties in the reference data. Due to the complex annotation process, large displacements are assigned a tolerance with regard to their magnitude. Overall results for all 200 test images are retrieved by averaging errors over valid reference values of foreground and background regions and the combination of both. The outlier percentage is reported for all ground truth pixels as well as for those regions, which do not leave the image domain.

5.3.2 Parameter Training and Sensitivity

The scene flow model described in Section 4.2 contains a number of parameters, which can be trained to adapt the model to specific data sets. For the investigations in this section, they are learned on the subset of 100 training images from KITTI'15. In order to obtain the model weights $\{\theta\}$ and adjusted truncation thresholds $\{\tau\}$, the training strategy for DiscreteFlow, as described in Section 5.2.1, is adapted to the more complex dependencies between the parameters.

	$\theta_{1,\text{stereo}}$	$\theta_{1,\text{flow}}$	$\theta_{1,\text{cross}}$	$\theta_{2,\text{stereo}}$	$\theta_{2,\text{flow}}$	$\theta_{2,\text{cross}}$	θ_3	θ_4	θ_5
KITTI'15	1.0	1.0	1.0	0.0176	0.7641	0.7641	0.3750	14.7857	83.1317

	C_{max}	C_{out}	$\tau_{1,\text{stereo}}$	$\tau_{1,\text{flow}}$	$\tau_{1,\text{cross}}$	τ_2	τ_3	α
KITTI'15	0.7943	0.3590	1.8209	3.9039	3.9039	2.5559	0.2594	0.1986

Table 5.8: *This table shows the values of the scene flow model parameters after training.*

To account for some of the correlations, for example between truncation thresholds and weights, some parameters are jointly optimized in pairs. This strategy implies quadratic growth of the number of samples in parameter space and is thus restricted to the most obvious dependencies. Because the direct search approach to parameter training is not guaranteed to converge, it is applied for a fixed number of iterations. The parameter values provided in Table 5.8 are the result of 10 iterations and are used in all experiments. Note that rounding the values to the first significant post decimal digit only results in a slight decrease of performance. In the table, truncation thresholds for sparse features are denoted by $\tau_{1,\text{x}}$ where $\text{x} \in \{\text{stereo}, \text{flow}, \text{cross}\}$. The truncation value of all dense census features is denoted by C_{max}. Points leaving the image domain are penalized with C_{out}. The provided values concerning census features refer to the normalized Hamming distance between descriptors.

The sensitivity of the scene flow model to the choice of weights θ is investigated in Figure 5.9. For each weight, a range from zero to 20 times the learned value is discretized and evaluated while all remaining parameters are kept constant. The red marker indicates the respective parameter value chosen after training. The plots show the outlier percentage with respect to the combined scene flow metric; all results are computed on the validation set of KITTI'15 to ensure an independent cross-validation. The first row of Figure 5.9 shows that the weight of the dense census features has a significant impact on the performance. Switching the dense features off degrades the performance. The weights of the sparse features are not that critical, as the overall performance suffers only slightly when setting their weights to zero and the gradient of the error function for increased parameter values is not as large as for the dense terms. Regarding the smoothness terms, depicted in the second row, the weight of the boundary term has a dominant influence while the sensitivity with respect to the weights of normal and motion smoothness terms is relatively low. These results are in accordance with the detailed quantitative evaluation of all model components presented in Table 5.9 in the next section.

5.3.3 Quantitative Results

The quantitative analysis of the scene flow approach consists of two major parts. First, the importance of the model components is investigated in respective ablation studies. Second, the full model is compared numerically to the results of state-of-the-art scene flow methods.

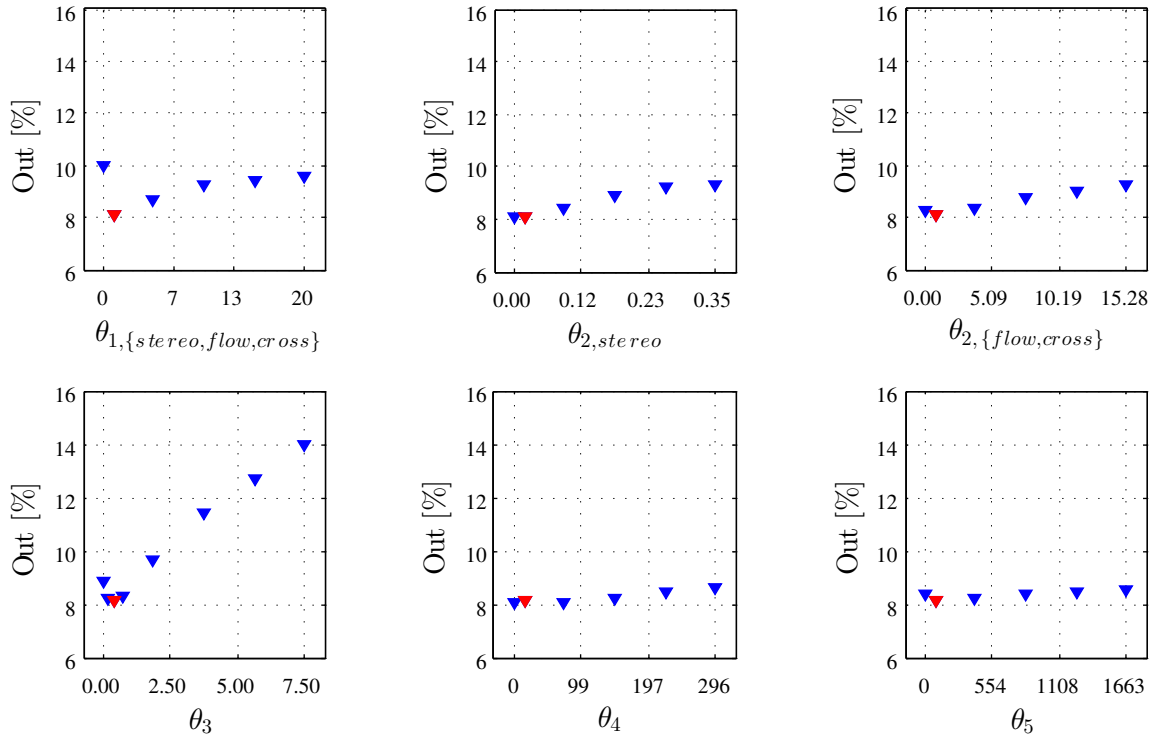


Figure 5.9: *Sensitivity of Object Scene Flow to the choice of parameters. The first row shows the impact of the data term weights of the census features (left), semi-dense disparity (center), and sparse flow and cross terms (right). The second row shows the impact of smoothness term weights for disparities in boundary pixels (left), orientation of neighboring normals (center) and the Potts model on neighboring labels (right).*

Ablation Studies

All model assumptions made to constrain the scene flow problem have been discussed in Section 4.2. Together they constitute the objective function. The following ablation studies are designed to show how well the model works when parts of the objective function are disabled or decisive parameters are set to sub-optimal values.

First, we assess the contribution of each individual term in the energy function (4.8). As before, evaluation is conducted on the validation portion of the scene flow data set. Table 5.9 shows the results when evaluating all annotated image locations. Results for non-occluded regions are provided in Appendix B. The columns show errors in terms of disparity ("D1", "D2"), flow ("Fl") and scene flow ("SF") using the conventions specified in Section 5.3.1. For each modality, the table provides results in terms of the static background ("bg"), individually moving foreground objects ("fg") as well as the combination of both ("bg&fg"). The first three rows of the table show the results of Object Scene Flow when only evaluating data terms. The overall error is comparable when using only sparse *or* dense features, and is reduced significantly using the combination of

	D1			D2			Fl			SF		
	bg	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*
Data (SPSS+SpF)	3.8	12.4	5.2	4.8	21.4	7.4	5.5	25.1	8.5	6.7	30.4	10.4
Data (Census)	4.8	12.8	6.0	5.8	14.4	7.1	6.1	19.0	8.1	7.6	25.1	10.3
Data (All)	3.9	11.0	5.0	4.8	12.9	6.1	5.4	17.4	7.3	6.6	23.0	9.1
Data (All) + Smooth (Boundary)	**3.5**	**8.6**	**4.3**	**4.4**	10.7	5.4	5.1	16.1	6.8	6.2	20.3	8.4
Data (All) + Smooth (Normal)	3.8	10.7	4.9	4.7	12.6	6.0	5.4	17.2	7.2	6.5	22.6	9.0
Data (All) + Smooth (Object)	4.0	12.1	5.3	4.9	13.2	6.2	5.4	17.2	7.2	6.6	22.7	9.1
Data (SPSS+SpF) + Smooth (All)	4.6	15.1	6.2	5.7	16.1	7.3	6.2	18.1	8.0	7.6	23.1	10.0
Data (Census) + Smooth (All)	3.6	8.9	4.4	4.5	**10.0**	5.4	5.1	15.5	6.7	6.2	19.4	8.2
Data (All) + Smooth (All)	**3.5**	9.2	4.4	**4.4**	10.6	**5.3**	**5.0**	**15.1**	**6.5**	**6.0**	**19.2**	**8.1**

Table 5.9: *Influence of scene flow model components. This table shows the error rates for disparities in the reference frame (D1) and the target frame (D2), optical flow (Fl) and scene flow (SF) averaged over all 100 validation images. For each modality, the outlier percentage is reported for the background region (bg), all foreground objects (fg) as well as all annotated pixels in the image (bg&fg). The evaluation is conducted on the validation portion of KITTI'15*

both. The next three rows show results for different combinations of all data terms and selected smoothness terms. It can be seen that the boundary term is the strongest pairwise cue on its own. In combination with all data terms, it produces the lowest error rates for disparities in the first frame. The remaining pairwise terms encourage consistently moving objects and contribute to D2, Fl and SF. Again, the combination of all pairwise terms, shown in the bottom row, yields the overall best scene flow results. The last three rows show the two groups of data terms together with all smoothness terms and the full model at the bottom. In combination with the pairwise terms, the dense census features almost reach the final result. However, the combination of all features yields a slight improvement.

Next, we investigate the performance of the full model with respect to the size of the object set and the number of iterations in Figure 5.10. Towards this goal, the number of allowed object hypotheses in our model is varied from 1 to 15. Allowing only one object restricts the model to completely static scenes. The left panel of Figure 5.10 affirms our assumption that the outdoor scenes we consider can be described sufficiently well by a small number of rigidly moving objects. The overall error, shown in red, drops significantly from 1 to 5 objects. It improves slightly to its minimal value at 10 objects and then starts to rise again. A moderate number of ten objects accounts for two phenomena: On the one hand, it covers complex scenes with many visible objects and distinct motions. On the other, it ensures a large enough number of object hypotheses to allow for some false detections from the motion-based segmentation and still cover the true objects. As can be expected, the background error is not affected by the number of objects since it typically corresponds to the largest object. The right panel of Figure 5.10 shows the performance of our method with respect to the number of iterations in the particle-based optimization framework. This plot shows that the error rate reduces significantly within the first 5 iterations and that 10 iterations are sufficient to achieve almost optimal performance under our model.

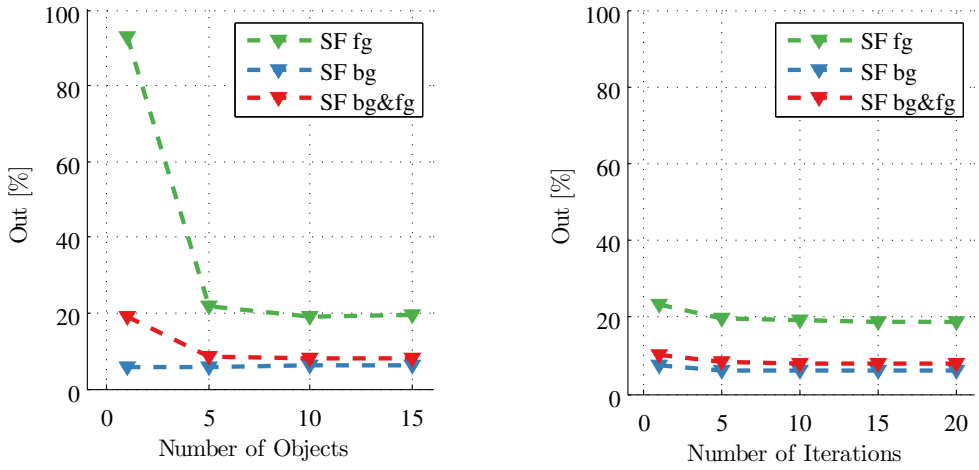


Figure 5.10: *This figure shows the scene flow errors of the proposed method with respect to the number of object proposals and iterations of the particle-based optimization. Different colors encode the results for foreground regions (green), background regions (blue) and the combined metric (red). The evaluation is conducted on the validation set of KITTI'15.*

Based on these results we chose the parameters for the following experiments. We use 10 shape particles per superpixel, 10 objects and 5 motion particles per object. As a tradeoff between runtime and overall accuracy 10 iterations of the particle-based energy minimization are performed. All motion particles and half of the shape particles are drawn from a normal distribution centered at the MAP solution of the last iteration or the initialization, respectively. The remaining shape particles are proposed using the plane parameters from spatially neighboring superpixels. These are randomly sampled conditioned on the distance of superixel centers. Both strategies complement each other and we found their combination important for efficiently exploring the search space.

Comparison to Related Methods

To evaluate the performance of the complete scene flow model both KITTI data sets are processed. While the more recent version provides appropriate reference data there are only a moderate number of baselines listed. More related approaches have been submitted to the 2012 version of the benchmark. The corresponding results are provided in the second part of this section.

Table 5.10 compares the error rates of the proposed Object Scene Flow (*OSF*) to several baselines on all annotated image locations of the KITTI'15 test data. It contains all methods, which were submitted and published by the end of 2015. Results for non-occluded regions are provided in Appendix B. To ensure a fair comparison based on all annotated pixels the results of sparse and semi-dense methods are interpolated using a standard routine provided by the KITTI development kit. Besides the classic variational approach (VSF) of [Huguet and Devernay, 2007], the table also

	D1			D2			Fl			SF			Run-
	bg	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	time[s]
OSF	**4.11**	**11.12**	**5.28**	**5.01**	**17.28**	**7.06**	**5.38**	**21.50**	**8.06**	**6.68**	**27.58**	**10.16**	390
OSF	4.54	12.03	5.79	5.45	19.41	7.77	5.62	22.17	8.37	7.01	28.76	10.63	50 min
PRSF	4.74	13.74	6.24	11.14	20.47	12.69	11.73	27.73	14.39	13.49	33.72	16.85	150
SGM+SF	5.15	15.29	6.84	14.10	23.13	15.60	20.91	28.90	22.24	23.09	37.12	25.43	>600
SGM+C+NL	5.15	15.29	6.84	28.77	25.65	28.25	34.24	45.40	36.10	38.21	53.04	40.68	270
SGM+LDOF	5.15	15.29	6.84	29.58	23.48	28.56	40.81	35.42	39.91	43.99	44.79	44.12	86
HWBSF	19.61	22.69	20.12	35.72	28.15	34.46	40.74	35.53	39.87	46.42	43.99	46.02	420
GCSF	11.64	27.11	14.21	32.94	35.77	33.41	47.38	45.08	47.00	52.92	59.11	53.95	3
VSF	27.31	21.72	26.38	59.51	44.93	57.08	50.06	47.57	49.64	67.69	64.03	67.08	>600

Table 5.10: *Outlier percentage on the scene flow test data of KITTI'15. This table shows the error rates for disparities in the reference frame (D1) and the target frame (D2), optical flow (Fl) and scene flow (SF) averaged over all 200 test images. For each modality, the outlier percentage is reported for the background region (bg), all foreground objects (fg) as well as all annotated pixels in the image (bg&fg). The table contains all methods, which were submitted and published by the end of 2015.*

provides results for the sparse scene flow method of [Cech et al., 2011] (GCSF). For the initial release of the benchmark, further baselines were constructed by combining two state-of-the-art optical flow algorithms with disparity estimates in both frames obtained using semi-global matching (SGM) [Hirschmüller, 2008]. On the one hand, SGM is combined with large displacement optical flow (LDOF) [Brox and Malik, 2011], and on the other with a classical hierarchical variational approach with non-local regularization (C+NL) [Sun et al., 2013]. As a representative for RGB-D based algorithms, the results of Sphere Flow (SGM+SF) by [Hornacek et al., 2014] are provided. To emulate the required depth component 3D object points were reconstructed from all valid pixels of the SGM disparity maps. The results of the piece-wise rigid scene flow (PRSF) approach by [Vogel et al., 2013b] were computed with the original parameter setting, which was trained on KITTI'12. As before, the table contains a duplicate entry of OSF in the second row that provides the results of the first version published in [Menze and Geiger, 2015]. In contrast to the variant described in this thesis disparity observations were originally computed using SGM instead of SPS-Stereo and the sparse optical flow matches from [Geiger et al., 2011] were used instead of DiscreteFlow. The improved observations used in this work allow for a reduced number of shape particles and iterations during inference. A decrease of the outlier percentage in all evaluated categories is accompanied by a significant reduction of the required runtime.

The proposed approach strictly outperforms all baselines with PRSF being the closest competitor. Both methods based on discrete inference arrive at moderate error rates below 20%. The performance of variational methods is significantly worse. The lowest error rate based on continuous optimization is achieved by the combination of SGM and C+NL, which exceeds the result of OSF by a factor of four. A visual comparison of the approaches is provided in the next section.

	Setting	Outliers [%] **Noc**	All	EPE [px] Noc	All	Run-time [s]
Displets v2		**2.37**	3.09	**0.7**	0.8	265
Displets		2.47	3.27	**0.7**	0.9	265
MC-CNN		2.61	3.84	0.8	1.0	100
PRSM	fl mv	2.78	**3.00**	**0.7**	**0.7**	300
SPS-StFl	fl ms	2.83	3.64	0.8	0.9	35
VC-SF	fl mv	3.05	3.31	0.8	0.8	300
Deep Embed		3.10	4.24	0.9	1.1	3
OSF	fl	3.21	3.97	0.8	1.0	390
OSF	fl	3.28	4.07	0.8	0.9	50 min
CoR		3.30	4.10	0.8	0.9	6
SPS-St		3.39	4.41	0.9	1.0	2
PCBP-SS		3.40	4.72	0.8	1.0	300
DDS-SS		3.83	4.59	0.9	1.0	60
StereoSLIC		3.92	5.11	0.9	1.0	2.3
PR-Sf+E	fl	4.02	4.87	0.9	1.0	200
PCBP		4.04	5.37	0.9	1.1	300
CSPMS		4.13	5.92	1.2	1.6	6
MBM		4.35	5.43	1.0	1.1	0.2
PRSF	fl	4.36	5.22	0.9	1.1	150

(a) **KITTI'12 Disparity First Frame**

	Setting	Outliers [%] **Noc**	All	EPE [px] Noc	All	Run-time [s]
PRSM	st mv	**2.46**	**4.23**	**0.7**	**1.0**	300
VC-SF	st mv	2.72	4.84	0.8	1.3	300
SPS-StFl	st ms	2.82	5.61	0.8	1.3	35
OSF	st	3.25	6.33	1.3	2.0	390
SPS-Fl	ms	3.38	10.06	0.9	2.9	11
OSF	st	3.47	6.34	1.0	1.5	50 min
PR-Sf+E	st	3.57	7.07	0.9	1.6	200
PCBP-Flow	ms	3.64	8.28	0.9	2.2	180
PRSF	st	3.76	7.39	1.2	2.8	150
MotionSLIC	ms	3.91	10.56	0.9	2.7	11
DiscreteFlow		5.74	14.29	1.2	3.1	180
PH-Flow		5.76	10.57	1.3	2.9	800
Flow Fields		5.77	14.01	1.4	3.5	23
NLTGV-SC		5.93	11.96	1.6	3.8	16
DDS-DF		6.03	13.08	1.6	4.2	60
TGV2ADCSIFT		6.20	15.15	1.5	4.5	12
BTF-ILLUM		6.52	11.03	1.5	2.8	80

(b) **KITTI'12 Optical Flow**

Setting | fl: Flow | mv: Multiview | ms: Motion Stereo | st: Stereo

Table 5.11: *Outlier percentage on the test data of KITTI'12. These tables show the error for disparities in the reference frame (a) and optical flow (b) averaged over all 195 test images. For each modality, the outlier percentage and the average endpoint error are reported for all annotated image sites (All) and non-occluded regions (Noc).*

A number of explicit scene flow approaches and closely related methods have been evaluated on the KITTI'12 data set. Table 5.11 provides the results on the respective test data. The table also contains specialized methods that compute only one of the sought entities, disparity *or* optical flow, to give a comprehensive overview. To distinguish them from combined methods, which compute both results, the column *Setting* contains information on which additional information is used. In particular, *fl* and *st* denote the usage of optical flow and stereo matching results. Processing of more than two successive images is indicated by the flag multi-view (*mv*). The respective methods PRSM and VC-SF are extensions of PRSF to longer image sequences. Considering joint motion estimation and reconstruction, they define the state-of-the-art on this benchmark. Another helpful information on this data set is the absence of individually moving objects. It allows imposing the epipolar constraint between the whole image regions of subsequent frames. Methods, which exploit this fact, are marked as motion stereo (*ms*) approaches. They also achieve top rankings among the combined methods. The reported runtime refers to different computer set-ups and can only serve as a rough orientation.

Comparing the proposed OSF to the two-frame variant of PRSF, there is an improvement in terms of the optical flow error in non-occluded regions of 14%. The disparity in the first frame is also estimated more accurately by OSF reducing the error rate by 26%. However, both methods do not achieve the performance of recent dedicated stereo methods.

5.3.4 Qualitative Results

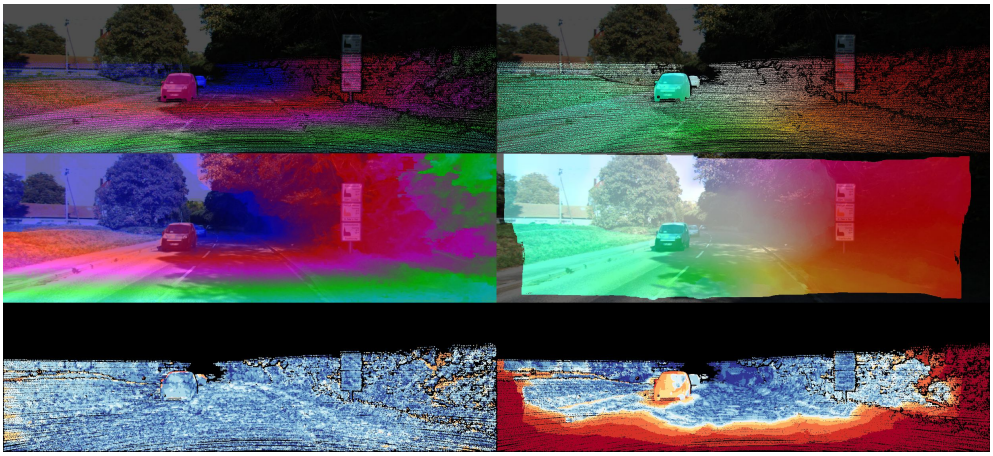
The first part of this section compares the results of Object Scene Flow to those of competing methods on a representative scene. The second part describes exemplary results of OSF in depth.

Figure 5.11 depicts the results of OSF and PRSF next to the best performing variational baseline. Each panel is split into a visualization of disparity in the reference frame, shown in the left part, and optical flow, shown on the right. The first row depicts reference data superimposed on the first image of the left camera; the second row provides the estimated results. The evaluation of both entities is given in the third row in terms of a color-coded error map. Note that the percentage error is mapped so that inliers according to the scene flow metric are depicted in blue shades. The optical flow field estimated by C+NL is only correct near the center of the image where the observed displacements are conceptually small. However, the combined baseline cannot benefit from the accurate reconstruction which is a systematic disadvantage compared to the integrated scene flow methods. The latter are able to propagate motion estimates based on the recovered surfaces. Although PRSF correctly reconstructs the geometry of the left car, it misses the associated motion. This could be due to inaccurate motion proposals. A background region at the left image boundary is also associated with an erroneous motion hypothesis. Image evidence is scarce in such regions as many object points, which are seen in the reference frame, are not visible in the subsequent image due to the ego-motion of the camera. The proposed Object Scene

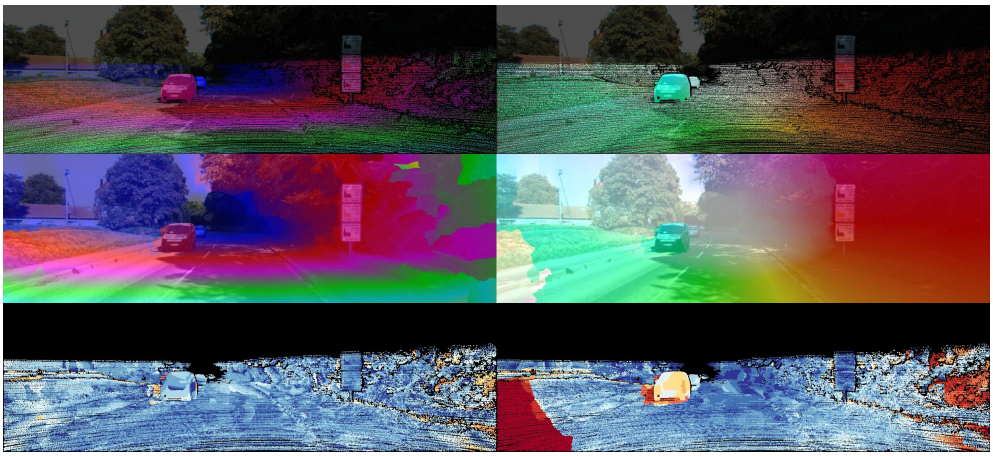
Flow assigns all background segments correctly. The optical flow result on the left car exhibits some errors. This indicates that estimating its rigid body motion helps to recover the observed displacement field. In this example, the transformation parameters were not recovered completely correct.

The next two figures provide qualitative results computed on four of the training sequences of KITTI'15. Figure 5.12 contains two examples on which the proposed method works well. As evidenced by the error images, it is able to recover the correct disparity and optical flow in a variety of challenging situations. The effect of sub-optimal segmentation is evidenced around the outline of foreground objects and on the traffic lights in both depicted scenes. It causes bleeding artifacts around the cars but most of the annotated disparities and displacements on the narrow poles are recovered correctly. The latter is important for approaches that make use of permanently installed infrastructure, e.g. to refine the position and orientation of the observing vehicle. Even objects that are not perfectly rigid are detected and assigned plausible estimates. An obvious example is the bicyclist in panel (a) of Figure 5.12. It moves at the same speed as the observing vehicle. Therefore, the optical flow map shows brighter colors indicating smaller displacements compared to the surrounding background. As neither the laser scanner nor rigid CAD models provide appropriate annotations for articulated objects, the respective image region is excluded from the evaluation. In panel (b), there are some problems in background areas close to the image boundaries. The leftmost region leaves the image domain of the three remaining frames. In this case, the smoothness terms are not able to extrapolate the disparity estimate correctly. The challenges on the bottom right are more subtle. Here, a small stretch of grass stands out from the otherwise flat ground. The rightmost portion of it is supported by enough image evidence to induce a depth jump, while the rest is smoothed by the slanted-plane model. The average scene flow error is provided in the respective captions. For these examples, it lies below the test set average reported in Table 5.10.

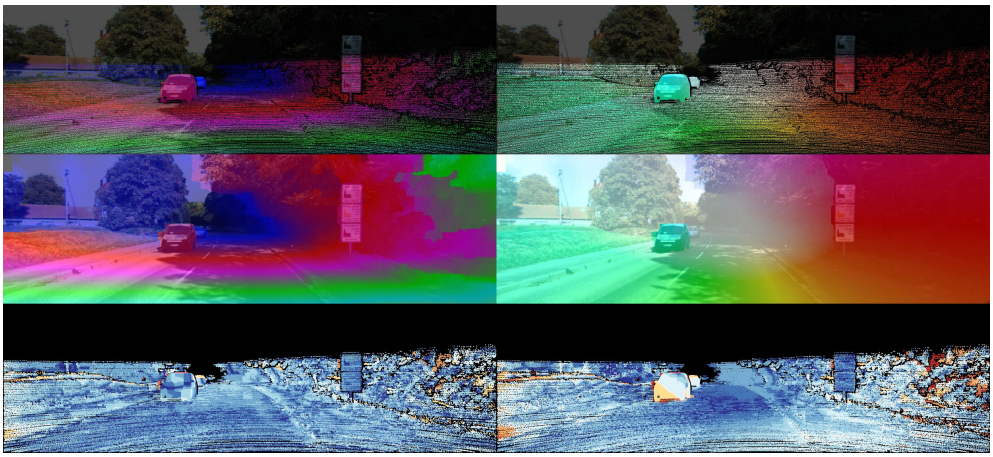
Figure 5.13 provides results with scene flow errors above the test set average. Panel (a) shows some gross errors in bushy regions next to the road. Small twigs and leaves contradicting the slanted plane assumption are responsible for them. Such errors are caused by the chosen model but in many applications, such very fine details can be considered irrelevant. On the more important, persistent parts of the scene, like the traffic sign and the visible tree trunks, OSF retrieves mainly correct disparities and displacements. The disparity map on the left shows a faithful segmentation of the delivery truck. While the reconstruction is largely successful, the optical flow result is less accurate around the roof of the vehicle. The example in panel (b) is shown as a basis for the comparison of all three proposed methods in Section 5.4.2. Difficult lighting conditions, reflecting surfaces and a quickly moving object on the opposite lane render this example especially challenging. Striking errors in the reconstruction occur on the leftmost vehicle and the bright area next to it. In addition, OSF fails to recover the motion of the approaching vehicles.



(a) SGM + C+NL



(b) PRSF



(c) OSF

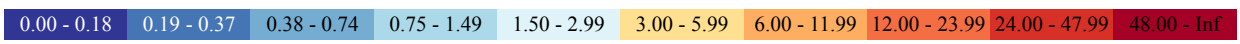
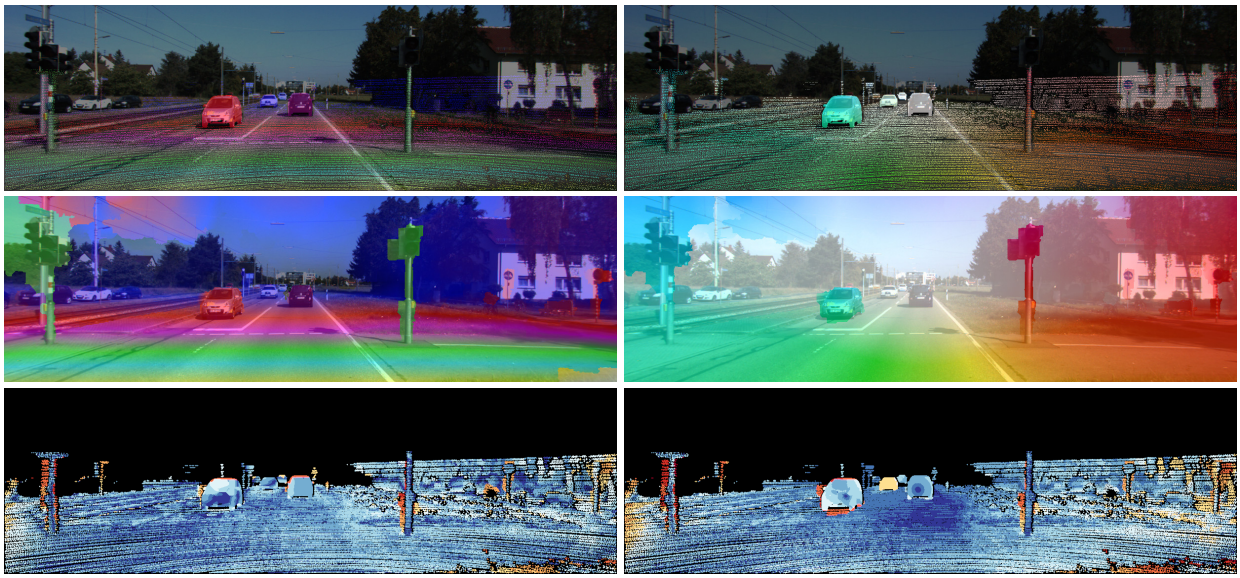


Figure 5.11: *Comparison of Results. Each subfigure shows from top-to-bottom: The disparity and optical flow ground truth in the reference view, the disparity map (D1) and optical flow map (Fl) estimated by the specified algorithm, and the respective error images using the logarithmic color scheme depicted in the legend.*

(a) *Scene flow outliers: 3.49%*



(b) *Scene flow outliers: 7.46%*

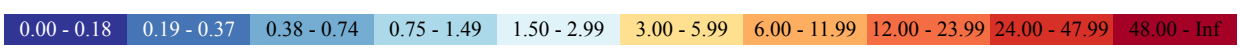
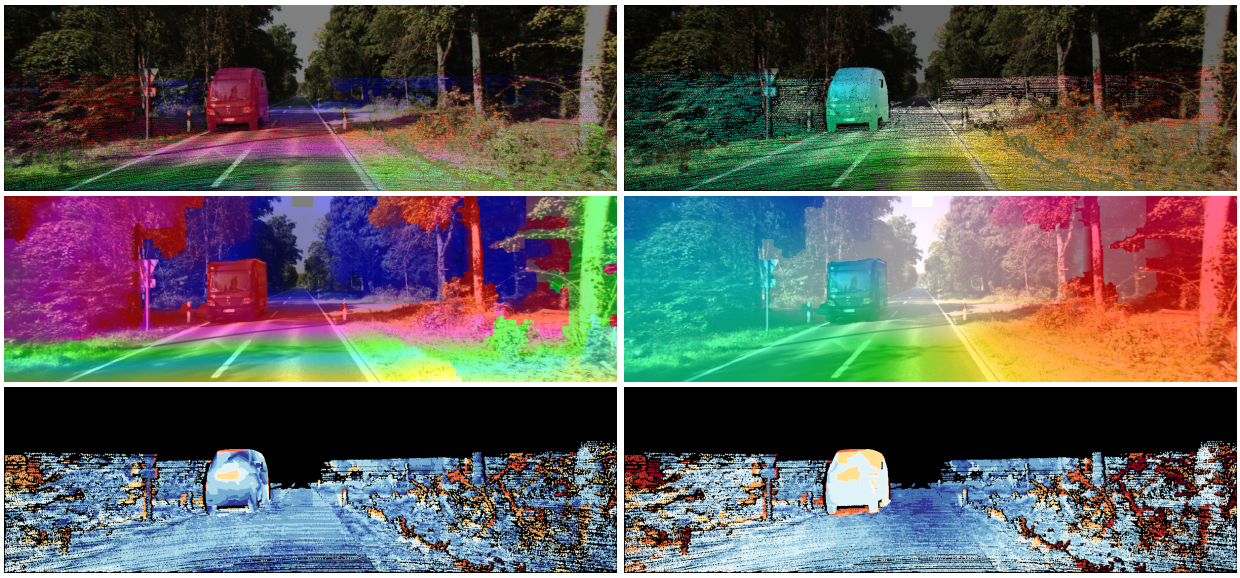
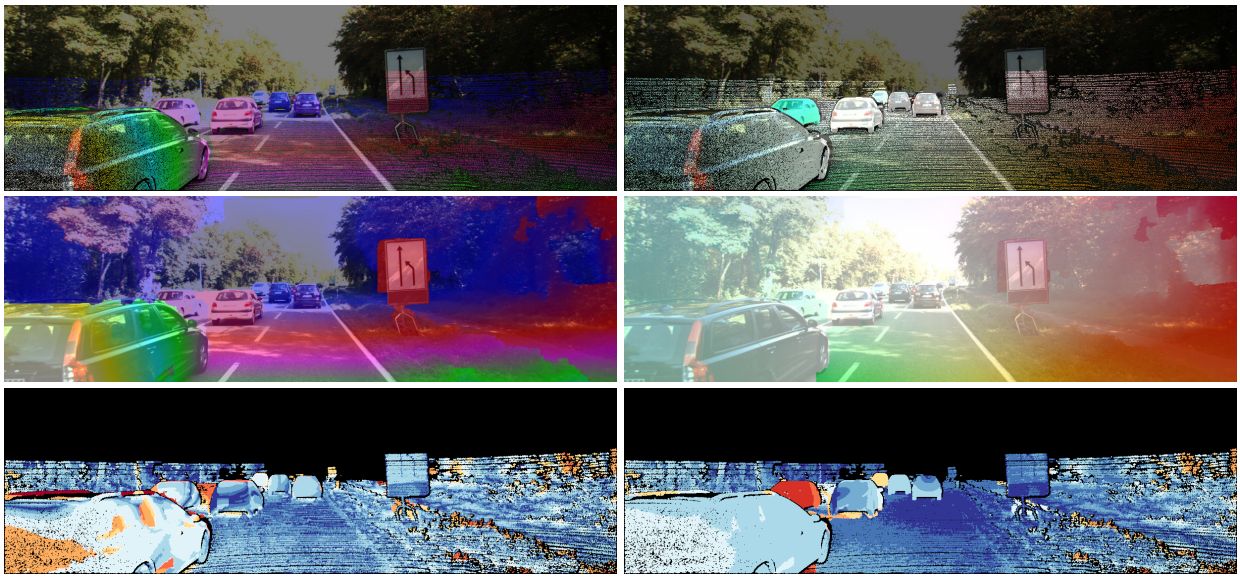


Figure 5.12: *Qualitiative results of the scene flow approach. Each sub-figure shows from top-to-bottom: The disparity and optical flow ground truth in the reference view, the disparity map (D1) and optical flow map (Fl) estimated by our scene flow algorithm, and the respective error images using the logarithmic color scheme depicted in the legend.*



(a) *Scene flow outliers: 13.24%*



(b) *Scene flow outliers: 14.10%*

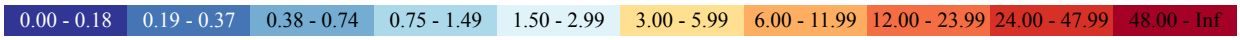


Figure 5.13: *Qualitiative results of the scene flow approach. Each sub-figure shows from top-to-bottom: The disparity and optical flow ground truth in the reference view, the disparity map (D1) and optical flow map (Fl) estimated by our scene flow algorithm, and the respective error images using the logarithmic color scheme depicted in the legend.*

5.4 Joint 3D Estimation of Vehicles and Scene Flow

To investigate the contribution of the model-based extension, both training and test data of KITTI'15 are processed. Section 5.4.1 reports numeric results addressing different aspects of the approach before Section 5.4.2 provides detailed qualitative results and a visual comparison of results from all three proposed methods.

5.4.1 Quantitative Evaluation

Aiming at a parametrized reconstruction of cars, the model-based extension of the scene flow approach is tailored to traffic scenes containing individually moving objects. Considering the publicly available data sets described in Section 5.1, the proposed method can only be evaluated on data from KITTI'15. By design, it is not able to detect and reconstruct static cars as contained in KITTI'12, e.g. parking next to the road. This section provides numeric results assessing the ability to adapt the active shape model to observed cars and comparing the error rates on the test set to those achieved with the basic model.

Initialization	0.52
After optimization	**0.59**

Table 5.12: *Model Coverage. This table shows the intersection-over-union criterion, averaged over all detected foreground objects in the 200 training images of KITTI'15 before and after optimization.*

Shape Adaption

To quantify the improvement gained by optimizing the model parameters, we evaluate the established intersection-over-union (IOU) criterion, which is frequently used for evaluating segmentation and object detection in the literature. It evaluates the consistency of annotated reference data and image regions assigned to objects by respective detectors. Here, we compare the ground truth mask of the annotated objects to the mask of the projected 3D model as inferred by the proposed method. Objects without successful initialization are discarded as they cause a constant offset in the considered metric. On the training data, the motion-based segmentation is able to detect 258 of 431 annotated cars correctly, corresponding to 60% true positive detections. The reported intersection-over-union score is averaged over all detected cars. Table 5.12 compares the results after initialization to our final results. The evaluation is conducted on all 200 training images from KITTI'15. The average intersection-over-union score is improved by 12%.

	D1			D2			Fl			SF		
	bg	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*
OSF	**4.11**	**11.12**	**5.28**	**5.01**	**17.28**	**7.06**	**5.38**	**21.50**	**8.06**	**6.68**	**27.58**	**10.16**
Model-based OSF	4.47	11.62	5.66	5.70	18.30	7.80	6.03	23.32	8.90	7.53	29.35	11.16

Table 5.13: *This table shows the comparison of the basic scene flow approach to the model-based extension on the test data of KITTI'15. The percentage of outliers is specified with respect to disparity estimates in the subsequent stereo pairs (D1,D2), optical flow in the reference frame (Fl) and the complete scene flow vectors (SF).*

Scene Flow Error

The overall performance of the model-based scene flow approach is investigated on the test data of KITTI'15. Table 5.13 shows the quantitative effect of incorporating the 3D object model in terms of the scene flow error metric discussed in Section 5.3.1. As a baseline, we optimize Equation 4.23 without the shape consistency term κ and sample only motion particles for the objects. This setup corresponds to Object Scene Flow. In contrast, model-based OSF also optimizes shape and pose parameters of the 3D model as described in Section 4.3.1. Table 5.13 shows that the performance decreases moderately in all categories. This is to be expected as the model can adapt to most but not to all depicted objects. Examples of both situations are discussed in the next section.

5.4.2 Qualitative Results

To give an impression of the reconstruction results it is important to complement the numeric analysis in the preceding section with exemplary qualitative results. The contents of this section are twofold. First, the effect of incorporating the ASM is visualized in detail. Second, the results of all three proposed approaches will be compared to each other in depth based on a dedicated figure.

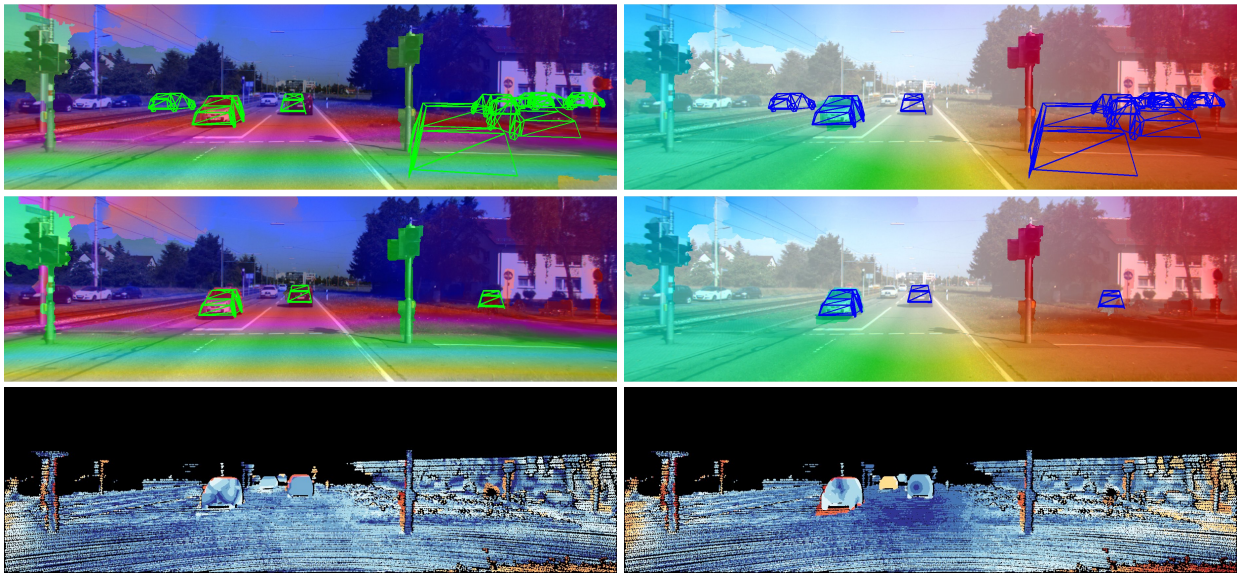
Figures 5.14 to 5.16 illustrate resulting disparity and optical flow maps together with wire-frame renderings of the object models. Note that only fully visible faces of the CAD models are rendered. All results are superimposed to the respective reference views of six representative scenes. The top row of each sub-figure depicts the layout after initialization as described above. In most cases, the shapes do not match the observed cars and there are some significant positional offsets. In addition, there are many spurious objects initialized due to wrong object hypotheses. The center row of each panel shows our reconstruction results after optimizing the energy function (4.23). Objects, which are not assigned to any of the superpixels, are considered absent and thus not drawn. The last row provides color-coded error maps as before.

For all examples shown in Figure 5.14, the model position is successfully aligned with the observed object and the shape of the model is faithfully adapted to the depicted cars. However, both frames

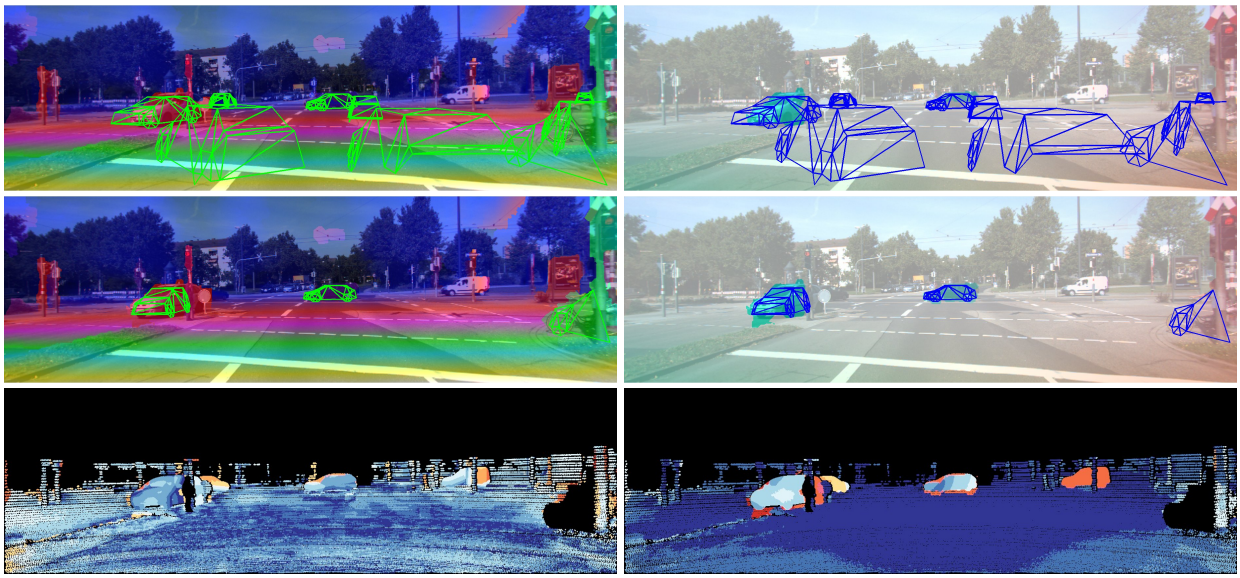
contain a bicyclist, which is correctly detected as an individually moving object in both cases. Object models are initialized on the bicyclists and approximately adjusted to the outline of the depicted objects within the range of possible deformations. This underlines the flexibility of the model but does not fully achieve the aspired goal. Further, spurious hypotheses are initialized next to the road and on some road markings. They are successfully removed, demonstrating the intrinsic model selection capability of the approach.

Figure 5.15 compares one very successful solution to a representative failure case of the method. Panel (a) shows one successfully reconstructed car in the foreground. The initialization also contains one spurious object around the traffic light. It is correctly removed during optimization. The observed scene fulfills most of the model assumptions. Apart from small bleeding artifacts, the depicted surfaces can be faithfully reconstructed based on image segments. The background motion, which corresponds to an identity transformation in this case, is clearly different from the observed motion of the only individually moving object. Under such conditions, the motion-based segmentation works as expected. Panel (b) picks up one of the examples from Section 5.3.4. In addition to the bosky regions next to the road that violate the slanted plane assumption, the depicted van exceeds the adaptive capability of the active shape model. Therefore, two models are fitted to the consistently moving region showing the van. The flow error map on the right confirms that the upper part of the vehicle is not assigned to the reconstructed object. This is the desired behavior, as the respective image segments are not covered by the object model. Under the proposed setup, failure cases like this lead to the increased average errors in Table 5.13. Another issue, which is emphasized by the model-based extension, is the persistence of erroneous objects. Complex scene geometry, as encountered next to the road in this example, can randomly generate groups of consistently deviating displacements. By design, such motion cues are likely to initialize false positive object hypotheses. Where such objects can be adapted to coincide with a sufficient amount of image evidence, they remain present even after optimization of the random field model.

False positive hypotheses are also contained in panel (a) of Figure 5.16. In this case, the corresponding image regions are correctly assigned to the background during inference. This example demonstrates the effect of partial occlusions. While small parts of the car in the center are occluded by the traffic sign, major parts of the rightmost vehicle are outside the reference view. As this car vanishes in the second image pair it is not contained in the reference data. The respective object model is nevertheless optimized to a plausible result. Both models on the leftmost cars are retrieved slightly too small causing the intersection-over-union criterion to decrease. Panel (b) shows the detailed results of the model-based scene flow approach on the example which is picked up in Figure 5.17 to compare the results of all proposed methods.



(a) *0.62 (init.) / 0.75 (opt.)*



(b) *0.56 (init.) / 0.63 (opt.)*

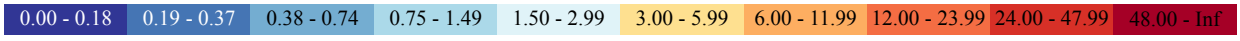
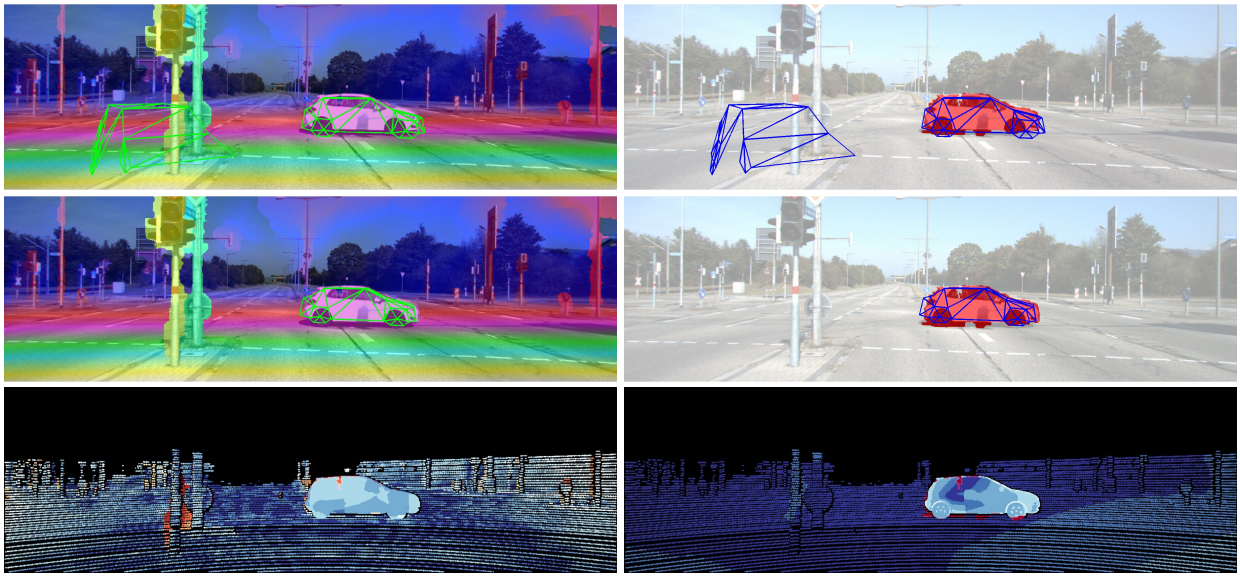
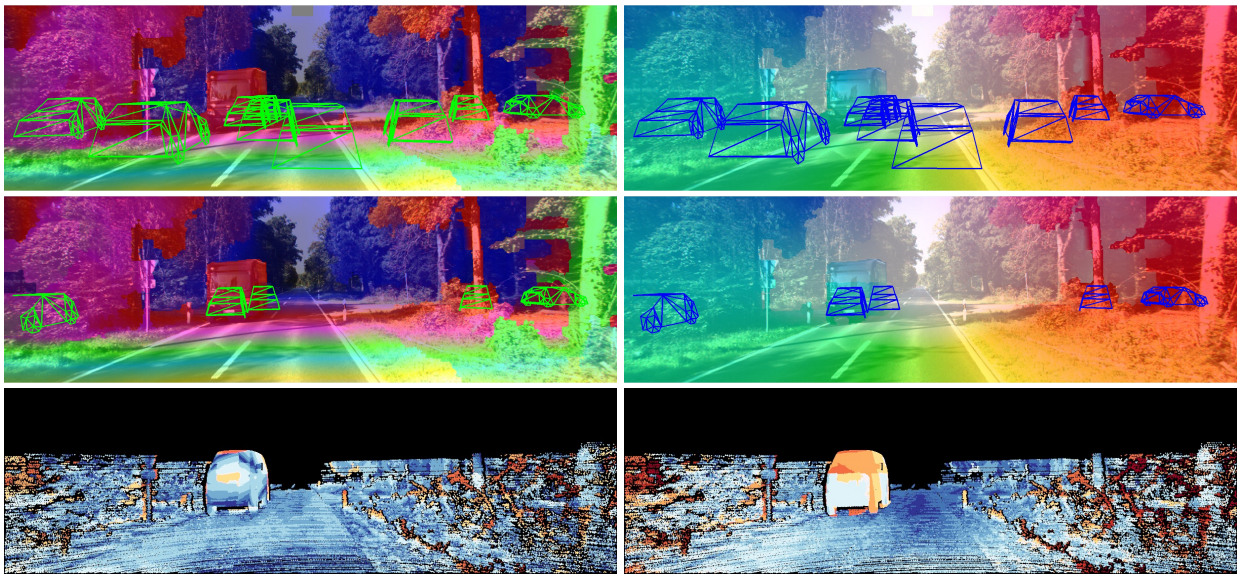


Figure 5.14: *Qualitative results of model-based scene flow estimation. Each panel shows our results at initialization (top row) and after optimization (center row). The reference view is superimposed with the color-coded disparity (left) and optical flow map (right). Object models are depicted as green and blue wire-frames. Error maps are provided in the third row. The numbers in the captions specify the value of the intersection-over-union criterion at initialization and after optimization.*



(a) *0.75 (init.) / 0.83 (opt.)*



(b) *0.20 (init.) / 0.19 (opt.)*

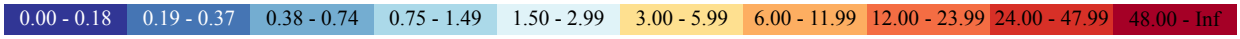
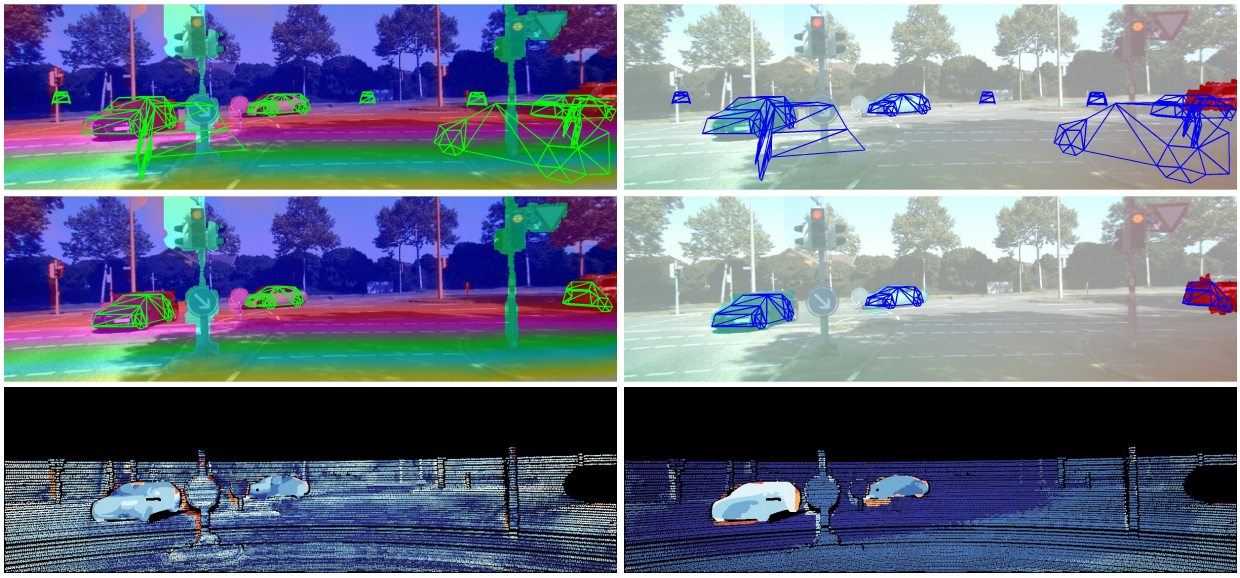
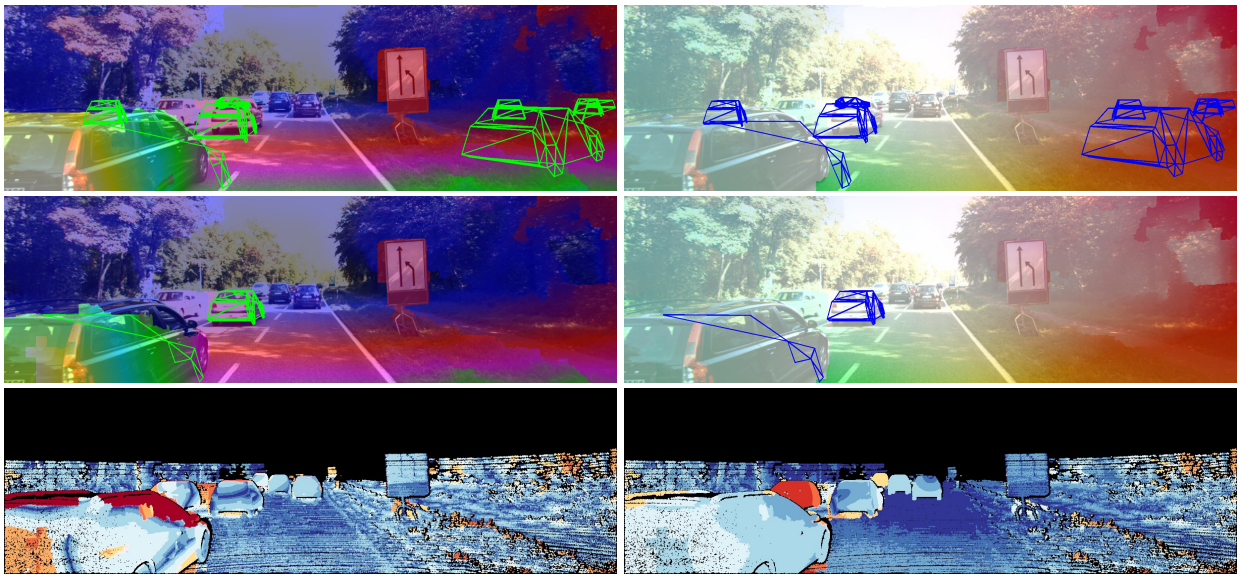


Figure 5.15: *Qualitative results of model-based scene flow estimation. Each panel shows our results at initialization (top row) and after optimization (center row). The reference view is superimposed with the color-coded disparity (left) and optical flow map (right). Object models are depicted as green and blue wire-frames. Error maps are provided in the third row. The numbers in the captions specify the value of the intersection-over-union criterion at initialization and after optimization.*



(a) *0.54 (init.) / 0.50 (opt.)*



(b) *0.37 (init.) / 0.41 (opt.)*

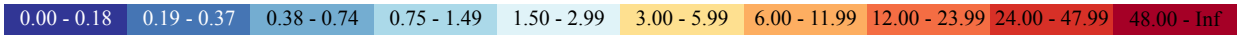
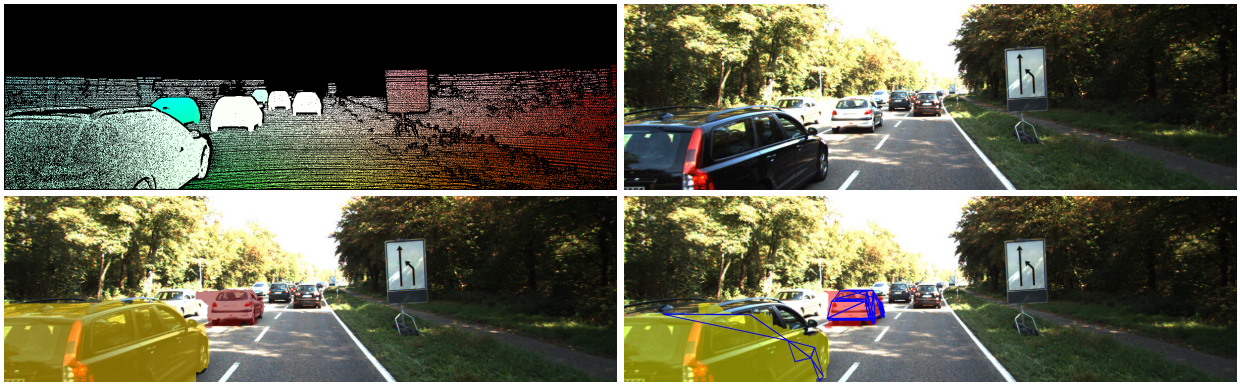


Figure 5.16: *Qualitative results of model-based scene flow estimation. Each panel shows our results at initialization (top row) and after optimization (center row). The reference view is superimposed with the color-coded disparity (left) and optical flow map (right). Object models are depicted as green and blue wire-frames. Error maps are provided in the third row. The numbers in the captions specify the value of the intersection-over-union criterion at initialization and after optimization.*

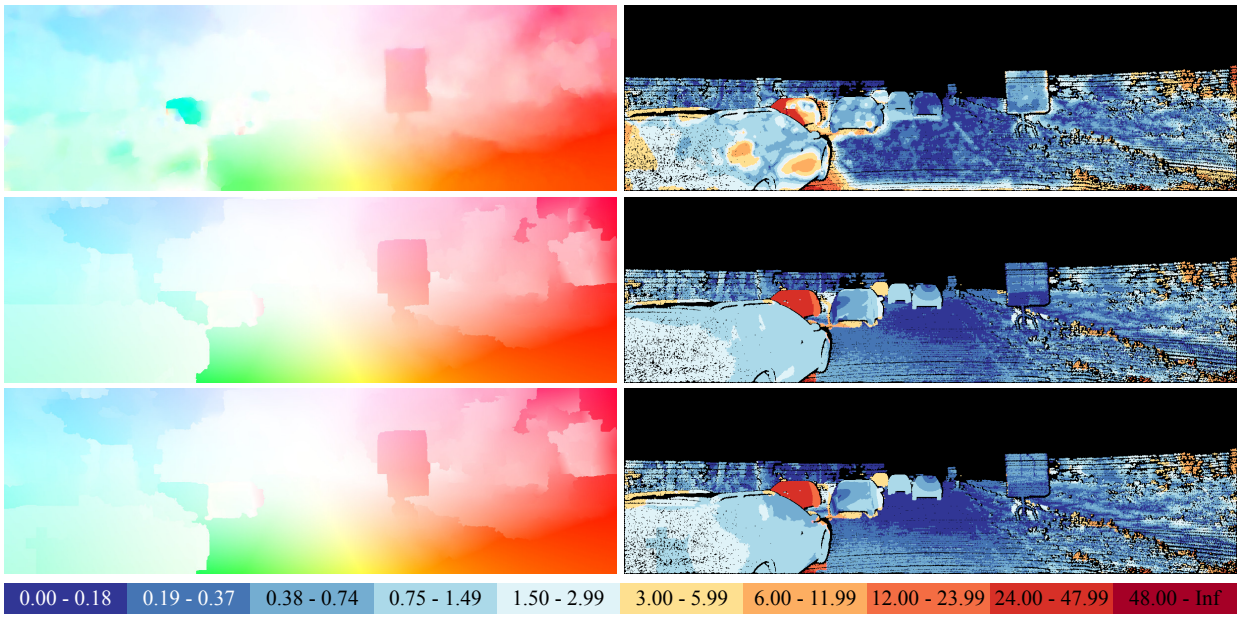
Figure 5.17 provides a visual comparison of the results of all three approaches described in this thesis. Panel (a) comprises reference data and segmentation results. The first row contains optical flow reference data and the first image from the left camera. Strong shadows and very bright regions render the imaging conditions very challenging. Containing six individually moving cars, the depicted scene is one of the most complex in the data set. However, the velocity of the vehicles is rather low as they approach a traffic jam. One result of the scene flow approaches is the assignment of object labels to the superpixels. These segmentations are shown in the second row. The labeling from Object Scene Flow is given on the left; it contains correct segmentations of two cars. Two more vehicles are moving in the same direction as the observing car. As they are located close to the center of the image and move slowly, they are missed by the motion-based segmentation. This is due to the fact that the observed motion agrees with the background motion within the respective threshold. Another two vehicles on the opposite lane are missed as well. This happens although the optical flow map from DiscreteFlow, shown in the next row, contains correctly estimated displacements on both cars. One of them appears relatively small in the input image and is likely to be missed due to larger, erroneous object hypotheses. The other is depicted in a saturated area, which does not provide strong image evidence. The smoothness terms of the scene flow model are designed to assign such ambiguous image segments to the surrounding object, which is background in this case. The image on the right shows the result of the model-based extension. It contains foreground segments on the same cars as before. This is to be expected because the algorithm generating object hypotheses is the same. The reconstruction of the cars is visualized by the blue wire-frame models. The central object is recovered correctly. The model on the leftmost car is initialized with a significant angular error. During inference, the particle-based optimization is not able to recover from the poor approximate values. Figure 5.13 shows that the reconstruction also suffers from erroneous disparity estimates on the trunk lid of this car. Although there are plausible explanations for the observed effects, this example can be considered a failure case showing the limitations of the scene flow approaches.

Panel (b) depicts the results and optical flow error maps for DiscreteFlow, Object Scene Flow and the model-based extension in turn. As described above, the optical flow method is able to correctly recover most parts of the cars on the opposite lane. Motion discontinuities are overly smoothed in the inhomogeneous region left of the image center. This effect can be attributed partly to the variational refinement. In the aforementioned region, it lacks the support of discrete matches as they are rejected by the outlier detection step. However, a bias towards very smooth motion boundaries can also be observed around the traffic sign on the right where discrete matches are available. Furthermore, DiscreteFlow encounters problems on reflecting parts of the leftmost vehicle. The image information in such regions changes over time due to the reflection and not according to the motion field in the geometric sense. Most image matching techniques have problems in such situations. The slanted plane model of the scene flow approaches helps to overcome these problems as evidenced in the flow map of Object Scene Flow in the second row. Thanks to the regularization within the patches and smoothness constraints between them, the

surface of the car is correctly recovered even in this challenging region. Assigning the correct object label leads to the faithfully recovered displacements. The last row of panel (c) shows the results of the model-based extension. Clearly, the segmentation is affected by the sub-optimal reconstruction of the leftmost car. According to the design of the shape and pose consistency term the labels of image segments are influenced by the model coverage. In this case, erroneous background labels around the windshield of the leftmost car do not induce errors in the estimated optical flow. The corresponding surface patches are pushed to the background leading to gross errors in disparity but consistent motion vectors. In the scene flow evaluation, described in Section 5.3.1, the corresponding image regions are faithfully penalized as outliers while an isolated evaluation of disparity and optical flow will only reveal a part of the problem.



(a) *Reference data and input image (top), retrieved objects (bottom)*



(b) *Retrieved optical flow and error maps*

Figure 5.17: *Comparison of the proposed methods. The object labels assigned by the scene flow approaches are shown in the second row of panel (a). Optical flow and error maps are depicted in panel (b): DiscreteFlow (top), Object Scene Flow (center), model-based extension (bottom).*

Chapter 6

Discussion

In this chapter, we will discuss the results provided in Chapter 5 in detail. Section 6.1 investigates the results of the optical flow approach. The experimental results of Object Scene Flow are evaluated in Section 6.2. It compares the proposed method to related algorithms for joint motion estimation and reconstruction as well as the novel optical flow method described in Section 4.1. Insights into strengths and weaknesses of the model-based scene flow approach are the subject of Section 6.3

6.1 Optical Flow by Discrete Optimization

For the sake of clarity, the discussion of the optical flow results is structured according to the investigations in Section 5.2.

Parameter Training and Sensitivity

To keep the computational burden of parameter training tractable, a simple line search algorithm is applied as described in Section 5.2.1. Each parameter is optimized individually conditioned on the fixed values of all remaining parameters. Correlations between the sought variables are largely ignored. Investigations of the sensitivity, as summarized in Figure 5.3, indicate that the objective function used for training is non-convex. Consequently, gradient-based training would depend critically on initial values. It would likely find very local minima. The investigated error metric varies moderately with the smoothness weight and the size of the label set confirming a certain robustness of the model. For the relative weight of smoothness and data term, the minimum differs between training and validation portion. This issue could be addressed by a more sophisticated

cross-validation strategy but the effect is limited and it can be expected that the performance on the test data will improve slightly at most.

Pilot Studies

The pilot studies confirm the validity of two core assumptions behind the optical flow approach. First, the oracle experiment confirms the completeness of the proposal set. Both relevant error metrics, the endpoint error and the outlier percentage, of the oracle are significantly smaller than those computed with DiscreteFlow and competing methods. This result shows that the initialization strategy based on DAISY descriptors is suitable to generate high-quality proposal sets for the investigated data. In addition, it indicates room for improvement in the model components as well as in the inference procedure. The provided qualitative results show that there are regions where discrete optimization returns erroneous matches although correct proposals are available. In these cases, the chosen model and its parameter values do not allow to infer the correct labels. The removal of small isolated segments implements a tradeoff between robustness to noisy matches and rejection of correct associations. Image regions where no matches prevail are interpolated by EpicFlow post-processing without further guidance. The strategy for outlier removal could be improved by complementing the segment size with additional quality measures.

The second experiment described in the pilot studies validates the assumption of a diverse proposal set for the investigated benchmarks. While these prerequisites might not be met by data from more realistic imagery or from different application domains, the processed images are amenable to the presented strategy to increase computational efficiency. Pre-computation of the relevant pairwise penalties, which fall below the truncation threshold, drastically reduces the computational efforts, as only up to 2.5% of the terms actually have to be evaluated. Furthermore, a violation of this assumption will not cause the described method to fail but will lead to significantly increased computation time. For specific applications with less diverse particle sets, it might be possible to compensate this effect with an increased stride length between the nodes of the random field model.

As DiscreteFlow replaces DeepMatching in the EpicFlow pipeline, the third pilot study directly compares the performance of both image matching approaches. Table 5.4 shows that the developed strategy clearly outperforms DeepMatching. As DiscreteFlow is based on the DAISY descriptor, it will not be as robust against wide baselines and strong changes in the viewing angle as DeepMatching but it handles the displacements present in the investigated diverse data sets very well.

An advantage of the presented discrete optimization approach is that, by design, a regular grid of matches is computed. This is not to be confused with hierarchical matching, as the target image is not re-sampled. The sought displacement vectors are not re-scaled but optimized at full

resolution. The distribution of matches is improved compared to local feature matching, where associations tend to agglomerate in image regions with strong texture. The formulation in terms of a regular random field of matches ensures a uniform distribution of associations throughout the reference image. On the other hand, choosing the image locations considered in matching according to this regular structure discards the accuracy potential of distinct, good features to track. Up to now, this algorithmic shortcoming is compensated by the overall strategy of first computing integer-valued displacements on the grid and then refining them to sub-pixel accuracy using a variational method, which takes into account dense image information. One possible extension to the current approach would be to allow the investigated image sites to move slightly, depending on local image structure. This is one promising starting point for future work.

Comparison to Related Methods

The quantitative evaluation in Section 5.2 shows that, in combination with the variational post-processing step, DiscreteFlow achieves state-of-the-art performance. On all three investigated data sets, it is ranked in the leading position except for the final pass of Sintel where Flow Fields perform best. Compared to the most closely related approach EpicFlow the presented method performs slightly better on Sintel. On KITTI'12 and '15, the margin increases to 27% and 34%. At the time of writing, Flow Fields is the top performing method on the final portion of MPI Sintel. It reduces the average endpoint error by 4% compared to DiscreteFlow. Flow Fields builds on the same processing pipeline as DiscreteFlow but follows a slightly different strategy to initialize continuous post-processing. The method was developed concurrently with the presented approach, which shows that the line of research followed in this thesis is indeed promising.

Table 5.11 shows that a significant reduction of the error rates results from the integration of additional constraints. Motion stereo approaches reduce displacement estimation to one-dimensional matching along the epipolar lines. The demonstrated error rates suggest that, in scenarios according to the assumption of static scenes, this valuable additional constraint should not be discarded. It could be easily incorporated in the initialization stage of the proposed method by reducing the search space accordingly.

The most interesting extension of the method would be to avoid the variational post-processing step and instead further refine the results by discrete optimization. This idea requires the addition of a more sophisticated mechanism to propagate promising matches to ambiguous regions. Inspiration on this issue can be drawn from methods like Generalized PatchMatch, which perform a propagation step to establish dense nearest neighbor fields. Furthermore, it would be necessary to refine the results to sub-pixel accuracy. To this end, the particle belief propagation framework, applied in the scene flow approaches, could be transferred to this method. It will be very interesting to see the development of methods based on discrete inference, as the presented pilot studies hint at further room for improvement. Current results on various benchmark data sets confirm

that a very good initialization of the displacement field is important but they also show that, at the current state, a variational post-processing step is required to retrieve dense high-accuracy optical flow.

6.2 Object Scene Flow

This section discusses the experimental evaluation of Object Scene Flow. It starts with an assessment of the evaluation protocol and the training strategy. Then, it explains the insights gained from ablation studies and the comparison to related methods.

Evaluation Protocol

The error metric described in Section 5.3.1 allows for an evaluation of the three-dimensional displacement field. However, the outlier thresholds are chosen conservatively. Such relaxed thresholds are necessary to account for inaccuracies of the reference data. As they mainly rely on laser scans from a driving vehicle, the reference data is typically compromised by accumulated registration errors and measurement noise. This is in contrast to synthetic imagery, as used in MPI Sintel, where the completely deterministic generation process ensures sub-pixel accurate ground truth. Given the relatively short stereo baseline of 0.5 meters, a disparity error of 3 pixels in image space will quickly result in large depth errors in object space. However, the published benchmark fills a gap concerning realistic test data for displacement field estimation with individually moving objects.

Parameter Training and Sensitivity

Section 5.3.2 picks up the simple training strategy that has been applied to the optical flow model before. As evidenced by the sensitivity plots and the quantitative evaluation, it yields an appropriate parametrization of the model weights. A meaningful adaption of parameter values to training data is an important step in the development of vision algorithms. However, for the tasks addressed in this thesis substantial gains in performance can be expected from better models. The methodology can be developed towards convex objective functions, promising gains from more rigorous optimization. Parameter tuning largely depends on the quality and the amount of available training data, which is scarce in the domain of motion estimation as compared to applications like object recognition, which are successfully tackled by deep learning [LeCun et al., 2015].

Ablation Studies

The detailed ablation studies presented in Section 5.3.3 show the value of the individual terms of the proposed scene flow model. During inference, the dense census features are the most important observations. This result confirms the intuition that the particle-based refinement is primarily guided by the dense similarity measure. However, in combination with the sparse features, the overall error decreases slightly. This second group of observations is of critical importance during initialization as described in Section 4.2.4. Establishing semi dense displacement fields in the image domain, these initial observations are the result of a complete matching strategy tailored to the sought entity and contain more consolidated information than image-based similarity measures.

Comparison to Related Methods

The quantitative comparison of the proposed approach to the state-of-the-art in Section 5.3.3 confirms its competitive ability to produce accurate results. On KITTI'12, it is ranked within the top ten methods. OSF does not reach the performance of methods exploiting epipolar matching between subsequent frames. However, motion stereo relies on the very strong constraint of static scenes. It typically holds for stereo matching of simultaneously captured images. While this assumption is helpful in reconstruction, it cannot be generalized to optical flow estimation without a loss of generality. Consequently, there are no results of motion stereo methods submitted to the optical flow challenge of KITTI'15. It can be expected that individually moving objects will not be recovered correctly. The other group of approaches outperforming OSF processes image sequences of more than two subsequent frames. Thus, they are able to consistently track surface elements throughout a number of frames and impose additional smoothness constraints in the temporal domain. Considering the overarching objective of comprehensive scene understanding, this direction of research is promising and verified by very good results. The proposed method can conceptually be extended to process longer image sequences. Estimation of the background shape and motion would then correspond to the task of visual SLAM. An optimal trajectory of the camera can be retrieved using temporal filtering techniques, which would also be useful to extend object detection by a tracking component.

On the recent KITTI'15 scene flow benchmark, the proposed approach yields state-of-the-art performance. However, many of the related methods have not been submitted yet. One important observation is the significant gap in performance between classical variational methods and discrete optimization. The latter group of approaches yields significantly lower error rates. This result is consistent with the conclusions drawn from related work in Section 3.4. The differential formulation of variational approaches is better suited to slowly moving articulated surfaces. The test data investigated in Section 5.3, which can be considered realistic examples in the context of autonomous navigation, pose different challenges. Large parts of the observed scenes can be ap-

proximated by piece-wise planar surfaces, imposing strong regularization on reconstruction. Large displacements of individual objects are more accurately retrieved by three-dimensional motion estimation as implemented in the proposed Object Scene Flow approach.

On the investigated data, OSF compares favorably with respect to the most closely related approach PRSF. While the comparison on KITTI'15 can be considered as biased due to different training data, reduced error rates can also be shown on the annotated entities of KITTI'12.

It is also interesting to inspect the effect of joint shape and motion estimation on the initial values. Table 5.7 evaluates optical flow results on KITTI'15. Object Scene Flow more than halves the error rate of DiscreteFlow by reducing the overall outlier percentage form 20% to 8%. The gain in performance is larger in background areas. A consistent, and heavily over-determined, rigid body transformation is able to explain most of the observed background motion. A pre-requisite for this strategy to work is a faithful reconstruction of the static parts of the scene. Disparities are initialized using SPS-Stereo, which, on its own, yields an error rate of 5%. OSF is not able to significantly improve this metric. However, the reconstruction in the reference frame is not affected by the additional constraints on object motion and provides the required input to successful scene flow estimation.

Considering optical flow estimation, recent variational approaches show competitive performance on KITTI'12 and a continuous refinement stage is an integral part of the most successful methods on Sintel. For the reasons discussed above, this trend does not generalize to 3D scene flow estimation. This domain is currently dominated by approaches with expressive prior assumptions of piece-wise planarity and rigidity, which lend themselves to discrete optimization.

6.3 Joint 3D Estimation of Vehicles and Scene Flow

Two metrics are investigated in Section 5.4.1 to evaluate the benefits of incorporating high-level prior information about the observed objects. The increased intersection-over-union measure confirms a better coverage of the detected objects. Averaged over all detected objects, the IOU metric is improved by 12%. This average result is influenced by a number of objects that cannot be reconstructed faithfully based on the employed active shape model. One example is discussed in the corresponding qualitative results. The coverage of a delivery van is compromised because it is erroneously decomposed into two smaller objects. Due to this effect, the overall error metrics consistently increase by a moderate amount.

These results provide a proof of concept of the model extension and strongly suggest the incorporation of a more flexible model. As long as the focus is on individually moving objects, the proposed motion-based segmentation is able to detect a reasonable percentage of relevant instances. On the investigated training data, it correctly initializes 60% of the annotated cars as object hypothe-

ses. By design, motion cues do not allow to retrieve objects with apparent motion close to the background.

One of the limiting factors considering the reconstruction of successfully detected objects is the ASM, which covers a broad range of passenger cars but cannot generalize to trucks and vans. To address this issue, the ASM could be trained on a more comprehensive set of input models. However, this strategy might deteriorate the expressiveness of the prior information and require more shape parameters to access the additional information in the ASM. A more promising extension would be the integration of a class-specific object detector. It could help to prune false-positive object hypotheses, as encountered in Section 5.4.2 and provide a more reliable initialization. This idea could be extended by training the detector on different types of objects.

The experimental evaluation also confirms that, like the basic scene flow approach, the model-based extension depends on the quality of the initialization. One representative failure case is shown in Figure 5.17. Thus far, a relatively simple initialization procedure yields promising results. An object detector providing a rough segmentation of the detections would help to further improve the approximate values.

Chapter 7

Conclusions and Outlook

Following the goal of two- and three-dimensional motion estimation, the present thesis addressed intertwined issues at several levels of perception. Starting from optical flow computation, which is a classic instance of early vision, we discussed graphical models and discrete inference techniques for successive levels of semantic interpretation. This section draws conclusions from the presented investigations and provides and outlook to the most promising future directions.

Discrete optimization has been shown to yield promising solutions at all levels of perception examined in this work. The presented approaches provide accurate models and suitable approximate optimization strategies that allow for the image-based estimation of two- and three-dimensional motion fields. While the computational complexity of the resulting algorithms allows for tractable inference, it still exceeds the requirements of real-time applications. This issue needs to be addressed in the future to increase the applicability of the presented ideas as building blocks of more comprehensive vision systems.

An appropriate mathematical model based on reasonable assumptions is the core of DiscreteFlow and Object Scene Flow. The proposed energy functions for both tasks are designed to faithfully model discontinuities as encountered in realistic applications. Truncated penalty functions allow for gross errors in the observations and abrupt changes in the estimated motion field. For the proposed optical flow model, it was shown how they additionally reduce the computational complexity of discrete inference on large, unordered label sets. Consequently, the objective functions are highly non-convex and their optimization can only be conducted approximately. The usefulness of the developed methods is shown as they perform favorably or comparable with the state-of-the-art on several challenging benchmark data sets. However, the success of the optimization significantly depends on the initialization of the models. A closer coupling of the initialization stage and the energy minimization should be investigated. This could allow for improved results

after less iterations of the optimization and further increase the computational efficiency of the methods.

Concerning dense optical flow estimation, the developed method is competitive in combination with a variational refinement step. So far, this post-processing is necessary to interpolate and refine the flow field. It should be revised in future work as the discussed experimental results hint at potential for further gains in performance. One possible approach would be an iterative extension of the discrete inference scheme that could refine the discrete proposals to sub-pixel accuracy and propagate reliable estimates into ambiguous image regions. In the current version, the integer-valued flow field, which is the result of discrete inference, outperforms DeepMatching on the investigated validation set of Sintel. It provides accurate initial values for continuous refinement and yields valuable motion observations forming the basis for the scene flow approach.

Direct access to the three-dimensional structure and motion of the observed scene is provided by Object Scene Flow. At an appropriate scale, many relevant scenes can be decomposed into a finite set of individually moving objects. This basic assumption of the developed scene flow model holds for the processed outdoor image sequences and leads to improved results on the challenging KITTI'15 data set. The efficient parametrization of the problem maintains the necessary flexibility of the model while imposing valuable constraints. The method is also shown to handle the static scenes of KITTI'12 successfully. They can be regarded a less complex special case of the problem at hand. Under these circumstances, the performance of the proposed methodology is surpassed by methods exploiting additional information like the epipolar constraint or longer image sequences.

The developed scene flow model provides the flexibility to introduce high-level object knowledge. Based on an active shape model the extended approach is able to retrieve parametrized reconstructions of distinct objects. This information is valuable for the addressed applications but the reduced generality of the extended approach leads to a moderate increase of error rates. The presented results can be considered a first proof of concept. This line of research can be advanced further by incorporating high-level observations like semantic segmentation, and higher-order interactions between the inferred objects. Such more complex relations in the graphical model increase the demand for advanced optimization techniques.

The presented experimental investigations demonstrate the value of the discussed contributions. Resuming this line of work and addressing the outlined challenges may provide valuable components of comprehensive scene understanding in the future.

Bibliography

[Achanta et al., 2012] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. and Susstrunk, S., 2012. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 34(11), pp. 2274–2282.

[Anandan, 1989] Anandan, P., 1989. A computational framework and an algorithm for the measurement of visual motion. International Journal of Computer Vision (IJCV) 2(3), pp. 283–310.

[Bailer et al., 2015] Bailer, C., Taetz, B. and Stricker, D., 2015. Flow Fields: Dense Correspondence Fields for Highly Accurate Large Displacement Optical Flow Estimation. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV), pp. 4015–4023.

[Bao et al., 2013] Bao, S., Chandraker, M., Lin, Y. and Savarese, S., 2013. Dense Object Reconstruction with Semantic Priors. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1264–1271.

[Barber, 2012] Barber, D., 2012. Bayesian Reasoning and Machine Learning. Cambridge University Press.

[Barnes et al., 2010] Barnes, C., Shechtman, E., Goldman, D. B. and Finkelstein, A., 2010. The Generalized Patchmatch Correspondence Algorithm. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 29–43.

[Basha et al., 2013] Basha, T., Moses, Y. and Kiryati, N., 2013. Multi-view Scene Flow Estimation: A View Centered Variational Approach. International Journal of Computer Vision (IJCV) 101(1), pp. 6–21.

[Bishop, 2006] Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer-Verlag New York.

[Black and Anandan, 1991] Black, M. J. and Anandan, P., 1991. Robust dynamic motion estimation over time. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 296–302.

[Black and Anandan, 1993] Black, M. J. and Anandan, P., 1993. A framework for the robust estimation of optical flow. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV), pp. 231–236.

[Bleyer et al., 2011] Bleyer, M., Rother, C., Kohli, P., Scharstein, D. and Sinha, S., 2011. Object stereo - joint stereo matching and object segmentation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3081–3088.

[Bouguet, 2000] Bouguet, J., 2000. Pyramidal implementation of the lucas kanade feature tracker. Intel.

[Boykov et al., 2001] Boykov, Y., Veksler, O. and Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 23(11), pp. 1222 –1239.

[Braun et al., 1995] Braun, C., Kolbe, T. H., Lang, F., Schickler, W., Steinhage, V., Cremers, A. B., Förstner, W. and Plümer, L., 1995. Models for Photogrammetric Building Reconstruction. In: Computer & Graphics, Vol. 19, pp. 109–118.

[Braux-Zin et al., 2013] Braux-Zin, J., Dupont, R. and Bartoli, A., 2013. A General Dense Image Matching Framework Combining Direct and Feature-Based Costs. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV), pp. 185–192.

[Brox and Malik, 2011] Brox, T. and Malik, J., 2011. Large Displacement Optical Flow: Descriptor Matching in Variational Motion Estimation. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 33, pp. 500–513.

[Brox et al., 2004] Brox, T., Bruhn, A., Papenberg, N. and Weickert, J., 2004. High Accuracy Optical Flow Estimation Based on a Theory for Warping. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 25–36.

[Butler et al., 2012] Butler, D. J., Wulff, J., Stanley, G. B. and Black, M. J., 2012. A Naturalistic Open Source Movie for Optical Flow Evaluation. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 611–625.

[Cech et al., 2011] Cech, J., Sanchez-Riera, J. and Horaud, R. P., 2011. Scene flow estimation by growing correspondence seeds. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3129 – 3136.

[Chen and Koltun, 2014] Chen, Q. and Koltun, V., 2014. Fast MRF Optimization with Application to Depth Reconstruction. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3914–3921.

[Cootes et al., 1995] Cootes, T. F., Taylor, C. J., Cooper, D. H. and Graham, J., 1995. Active Shape Models - Their Training and Application. Computer Vision and Image Understanding (CVIU) 61(1), pp. 38–59.

[Dame et al., 2013] Dame, A., Prisacariu, V., Ren, C. and Reid, I., 2013. Dense Reconstruction Using 3D Object Shape Priors. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1288–1295.

[Debevec et al., 1996] Debevec, P. E., Taylor, C. J. and Malik, J., 1996. Modeling and Rendering Architecture from Photographs: A hybrid geometry-and image-based approach. In: ACM Trans. on Graphics (SIGGRAPH), pp. 11–20.

[Demetz et al., 2014] Demetz, O., Stoll, M., Volz, S., Weickert, J. and Bruhn, A., 2014. Learning Brightness Transfer Functions for the Joint Recovery of I llumination Changes and Optical Flow. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 455–471.

[Dollár and Zitnick, 2013] Dollár, P. and Zitnick, C. L., 2013. Structured Forests for Fast Edge Detection. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV), pp. 1841–1848.

[Drayer and Brox, 2015] Drayer, B. and Brox, T., 2015. Combinatorial Regularization of Descriptor Matching for Optical Flow Estimation. In: Proc. of the British Machine Vision Conf. (BMVC).

[Farnebäck, 2003] Farnebäck, G., 2003. Two-frame motion estimation based on polynomial expansion. In: Image Analysis, Springer, pp. 363–370.

[Felzenszwalb and Huttenlocher, 2006] Felzenszwalb, P. and Huttenlocher, D., 2006. Efficient Belief Propagation for Early Vision. International Journal of Computer Vision (IJCV) 70(1), pp. 41–54.

[Fischer et al., 2015] Fischer, P., Dosovitskiy, A., Ilg, E., Häusser, P., Hazirbas, C., Smagt, V. G. P., Cremers, D. and Brox, T., 2015. FlowNet: Learning Optical Flow with Convolutional Networks. arXiv.org.

[Förstner, 2013] Förstner, W., 2013. Graphical Models in Geodesy and Photogrammetry. PFG Photogrammetrie, Fernerkundung, Geoinformation 2013(4), pp. 255–267.

[Fortun et al., 2014] Fortun, D., Bouthemy, P. and Kervrann, C., 2014. Aggregation of local parametric candidates with exemplar-based occlusion handling for optical flow. arXiv.org.

[Geiger and Wang, 2015] Geiger, A. and Wang, C., 2015. Joint 3D Object and Layout Inference from a Single RGB-D Image. In: J. Gall, P. Gehler and B. Leibe (eds), Pattern Recognition, Lecture Notes in Computer Science, Vol. 9358, Springer International Publishing, pp. 183–195.

[Geiger et al., 2012] Geiger, A., Lenz, P. and Urtasun, R., 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3354–3361.

[Geiger et al., 2011] Geiger, A., Ziegler, J. and Stiller, C., 2011. StereoScan: Dense 3D Reconstruction in Real-time. In: Proc. IEEE Intelligent Vehicles Symposium (IV), pp. 963–968.

[Güney and Geiger, 2015] Güney, F. and Geiger, A., 2015. Displets: Resolving Stereo Ambiguities using Object Knowledge. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 4165–4175.

[Herbst et al., 2013] Herbst, E., Ren, X. and Fox, D., 2013. RGB-D flow: Dense 3D motion estimation using color and depth. In: Proc. IEEE International Conf. on Robotics and Automation (ICRA), pp. 2276–2282.

[Hirschmüller, 2008] Hirschmüller, H., 2008. Stereo Processing by Semiglobal Matching and Mutual Information. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 30(2), pp. 328–341.

[Hooke and Jeeves, 1961] Hooke, R. and Jeeves, T. A., 1961. "Direct Search" Solution of Numerical and Statistical Problems. J. ACM 8(2), pp. 212–229.

[Horn, 1986] Horn, B., 1986. Robot vision. MIT Press.

[Horn and Schunck, 1981] Horn, B. K. P. and Schunck, B. G., 1981. Determining Optical Flow. Artificial Intelligence (AI) 17(1-3), pp. 185–203.

[Hornacek et al., 2014] Hornacek, M., Fitzgibbon, A. and Rother, C., 2014. SphereFlow: 6 DoF Scene Flow from RGB-D Pairs. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3526–3533.

[Huguet and Devernay, 2007] Huguet, F. and Devernay, F., 2007. A Variational Method for Scene Flow Estimation from Stereo Sequences. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV), pp. 1–7.

[Ihler and McAllester, 2009] Ihler, A. T. and McAllester, D. A., 2009. Particle Belief Propagation. In: Conference on Artificial Intelligence and Statistics (AISTATS), pp. 256–263.

[Isard and MacCormick, 2006] Isard, M. and MacCormick, J., 2006. Dense motion and disparity estimation via loopy belief propagation. In: Proc. of the Asian Conf. on Computer Vision (ACCV), Springer, pp. 32–41.

[Kappes et al., 2015] Kappes, J. H., Andres, B., Hamprecht, F. A., Schnörr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B. X., Kröger, T., Lellmann, J., Komodakis, N., Savchynskyy, B. and Rother, C., 2015. A Comparative Study of Modern Inference Techniques for Structured Discrete Energy Minimization Problems. International Journal of Computer Vision (IJCV) pp. 1–30.

[Kennedy and Taylor, 2014] Kennedy, R. and Taylor, C. J., 2014. Optical Flow with Geometric Occlusion Estimation and Fusion of Multiple Frames. In: Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR).

[Koch, 1999] Koch, K.-R., 1999. Parameter estimation and hypothesis testing in linear models. Springer Science & Business Media.

[Kolmogorov, 2006] Kolmogorov, V., 2006. Convergent Tree-Reweighted Message Passing for Energy Minimization. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 28(10), pp. 1568–1583.

[Kondermann et al., 2015] Kondermann, D., Nair, R., Meister, S., Mischler, W., Güssefeld, B., Honauer, K., Hofmann, S., Brenner, C. and Jähne, B., 2015. Stereo Ground Truth with Error Bars. In: Proc. of the Asian Conf. on Computer Vision (ACCV), pp. 595–610.

[Kumar and Hebert, 2006] Kumar, S. and Hebert, M., 2006. Discriminative random fields. International Journal of Computer Vision (IJCV) 68(2), pp. 179–201.

[Lafferty et al., 2001] Lafferty, J., McCallum, A. and Pereira, F. C., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the 18th International Conference on Machine Learning, pp. 282–289.

[LeCun et al., 2015] LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. Nature 521(7553), pp. 436–444.

[Leibe et al., 2006] Leibe, B., Cornelis, N., Cornelis, K. and Van Gool, L., 2006. Integrating Recognition and Reconstruction for Cognitive Traffic Scene Analysis from a Moving Vehicle. In: Pattern Recognition, Lecture Notes in Computer Science, Vol. 4174, Springer, pp. 192–201.

[Lempitsky et al., 2008] Lempitsky, V. S., Roth, S. and Rother, C., 2008. FusionFlow: Discrete-continuous optimization for optical flow estimation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1–8.

[Leordeanu and Hebert, 2008] Leordeanu, M. and Hebert, M., 2008. Smoothing-based Optimization. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1–8.

[Li, 2009] Li, S. Z., 2009. Markov Random Field Modeling in Image Analysis. Springer-Verlag London.

[Li et al., 2015] Li, Y., Min, D., Brown, M. S., Do, M. N. and Lu, J., 2015. SPM-BP: Sped-up PatchMatch Belief Propagation for Continuous MRFs. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV), pp. 4006–4014.

[Lowe, 2004] Lowe, D. G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision (IJCV) 60(2), pp. 91–110.

[McGlone, 2004] McGlone, J. C. (ed.), 2004. Manual of Photogrammetry Fifth Edition. American Society for Photogrammetry and Remote Sensing.

[Menze and Geiger, 2015] Menze, M. and Geiger, A., 2015. Object Scene Flow for Autonomous Vehicles. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3061–3070.

[Menze et al., 2015a] Menze, M., Heipke, C. and Geiger, A., 2015a. Discrete Optimization for Optical Flow. In: J. Gall, P. Gehler and B. Leibe (eds), Pattern Recognition, Lecture Notes in Computer Science, Vol. 9358, Springer International Publishing, pp. 16–28.

[Menze et al., 2015b] Menze, M., Heipke, C. and Geiger, A., 2015b. Joint 3D Estimation of Vehicles and Scene Flow. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences II-3/W5, pp. 427–434.

[Mozerov, 2013] Mozerov, M., 2013. Constrained Optical Flow Estimation as a Matching Problem. IEEE Trans. on Image Processing (TIP) 22(5), pp. 2044–2055.

[Nir et al., 2008] Nir, T., Bruckstein, A. M. and Kimmel, R., 2008. Over-Parameterized Variational Optical Flow. International Journal of Computer Vision (IJCV) 76(2), pp. 205–216.

[Poggio et al., 1985] Poggio, T., Torre, V. and Koch, C., 1985. Computational vision and regularization theory. Nature 317, pp. 314–319.

[Pons et al., 2007] Pons, J.-P., Keriven, R. and Faugeras, O., 2007. Multi-View Stereo Reconstruction and Scene Flow Estimation with a Global Image-Based Matching Score. International Journal of Computer Vision (IJCV) 72(2), pp. 179–193.

[Prince, 2012] Prince, S., 2012. Computer Vision: Models, Learning and Inference. Cambridge University Press.

[Quiroga et al., 2014] Quiroga, J., Brox, T., Devernay, F. and Crowley, J. L., 2014. Dense Semi-Rigid Scene Flow Estimation from RGB-D images. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 567–582.

[Rabe et al., 2010] Rabe, C., Mueller, T., Wedel, A. and Franke, U., 2010. Dense, Robust, and Accurate Motion Field Estimation from Stereo Image Sequences in Real-Time. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 582–595.

[Ranftl et al., 2014] Ranftl, R., Bredies, K. and Pock, T., 2014. Non-local Total Generalized Variation for Optical Flow Estimation. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 439–454.

[Revaud et al., 2015] Revaud, J., Weinzaepfel, P., Harchaoui, Z. and Schmid, C., 2015. EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

[Rother et al., 2007] Rother, C., Kolmogorov, V., Lempitsky, V. and Szummer, M., 2007. Optimizing Binary MRFs via Extended Roof Duality. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1–8.

[Stein, 2004] Stein, F., 2004. Efficient Computation of Optical Flow Using the Census Transform. In: C. Rasmussen, H. Bülthoff, B. Schölkopf and M. Giese (eds), Pattern Recognition, Lecture Notes in Computer Science, Vol. 3175, Springer Berlin Heidelberg, pp. 79–86.

[Sun et al., 2013] Sun, D., Roth, S. and Black, M. J., 2013. A Quantitative Analysis of Current Practices in Optical Flow Estimation and the Principles Behind Them. International Journal of Computer Vision (IJCV) 106(2), pp. 115–137.

[Szeliski, 2011] Szeliski, R., 2011. Computer Vision - Algorithms and Applications. Texts in Computer Science, Springer.

[Szeliski et al., 2008] Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M. and Rother, C., 2008. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 30(6), pp. 1068–1080.

[Thomas et al., 2007] Thomas, A., Ferrari, V., Leibe, B., Tuytelaars, T. and Van Gool, L., 2007. Depth-from-recognition: Inferring meta-data by cognitive feedback. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV), IEEE, pp. 1–8.

[Timofte and Van Gool, 2015] Timofte, R. and Van Gool, L., 2015. Sparse Flow: Sparse Matching for Small to Large Displacement Optical Flow. In: Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1100–1106.

[Tola et al., 2010] Tola, E., Lepetit, V. and Fua, P., 2010. DAISY: An efficient dense descriptor applied to wide baseline stereo. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 32(5), pp. 815–830.

[Valgaerts et al., 2010] Valgaerts, L., Bruhn, A., Zimmer, H., Weickert, J., Stoll, C. and Theobalt, C., 2010. Joint Estimation of Motion, Structure and Geometry from Stereo Sequences. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 568–581.

[Vedula et al., 1999] Vedula, S., Baker, S., Rander, P., Collins, R. and Kanade, T., 1999. Three-dimensional scene flow. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 722–729.

[Vedula et al., 2005] Vedula, S., Rander, P., Collins, R. and Kanade, T., 2005. Three-dimensional scene flow. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 27(3), pp. 475–480.

[Vogel et al., 2013a] Vogel, C., Roth, S. and Schindler, K., 2013a. An Evaluation of Data Costs for Optical Flow. In: Pattern Recognition, Lecture Notes in Computer Science, Vol. 8142, pp. 343–353.

[Vogel et al., 2014] Vogel, C., Roth, S. and Schindler, K., 2014. View-Consistent 3D Scene Flow Estimation over Multiple Frames. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 263–278.

[Vogel et al., 2011] Vogel, C., Schindler, K. and Roth, S., 2011. 3D scene flow estimation with a rigid motion prior. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV), pp. 1291–1298.

[Vogel et al., 2013b] Vogel, C., Schindler, K. and Roth, S., 2013b. Piecewise Rigid Scene Flow. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV), pp. 1377–1384.

[Vogel et al., 2015] Vogel, C., Schindler, K. and Roth, S., 2015. 3D Scene Flow Estimation with a Piecewise Rigid Scene Model. International Journal of Computer Vision (IJCV) 115(1), pp. 1–28.

[Wainwright et al., 2005] Wainwright, M. J., Jaakkola, T. S. and Willsky, A. S., 2005. Map estimation via agreement on trees: message-passing and linear programming. IEEE Trans. on Information Theory 51(11), pp. 3697–3717.

[Wedel and Cremers, 2011] Wedel, A. and Cremers, D., 2011. Stereo Scene Flow for 3D Motion Analysis. Springer-Verlag London.

[Wedel et al., 2008] Wedel, A., Rabe, C., Vaudrey, T., Brox, T., Franke, U. and Cremers, D., 2008. Efficient Dense Scene Flow from Sparse or Dense Stereo Data. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 739–751.

[Wei et al., 2014] Wei, D., Liu, C. and Freeman, W., 2014. A Data-driven Regularization Model for Stereo and Flow. In: Proc. of the International Conference on 3D Vision (3DV), pp. 277–284.

[Weinzaepfel et al., 2013] Weinzaepfel, P., Revaud, J., Harchaoui, Z. and Schmid, C., 2013. DeepFlow: Large Displacement Optical Flow with Deep Matching. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV), pp. 1385–1392.

[Weiss and Freeman, 2001] Weiss, Y. and Freeman, W. T., 2001. On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. IEEE Trans. on Information Theory 47(2), pp. 736–744.

[Wulff and Black, 2015] Wulff, J. and Black, M. J., 2015. Efficient Sparse-to-Dense Optical Flow Estimation using a Learned Basis and Layers. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

[Yamaguchi et al., 2012] Yamaguchi, K., Hazan, T., McAllester, D. and Urtasun, R., 2012. Continuous markov random fields for robust stereo estimation. In: Proc. of the European Conf. on Computer Vision (ECCV), Springer, pp. 45–58.

[Yamaguchi et al., 2013] Yamaguchi, K., McAllester, D. and Urtasun, R., 2013. Robust Monocular Epipolar Flow Estimation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1862–1869.

[Yamaguchi et al., 2014] Yamaguchi, K., McAllester, D. and Urtasun, R., 2014. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 756–771.

[Yang and Li, 2015] Yang, J. and Li, H., 2015. Dense, Accurate Optical Flow Estimation with Piecewise Parametric Model. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1019–1027.

[Zabih and Woodfill, 1994] Zabih, R. and Woodfill, J., 1994. Non-parametric local transforms for computing visual correspondence. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 151–158.

[Zach et al., 2007] Zach, C., Pock, T. and Bischof, H., 2007. A Duality Based Approach for Realtime TV-L1 Optical Flow. In: Pattern Recognition, Lecture Notes in Computer Science, Vol. 4713, pp. 214–223.

[Zia et al., 2013a] Zia, M., Stark, M. and Schindler, K., 2013a. Explicit Occlusion Modeling for 3D Object Class Representations. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3326–3333.

[Zia et al., 2015] Zia, M., Stark, M. and Schindler, K., 2015. Towards Scene Understanding with Detailed 3D Object Representations. International Journal of Computer Vision (IJCV) 112(2), pp. 188–203.

[Zia et al., 2013b] Zia, M., Stark, M., Schiele, B. and Schindler, K., 2013b. Detailed 3D Representations for Object Recognition and Modeling. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 35(11), pp. 2608–2623.

[Zimmer et al., 2011] Zimmer, H., Bruhn, A. and Weickert, J., 2011. Optic Flow in Harmony. International Journal of Computer Vision (IJCV) 93(3), pp. 368–388.

Appendix A

Plane Parametrization

The scene flow approach introduced in Section 4.2 relies on a parametrization of superpixels in terms of planes in three-dimensional space. For several steps it is important to re-parametrize these planes in terms of disparites in image space.

The basic parametrization of a plane in 3D in terms of the Hesse normal form reads as

$$0 = \mathbf{X}\,\mathbf{n}_0 - D$$

with D encoding the distance of the plane from the origin. $\mathbf{X}$ is a 3D point on the plane and $\mathbf{n}_0$ is the unit normal vector of the plane. A compact representation is thus given by the three vector (n_X, n_Y, n_Z) which is the unit normal vector divided by the distance from the origin

$$1 = X\,\frac{n_X^0}{D} + Y\,\frac{n_Y^0}{D} + Z\,\frac{n_Z^0}{D} = X\,n_X + Y\,n_Y + Z\,n_Z$$

An equivalent representation in terms of disparities d is given by

$$d = a\,x + b\,y + c$$

with image coordinates x, y and disparity plane parameters a, b, c.

For the normal case of stereo photogrammetry with undistorted images, given the interior orientation parameters and the stereo baseline, the relation between object points $(X, Y, Z)^T$ and image coordinates and disparity $(x, y, d)^T$ is defined as

$$\begin{aligned} x &= \frac{f \cdot X}{Z} + c_x, \quad X = (x - c_x) \cdot \frac{Z}{f} \\ y &= \frac{f \cdot Y}{Z} + c_y, \quad Y = (y - c_y) \cdot \frac{Z}{f} \\ d &= \frac{f \cdot L}{Z}, \qquad Z = \frac{f \cdot L}{d} \end{aligned}$$

Here, L denotes the length of the stereo baseline, f is the principal distance and c_x, c_y are the image coordinates of the principal point.

Inserting image coordinates into Equation A yields

$$\begin{aligned} 1 &= (x - c_x) \frac{n_X \cdot Z}{f} + (y - c_y) \frac{n_Y \cdot Z}{f} + Z \cdot n_Z \\ \frac{1}{Z} &= (x - c_x) \frac{n_X}{f} + (y - c_y) \frac{n_Y}{f} + n_Z \end{aligned}$$

and finally

$$\begin{aligned} d &= \frac{f \cdot L}{Z} \\ d &= (x - c_x) \cdot L \cdot n_X + (y - c_y) \cdot L \cdot n_Y + f \cdot L \cdot n_Z \\ d &= L \cdot n_X \cdot x + L \cdot n_Y \cdot y + f \cdot L \cdot n_Z - c_x \cdot L \cdot n_X - c_y \cdot L \cdot n_Y \end{aligned}$$

The conversion from 3D plane parameters to a disparity plane is given by

$$\begin{aligned} a &= L\, n_X \\ b &= L\, n_Y \\ c &= f\, L\, n_Z - c_x\, L\, n_X - c_y\, L\, n_Y \end{aligned}$$

In the same set-up, the elements of the 3D normal can be computed as

$$\begin{aligned} n_X &= \frac{a}{L} \\ n_Y &= \frac{b}{L} \\ n_Z &= \frac{c + c_x\, a + c_y\, b}{f \cdot L} \end{aligned}$$

Appendix B

Additional Quantitative Results

For completeness, Table B.1 provides the results for the non-occluded parts of the scene flow test data of KITTI'15. Table B.2 contains the results of the ablation studies explained in Section 5.3.3 on non-occluded pixels. The tables provide error rates for disparities in the reference frame (D1) and the target frame (D2), optical flow (Fl) and scene flow (SF) averaged over the non-occluded annotations. For each modality, the outlier percentage is reported for the background region (bg), all foreground objects (fg) as well as all annotated pixels in the image (bg+fg).

	D1			D2			Fl			SF			Run-
	bg	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	time[s]
OSF	**3.78**	**10.25**	**4.85**	**4.05**	**14.23**	**5.87**	**4.02**	**18.00**	**6.56**	**5.20**	**23.39**	**8.46**	390
OSF	4.14	11.12	5.29	4.49	16.33	6.61	4.21	18.65	6.83	5.52	24.58	8.93	50 min
PRSF	4.41	13.09	5.84	6.35	16.12	8.10	6.94	23.64	9.97	8.35	28.45	11.95	150
SGM+SF	4.75	14.22	6.31	8.34	18.71	10.20	13.36	25.21	15.51	15.28	32.33	18.33	>600
SGM+C+NL	4.75	14.22	6.31	15.72	20.79	16.63	23.03	41.92	26.46	26.22	48.61	30.23	270
SGM+LDOF	4.75	14.22	6.31	17.08	18.66	17.36	30.41	31.34	30.58	33.00	39.44	34.15	86
HWBSF	18.76	21.14	19.16	23.92	21.88	23.55	30.13	31.33	30.35	35.65	38.40	36.14	420
GCSF	11.24	26.26	13.72	21.88	31.66	23.63	38.12	41.53	38.74	43.64	55.02	45.68	3
VSF	26.38	19.88	25.31	52.30	40.83	50.24	41.15	44.16	41.70	61.14	60.38	61.00	>600

Table B.1: *Outlier percentage on the non-occluded parts of the scene flow test data of KITTI'15.*

	D1			D2			Fl			SF		
	bg	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*
Data (SPSS+SpF)	3.6	10.7	4.6	4.5	19.1	6.7	4.3	23.6	7.7	5.6	27.0	9.2
Data (Census)	4.3	11.4	5.4	5.2	13.0	6.4	4.4	17.9	6.7	5.9	22.8	8.8
Data (All)	3.6	10.0	4.6	4.5	11.7	5.6	4.0	16.3	6.1	5.2	20.7	7.8
Data (All) + Smooth (Boundary)	**3.2**	**7.4**	**3.9**	**4.1**	9.4	4.9	3.8	15.1	5.7	4.9	17.9	7.1
Data (All) + Smooth (Normal)	3.6	9.7	4.5	4.4	11.4	5.5	3.9	16.0	6.0	5.1	20.3	7.7
Data (All) + Smooth (Object)	3.8	11.0	4.8	4.6	11.9	5.7	3.9	16.0	6.1	5.2	20.4	7.8
Data (SPSS+SpF) + Smooth (All)	4.4	13.8	5.8	5.4	14.9	6.9	4.8	16.8	6.9	6.3	20.6	8.7
Data (Census) + Smooth (All)	3.3	7.7	4.0	4.2	**8.8**	4.9	3.8	14.4	5.7	5.0	16.7	6.9
Data (All) + Smooth (All)	**3.2**	7.9	**3.9**	**4.1**	9.2	**4.8**	**3.7**	**14.1**	**5.5**	**4.8**	**16.5**	**6.8**

Table B.2: *Outlier percentage for distinct model components on the validation portion of KITTI'15.*

Curriculum Vitae

Personal Information

Name	Menze, Till Moritz
Date of Birth	November 1982

Work Experience

September 2010 – January 2016	Institute of Photogrammetry and GeoInformation Leibniz Universität Hannover *Research Assistant*
March 2014 – August 2014	Max Planck Institute for Intelligent Systems, Tübingen Perceiving Systems Department *Six-month Doctoral Scholarship*
June 2008 – September 2010	Bertrandt Ingenieurbüro GmbH, Tappenbeck *Test Engineer*

Education

October 2007 – March 2008	Volkswagen AG, Wolfsburg Development Department *Diploma Thesis*
October 2003 – May 2008	Course Geodesy and GeoInformatics Leibniz Universität Hannover *Diploma*
August 2002 – May 2003	Civil Service
August 1995 – June 2002	High School

Acknowledgements

Since a dissertation is the result of valuable advice, intense collaboration and substantial support, I would like to thank a number of people who contributed to the success of this work.

I would like to express my deep gratitude to my supervisor Prof. Dr.-Ing. habil. Christian Heipke for very productive discussions and careful guidance during my studies and my time with the Institute of Photogrammetry and GeoInformation (IPI) in Hanover. Furthermore, I would like to thank Prof. Dr. Konrad Schindler and apl. Prof. Dr.-Ing. Claus Brenner for acting as referees and Prof. Dr.-Ing. habil. Monika Sester for chairing the examination committee.

I thank all my friends and colleagues at the IPI who made the institute a pleasant and productive place to be at all times. In particular, I would like to thank Joachim for his hospitality and very insightful discussions, and Tobias for sharing the office as well as all the smaller and larger challenges of a PhD student. In Hanover, I enjoyed breakfast and football with my fellow students Alexander, Nico, Sebastian and Jens-André, whom I thank for their support and advice.

During my doctoral studies, I had the opportunity to spend six month with the amazing Perceiving Systems Department at the Max Planck Institute for Intelligent Systems in Tübingen. I would like to thank all department members and collaborators for sharing their expertise and for a great time with the group. My special thanks go to Dr. Andreas Geiger. I am very grateful for his support and the successful collaboration, which reached far beyond my stay in Tübingen.

Finally, I would like to express my thanks to my family and friends for their understanding and loving support. I especially thank my girlfriend Julia for her unconditional love and encouragement throughout eventful times.