RESEARCH

Vladislav Golyanik

Robust Methods for Dense Monocular Non-Rigid 3D Reconstruction and Alignment of Point Clouds

Springer Vieweg

Robust Methods for Dense Monocular Non-Rigid 3D Reconstruction and Alignment of Point Clouds

Vladislav Golyanik

Robust Methods for Dense Monocular Non-Rigid 3D Reconstruction and Alignment of Point Clouds

Springer Vieweg

Vladislav Golyanik
Computer Graphics D4
Max Planck Institute for Informatics
Saarbruecken, Germany

Vom Fachbereich Informatik der Technischen Universität Kaiserslautern zur Verleihung des akademischen Grades Doktor der Ingenieurwissenschaften (Dr.-Ing.) genehmigte Dissertation

Datum der wissenschaftlichen Aussprache: 20. November 2019
Dekan: Prof. Dr. Stefan Deßloch
Vorsitzender der Promotionskommission: Prof. Dr. Hans Hagen
Erster Berichterstatter: Prof. Dr. Didier Stricker
Zweiter Berichterstatter: Prof. Dr. Reinhard Koch
Dritter Berichterstatter: Prof. Dr. Antonio Agudo

Technische Universität Kaiserslautern, 2019
D386

ISBN 978-3-658-30566-6 ISBN 978-3-658-30567-3 (eBook)
https://doi.org/10.1007/978-3-658-30567-3

This Springer Vieweg imprint is published by the registered company Springer Fachmedien Wiesbaden GmbH part of Springer Nature.
The registered company address is: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Acknowledgments

I am deeply and sincerely thankful to my doctoral supervisor Didier Stricker, the head of the Augmented Vision Department at the German Research Centre for Artificial Intelligence (DFKI) and a professor at the University of Kaiserslautern. He has always supported me in my research intentions, appreciated my independent work and strategically contributed to my development. He has also encouraged me to complete a research internship in a company.

The AV group has celebrated its tenth anniversary in August 2018. Its warm and collaborative environment has, without doubt, contributed to the accomplishment of this work. Therefore, I would like to acknowledge all colleagues and visiting researches of the AV group, who have been working in the AV group from 2014 until 2018, including my office mates, Adytia Tewari, Jilliam Maria Diaz Barros and Sarvenaz Salehi. A special thank goes to Leivy Michelly Kaul and Keonna Cunningham, for helping me to struggle through the office jungles and plenty of formalities.

I am deeply thankful to Bertram Taetz who was my postdoctoral colleague during my first months at DFKI. It was a unique experience to work with him which has effected my mindset in a way that I began to perceive computer vision in the continuous domain and from the perspective of applied mathematics. Since that time, I highly appreciate working with excellent mathematicians. I am deeply thankful to Gerd Reis who has often been helping me with improving paper drafts. Discussions with him have always been inspiring and rewarding.

During my sabbatical at NVIDIA Research in Santa Clara, I had the great fortune to work with Robert Maier who was an intern at NVIDIA Research at the same time as me, and Matthias Nießner who has been a visiting assistant professor at Stanford Univerity back then. I am thankful to my supervisors Kihwan Kim and Jan Kautz from NVIDIA Research for giving me the opportunity to work on a challenging topic and sharing the ideas. Moreover, working with them has helped me to strengthen my leadership qualities.

I am deeply thankful to all my further co-authors: Gabriele Bleser, Torben Fetzer, Soshi Shimada, Mitra Narsi, Oliver Wasenmüller, Jan C. Peters, Aman Shankar Mathur, Sk Aziz Ali, Mohammad Dawud Ansari, Tomonari Yosida, Kiran Varanasi, André Jonas and Christian Theobalt. I have enjoyed working with all of you, and I have learned from all of you. Furthermore, I would like to thank Stella Graßhof, Hanno Ackermann, Bodo Rosenhahn, Jörn Ostermann, Antonio Agudo, Francesc Moreno-Noguer, Daniel Cremers, Willi Frieden, Gabriele Steidl, Yongzhi Su, Alain Pagani, Norbert Schmitz, Marco Paladini, Deqing Sun and Thomas Breuel for collaborations, fruitful discussions and advice.

Last but not least, I am genuinely and sincerely grateful to my parents and my family. Without their patience and support, endurance and apprehension this work would have been barely possible. At all times, the music of Ludwig van Beethoven was helping me to make decisions and served as an inexhaustible source of patience, persistence and inspiration.

Vladislav Golyanik

Contents

List of Figures

List of Tables

Abstract

An accurate acquisition and processing of 3D point cloud data is an active research area in computer vision encompassing various unsolved problems. The thesis at hand addresses the jointly studied domains of dense non-rigid 3D reconstruction from monocular image sequences and point set registration under rigid as well as non-rigid transformations. Monocular non-rigid 3D reconstruction, which is in the focus of this work — known as non-rigid structure from motion (NRSfM) — relies on weak assumptions about the feasible deformation modes imposed on top of the motion and deformation cues. NRSfM and non-rigid point set registration are highly ill-posed problems in the sense of Hadamard.

The proposed dense NRSfM methods address the broad range of research questions including occlusion handling, scalability, interactive yet accurate processing as well as dense structure compression. For the occlusion handling and dealing with inaccurate point tracks, we propose a shape prior obtained on-the-fly and a new spatial regulariser — the coherency term. We also introduce a new model for NRSfM, which allows representing the recovered structure compactly.

The proposed point set registration methods aim at the enhanced registration accuracy for noisy data and samples with clustered outliers. For that reason, we embed prior correspondences into probabilistic point set registration and introduce a previously unexplored class of methods relying on principles of particle dynamics with simulated gravitational forces.

The thorough experimental evaluation confirms the efficiency and high accuracy of the proposed methods as well as the validity of the new ideas. By using the new principles, we advance the state of the art in dense monocular non-rigid 3D reconstruction and alignment of noisy point sets. Applications of the proposed NRSfM methods include (but are not limited to) 3D recovery and analysis of human and animal faces, endoscopic scenes and various other deformable surfaces. The proposed point set registration methods can be applied in robotics, automotive driving, face and shape recognition, and other areas. Apart from the abovementioned applications, we show how both method classes can be used for human appearance transfer, multiframe scene flow estimation from RGB-D as well as monocular image sequences. The developed methods offer numerous avenues for further investigation.

Zusammenfassung

Die genaue Eingabe und Verarbeitung von 3D Punktwolken ist ein aktives Forschungsfeld im maschinellen Sehen, das viele ungelöste Probleme umfasst. Die vorliegende Doktorarbeit befasst sich mit den in Zusammenwirkung erforschten Bereichen der dichten nicht-starren 3D Rekonstruktion aus monokularen Bildsequenzen, mit sowohl starren als auch nicht-starren Punktwolkenregistrierung. Die monokulare 3D Rekonstruktion, die im Fokus dieser Arbeit steht und die als nichtrigide Struktur aus Bewegung (NRSaB) bekannt ist, wertet, einerseits, Bewegungen und Deformationen aus, und, andererseits, verknüpft diese mit den zusätzlichen Annahmen und Vorwissen über die Szene und die Art der zulässigen Zustände.

Die eingeführten Verfahren zur dichten NRSaB gehen auf mehrere offene Fragen ein, und zwar auf die Behandlung von Verdeckungen, die Skalierbarkeit und die Anpassungsfähigkeit auf unterschiedliche Szenarien und Größenordnungen der Szenen, interaktive und präzise Verarbeitung, sowie die Komprimierung dichter 3D Geometrie. Zwecks der Behandlung von Verdeckungen und fehlerhafter Punktkorrespondenzen werden Verfahren mit dem am Anfang einer Bildsequenz gewonnenen Formvorwissen sowie einem neuen räumlichen Kohärenz Regularisierer vorgestellt. Darüber hinaus, leiten wir ein neues NRSaB Verfahren her, das die gewonnene Geometrie in eine kompakte Repräsentation überführt.

Die entwickelten Verfahren zur Punktwolkenregistrierung verfolgen das Ziel, verrauschte und partielle Eingabedaten mit höherer Präzision zu verarbeiten als die Vorgängermethoden. Dementsprechend schlagen wir vor, die im Vorfeld hergestellten Korrespondenzen ins probabilistische Framework für die Punktwolkenregistrierung zu integrieren, und, zweitens, präsentieren wir eine neue und bisher unerforschte Verfahrensklasse, welche die Teilchenbewegungen unter virtuellen Schwerkräften simuliert.

Durch gründliche und zahlreiche Experimente ist es uns gelungen, die Geltung der neuen Ideen sowie die Präzision und Robustheit der entwickelten Verfahren

zu bestätigen. Dank der neuen Prinzipien und Verfahren waren wir imstande, den Stand der Technik in beiden Bereichen der monokularen nicht-rigiden 3D Rekonstruktion sowie Punktwolkenregistrierung zu verbessern. Zu den Anwendungen neuer NRSaB Verfahren zählen die 3D Rekonstruktion von Menschen, Tieren und endoskopischer Aufnahmen sowie die Erfassung dünner Strukturen unterschiedlicher Herkunft. Die entwickelten Verfahren zur Punktwolkenregistrierung können unter anderem in Robotik, selbstfahrender Fahrzeugtechnik sowie Gesichts- und Formerkennung angewendet werden. Neben der erwähnten Gebieten wird in dieser Arbeit gezeigt, wie die beiden Verfahrensklassen zwecks der Übertragung des äußeren Erscheinungsbildes von Menschen sowie der Berechnung vom Szenenfluss aus Tiefenkamerabildern und monokularen Bildern angepasst werden können. Ferner, bieten die entwickelten Verfahren verschiedene Wege und Möglichkeiten zur Verbesserung und Weiterentwicklung, auf die am Schluss eingegangen wird.

1 Introduction

ONE of the objectives of computer vision is an accurate sensing of the real world and robust processing of the acquired data. Along with the material properties, knowledge of 3D geometry is a key component of complete scene description. 3D machine perception is the foundation for multiple applications which involve scene replication, scene understanding, localisation as well as a realistic superimposition of virtual contents, among others.

There are multiple sensors which come into question while designing a vision-based system including time-of-flight cameras, stereo cameras, lidars and sonars. A lightweight alternative to those is a single monocular camera. The advantages of a monocular camera are different designs and form-factors, affordability, relatively low electric energy consumption but also pervasiveness in modern electronic devices and wide acceptance in society. There are monocular cameras embedded in augmented reality glasses, helmets, mobile phones and tablets. Monocular cameras are central components in endoscopic surgery systems, surveillance systems, person identification systems, unmanned aerial and underwater vehicles, mobile robots, rovers for planetary explorations and autonomous cars. Thus, methods using monocular cameras for 3D sensing are of high relevance in a broad variety of systems and applications. Moreover, techniques for processing and analysis of the recovered raw 3D representations — often point sets and point clouds — are increasingly gaining relevance.

3D reconstruction is an extensively studied inverse problem in computer vision consisting in the recovery of the depth dimension of a scene lost during the imaging (together, the scene geometry), from single or multiple views. *Point set registration* is a computer vision problem of recovery the transformations aligning one or multiple point sets (raw 3D representations or 3D reconstructions) into a common coordinate frame or deforming the point sets so that their appearances match.

Depending on the available input and in many practical situations, 3D reconstruction can also be an ill-posed problem in the sense of Hadamard. Thus, 3D reconstruction from a single image is ill-posed, as multiple 3D scenes can result in the same 2D image. Additional prior knowledge is required to disambiguate the reconstruction such as a known object class, symmetry prior or a geometric prior. Starting from two views, additional constraints can be used ranging from epipolar geometry prior and trilinear constraints to the consistency constraints over multiple

V. Golyanik, *Robust Methods for Dense Monocular Non-Rigid 3D Reconstruction and Alignment of Point Clouds*,
https://doi.org/10.1007/978-3-658-30567-3_1

views. Moreover, the rigidity assumption disambiguates the problem well, and impressive results were achieved in 3D reconstruction under the rigidity assumption, both from multiple views and the sequences of monocular views.

If several different views of the same scene at the same time frame are available, the subclass of the techniques is referred to as *multiple view geometry reconstruction*. In contrast, if multiple monocular views of the scene over several time frames are available, the subclass of the techniques is called *structure from motion* (SfM). In a general context, the input data corresponds to an image sequence, and some methods operate on a set of tracked points over the input views. The difference between multiple view geometry and structure from motion under rigidity is often subtle. In many cases, the techniques can be applied interchangeably, though the information about whether the cameras are static or moving can be advantageous (*e.g.,* motion blur prone to a moving camera can be accounted for). Compared to multi-view reconstruction, SfM often assumes smaller frame-to-frame displacements, as those observed in a video sequence. Video sequences also allow for stronger priors such as temporal smoothness.

Apart from predominantly static environments and sceneries, our surroundings are inhabited by living species including ourselves which move and deform. Besides, there are rigidly moving manmade instruments and products violating the assumption about the static and conserved ambience. Thus, capturing and processing of dynamic scenes is a core capability of robust vision-based systems.

1.1 Monocular Non-Rigid Dynamic 3D Reconstruction

The situation changes considerably, if the rigidity assumption does not hold anymore, *i.e.,* the scene undergoes non-rigid deformations. In the case of multiple views, the observations are captured at the same moment of time, and the geometry is still related by spatial rigidity between the views. If captured at different time frames and with different camera poses, the scene is observed in different states and the temporal rigidity does not hold anymore. The class of methods assuming non-rigid scenes over a temporal sequence of views is specified as *non-rigid structure from motion* (NRSfM). In NRSfM, the camera is moving, whereas the scene is moving and deforming. Similar to rigid SfM, the input of NRSfM is a set of tracked points over the available views.

Though NRSfM is a highly ill-posed problem which is sometimes said to be equivalent to the reconstruction from a single view, additional constraints can help to disambiguate it. *Real-world objects do not deform arbitrarily and rather follow a* *certain deformation pattern*. The deformation pattern is also often associated with periodicity, which implies that the scene states are repeated in a temporally-disjoint manner. Moreover, an average or middle state can be distinguished among all observed states. Additionally, it is more probable that the states in neighbouring frames are more similar than states in frames that are temporally further apart.

There is a substantial difference between single view rigid reconstruction and NRSfM which has crystallised out. In the single view rigid reconstruction, it is valid and common to assume a specific object class, and supervised learning methods are often applied. NRSfM, in contrast, assumes that no prior shape information about the observed scene is available, and solely relies on motion and deformation cues to obtain 3D surface reconstructions from monocular image sequences. This makes NRSfM capable of handling equally well — depending on the accuracy of correspondences — thin surfaces of different kinds (flags, sails, *etc.*), human and animal faces, clothes and body tissues in medical contexts.

Several new method classes have emerged which constrain the context of NRSfM, such as those assuming an accurate reconstruction of at least one of the frames in the sequence (template-based methods) and those assuming a pre-defined deformation model but different material properties.

1.2 Point Set Registration

When 3D surface recovery is complete, there are multiple ways how the dynamic reconstruction can be processed and analysed. One of the essential pre-processing steps is changing the reference frame or pose of the reconstruction for the further comparison, deformation transfer or recognition. This operation can be performed with *rigid point set registration* if the orientation of the reference frame or object is known. The comparison and deformation transfer can be accomplished with *non-rigid point set registration*.

The objective of point set registration is to align two or several point sets, *i.e.,* to recover a transformation which registers a *template* point set to an unaltered *reference*. A point set is an unordered set of coordinates (2D or 3D), with no further information available. As a representation of a shape, it can contain noise and clustered outliers, and some parts can be missing. Point set registration should not be confused with mesh registration methods (meshes are more complete shape representations consisting of points, triangles, normals *etc.*). 3D reconstructions obtained with NRSfM often represent point sets.

In the rigid case, the transformation is parametrised by the variables of rigid body motion with six degrees of freedom (three for rotation and three for translation). During a rigid transformation, no deformation is happening, and all distances between the points are preserved. Transformation of every point is given by the same rotation and translation. In the non-rigid case, due to deformations, distances between the points are not preserved, and the transformation is described by a general per-point displacement field. As monocular deformable reconstruction, *non-rigid point set registration relies on prior knowledge that real-world objects and scenes do not deform arbitrarily but rather follow some deformation rules and patterns.* One of the most commonly used and reasonable constraints in non-rigid point set registration is the topology the preserving constraint which states that point topology must be preserved despite the distances between the points are changing. It prevents intersections between the displacements, and, as a consequence, self-intersections of the surfaces represented by the points sets (though, point sets can also represent volumetric structures). Similarly to NRSfM, non-rigid point set registration is an ill-posed problem in the sense of Hadamard.

Despite the progress in point set registration which enabled various practical applications, one of the central research questions in point set registration remains improvement of the robustness to noise and disturbing effects in the data (missing parts and clustered outliers). Moreover, processing of large point sets is an ever-relevant problem (in other words, point sets containing one-two orders of magnitude more points than what is considered as a standard nowadays; the contemporary standard in NRSfM is around $30k$ points). It is addressed with faster hardware, parallelisation as well as data structures for acceleration. In contrast to methods for processing of synthetic 3D data (computer graphics), methods for processing of raw sensor inputs have to cope and consider noise and incompleteness of the data.

3D reconstruction and point set registration exhibit similarities. Thus, common types of assumptions and constraints can be applied to disambiguate them (*e.g.,* rigidity assumption and shape priors), and for handling non-rigid deformations, regularisation of displacement fields is required. In this thesis, the study of 3D reconstruction and point set registration is conducted jointly. As will be shown throughout the thesis, both related research fields facilitate and enrich each other with ideas. Point set registration provides tools for 3D reconstruction (algorithmic and evaluation tools), 3D reconstruction provides data for optimal evaluation of point set registration, and multiple concepts can be borrowed from one field to another one (regularisation of the displacement fields).

1.3 Scope of the Thesis

This thesis focuses on robust methods for dense monocular non-rigid 3D reconstruction from uncalibrated views and alignment of point cloud data.

The methods for dense monocular non-rigid 3D reconstruction should not assume that the calibration is known, though if it is known, the algorithms could optionally use the calibration parameters. Moreover, the new approaches should reconstruct the scene per-point densely, and, optionally, allow sparse reconstruction. The requirements to the new methods include robustness to self- and external occlusions, scalability, higher accuracy and lower runtime compared to the existing methods. Some of the requirements are not necessary facilitating towards the other ones, *i.e.,* it is more challenging to develop a scalable, accurate and fast method at the same time. Efficient NRSfM methods in conjunction with robust methods for dense correspondences would enable new applications based on commodity hardware.

Thanks to the point set registration, the reconstructed scenes can be compared to some reference data or the recovered appearance can be transferred to some other representations usable in different application scenarios. Thus, both method classes can be used in a single 3D reconstruction and processing pipeline.

There is also another reason to study the fields of monocular 3D reconstruction and point cloud alignment jointly. Even though the underlying methods pursue different goals and assume different input data, both fields are still related to each other, so that cross-fertilisation and exploitation of synergies is possible. Thus, non-rigid registration can help in the joint evaluation of NRSfM and correspondence establishment approaches, as will be shown in §5. Moreover, due to the handling of deformable structures in both algorithm classes, we proposed a new spatial regulariser (coherency constraint, §6) for NRSfM which was previously used exclusively in non-rigid point set registration.

The work at hand was also inspired by the maturing research area of augmented reality. Augmented reality is an interdisciplinary research field on the intersection of computer vision, computer graphics and hardware systems (which include material science, physics, mechatronics and electronics). The goal of augmented reality is to extend and enhance the perceived reality through useful virtual contents. Virtual contents should be realistic and indistinguishable from the real ones. Along with the realistic rendering, accurate placement of virtual contents is one of the quality factors. The acquisition of geometry of deformable objects with efficient methods for processing of the reconstructions is highly relevant for augmented reality as well. Both method classes addressed in this thesis — NRSfM and point set registration — can be used in augmented reality systems in a pipeline for 3D reconstruction and data processing with a single monocular moving camera.

1.4 Overview of the Contributions

The primary subject of the dissertation is dynamic 3D reconstruction of non-rigidly deforming scenes from monocular image sequences as well as processing of point sets. The main considered algorithm classes are non-rigid structure from motion (NRSfM) and point set registration (PSR). NRSfM is a highly ill-posed inverse problem. The input of NRSfM is a set of point tracks over several unsynchronised and uncalibrated views, and the objective is the recovery of the observed non-rigid 3D geometry. Thus, NRSfM uses motion and deformation cues as well as additional weak prior assumptions about the type of valid deformations for 3D recovery. In PSR, the inputs are two point sets with a different number of points, and the objective is the alignment of those into a common reference frame (in the rigid case) or the recovery of the displacements and correspondences non-rigidly aligning the inputs (in the non-rigid case). The two fields were studied jointly and complemented each other.

NRSfM and Monocular Surface Recovery

In the field of NRSfM, the thesis features the following contributions:

- First, a new dense variational NRSfM technique for handling large occlusions and inaccuracies in the data was proposed — *Shape Prior Variational Approach* (SPVA). SPVA estimates a shape prior from several first unoccluded frames of the sequence on-the-fly and guides the reconstruction by the occlusion tensor. The occlusion tensor is computed from the initial dense flow fields and indicates occlusion probabilities for every frame. The method allows for the reconstruction of scenes where large occlusions are expected (*e.g.*, in medical contexts). The method is parallelisable and is implemented on a GPU. The experimental results show the state-of-the-art accuracy on challenging sequences for which a shape prior can be obtained.

- Second, a new method with an intrinsic dynamic shape prior for 3D reconstruction and compression of sequences with temporally-disjoint rigidity is introduced. Temporally-disjoint rigidity occurs in most real video sequences, *i.e.,* the phenomenon of state reoccurrence. The repeating states can be separated by an arbitrary number of other states and can reappear in different poses. Our *Dynamic Shape Prior Reconstruction* (DSPR) approach takes advantage of temporally-disjoint rigidity and allows for dense reconstructions with low latencies. Experiments demonstrate that DSPR can operate on inaccurate correspondences.

- Third, a new spatial regulariser — the *coherency term* — for dense NRSfM is proposed which has allowed handling of large occlusions without a shape prior. The coherency term was adopted from the motion coherence theory. Before, the coherency term was used in non-rigid point set registration. We have shown how to minimise energy with the coherency term in the context of NRSfM.
- Fourth, we have addressed the problem of structure compressibility in the sense of data compression theory in the context of NRSfM and proposed a new *High-Dimensional Space Model* (HDSM) for NRSfM. In HDSM, non-rigid geometry in 3D is encoded as multiple projections of a single high dimensional structure onto different 3D subspaces. The proposed representation in combination with the factorisation-based (decoupled) formulation for camera pose and shape recovery allows compressing the structure in the high dimensional space. The resulting method encompassing handling of inaccurate point tracks with the coherency term and structure compression is known as *Lifted Coherent Depth Fields* (L-CDF).
- Fifth, we propose a new fast technique for dense NRSfM — *Accelerated Metric Projections* (AMP) — which allows to factorise dense batches of point tracks in seconds on a CPU. At the moment of publication, AMP was the fastest dense NRSfM method delivering high reconstruction accuracy. We have shown in AMP how to minimise a quadratic function on a set of orthonormal matrices using an efficient semidefinite programming solver. The method allows an arbitrary reshuffling of the per-frame measurements which can be advantageous in the cases when temporal information cannot be maintained.
- Sixth, we have addressed the question of scalability in the context of NRSfM. The core characteristic of the resulting robust NRSfM technique — *Scalable Monocular Surface Reconstruction* (SMSR) — is the steady high accuracy across a large variety of dense and sparse datasets with reasonable runtime and linear scalability w.r.t. the number of points. In SMSR, the camera pose is updated with singular value thresholding and proximal gradient techniques, whereas the surface is estimated by alternating direction method of multipliers.
- Seventh, we found a new way to regress non-rigid geometry with a trained encoder-decoder deep neural network. In the *Hybrid Deformation Model Network* (HDM-Net), the deformation model is learned from synthetic data in a supervised manner. Among contributions of HDM-Net is a new way to perform a convolution on a point set instead of a volumetric representation, an isometric loss and a contour loss. Moreover, the inference of a surface with over $5k$ points takes around 5 ms. Results on real images demonstrated the potential of the proposed architecture for augmented reality applications.