Fortschritt-Berichte VDI

VDI

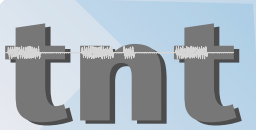
Reihe 10

Informatik/
Kommunikation

Timo von Marcard, M. Sc.,
Hemmingen

Nr. 866

Human Motion Capture with Sparse Inertial Sensors and Video



Institut für Informationsverarbeitung
www.tnt.uni-hannover.de

Human Motion Capture with Sparse Inertial Sensors and Video

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades
Doktor-Ingenieur
(abgekürzt: Dr.-Ing.)
genehmigte Dissertation

von

Timo von Marcard, M. Sc.
geboren am 31. März 1984 in Gießen, Deutschland

2019

1. Referent:	Prof. Dr.-Ing. Bodo Rosenhahn
2. Referent:	Prof. Dr.-Ing. Marcus Magnor
Vorsitzender:	Prof. Dr.-Ing. Jörn Ostermann

Tag der Promotion: 16. Oktober 2019

Fortschritt-Berichte VDI

Reihe 10

Informatik/
Kommunikation

Timo von Marcard, M. Sc.,
Hemmingen

Nr. 866

Human Motion Capture with Sparse Inertial Sensors and Video

Institut für Informationsverarbeitung
www.tnt.uni-hannover.de

von Marcard, Timo
Human Motion Capture with Sparse Inertial Sensors and Video
Fortschr.-Ber. VDI Reihe 10 Nr. 866. Düsseldorf: VDI Verlag 2019.
124 Seiten, 47 Bilder, 14 Tabellen.
ISBN 978-3-18-386610-6, ISSN 0178-9627,
€ 48,00/VDI-Mitgliederpreis € 43,20.

Keywords: Human Pose Estimation – Inertial Sensors – Video – Non-static Camera – Model-based Optimization – Sparse Sensors

This thesis explores approaches to capture human motions with a small number of sensors. In the first part of this thesis an approach is presented that reconstructs the body pose from only six inertial sensors. Instead of relying on pre-recorded motion databases, a global optimization problem is solved to maximize the consistency of measurements and model over an entire recording sequence. The second part of this thesis deals with a hybrid approach to fuse visual information from a single hand-held camera with inertial sensor data. First, a discrete optimization problem is solved to automatically associate people detections in the video with inertial sensor data. Then, a global optimization problem is formulated to combine visual and inertial information. The proposed approach enables capturing of multiple interacting people and works even if many more people are visible in the camera image. In addition, systematic inertial sensor errors can be compensated, leading to a substantial increase in accuracy.

Bibliographische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet unter www.dnb.de abrufbar.

Bibliographic information published by the Deutsche Bibliothek
(German National Library)
The Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliographie (German National Bibliography); detailed bibliographic data is available via Internet at www.dnb.de.

Als Manuskript gedruckt. Printed in Germany.
ISSN 0178-9627
ISBN 978-3-18-386610-6

Acknowledgments

This thesis was written in the course of my activity as a scientific research assistant at the *Institut für Informationsverarbeitung* of the Leibniz University Hannover.

First of all, I would like to thank my doctoral advisor Prof. Dr.-Ing. Bodo Rosenhahn for the opportunity to work in his group and for his excellent supervision, support, and encouragement. Also, many thanks to him and Prof. Dr.-Ing. Jörn Ostermann for the great working conditions at the institute.

I am also very grateful to Dr.-Ing. Gerard Pons-Moll for many inspiring discussions and his efforts to make me grow as a researcher. His ideas and experience contributed a lot to make this thesis a success.

I thank Prof. Dr.-Ing. Marcus Magnor for being the second examiner of this thesis and Prof. Dr.-Ing. Jörn Ostermann for being the chair of the defense committee.

Special thanks go to all my colleagues at the *Institut für Informationsverarbeitung*. I had a fantastic time at the institute. Specifically, I would like to thank the administrative staff for all the support in technical and administrative matters. A special thanks also goes to my office mate and friend Dipl.-Math. Roberto Henschel for substantially improving my math skills and for making our office such a great place to waste time.

Finally, I would like to thank my family. Every single day, my wife Kathrin and my children Leni and Minna show me that there are more important things in life than work. Also, I would like to thank my parents for their unconditional support.

Contents

Acronyms

2D	*two-dimensional*
3D	*three-dimensional*
3DPW	*three-dimensional poses in the wild*
CNN	*Convolutional Neural Network*
DoF	*Degrees of Freedom*
GPS	*Global Positioning System*
IMU	*Inertial Measurement Unit*
MEMS	*Micro-Electro-Mechanical Systems*
MoCap	*Motion Capture*
MPJAE	*Mean Per Joint Angular Error*
MPJPE	*Mean Per Joint Position Error*
NN	*Neural Network*
RMS	*Root Mean Square*
SIP	*Sparse Inertial Poser*
SMPL	*Skinned Multi-Person Linear*
SO(3)	*Special Orthogonal Group of dimension three*
SE(3)	*Special Euclidean Group of dimension three*
VIP	*Video Inertial Poser*

Notation

Numbers and Arrays

a	A scalar (integer or real)
$\mathbf{a}$	A vector
$\mathbf{A}$	A matrix
$\mathbf{A}^T$	Transpose of matrix $\mathbf{A}$
$\langle \mathbf{a}, \mathbf{b} \rangle$	Scalar product of $\mathbf{a}$ and $\mathbf{b}$
$\mathbf{a} \times \mathbf{b}$	Cross product of $\mathbf{a}$ and $\mathbf{b}$
$\frac{\partial y}{\partial x}$	Partial derivative of y with respect to x
$\frac{\partial f}{\partial \mathbf{x}}$	Jacobian matrix $\mathbf{J} \in \mathbb{R}^{m \times n}$ of $f \colon \mathbb{R}^n \to \mathbb{R}^m$
$\bar{\mathbf{a}}$	Homogeneous representation of vector $\mathbf{a}$
$\tilde{\mathbf{a}}$	Estimate of vector quantity $\mathbf{a}$
$\|\|\mathbf{a}\|\|$	L^2-norm of $\mathbf{a}$
$\det(\mathbf{A})$	Determinant of $\mathbf{A}$
$\mathbf{I}_n$	Identity matrix with n rows and n columns
$\mathbf{0}_{n \times m}$	Zero matrix with n rows and m columns

Symbols

θ	A scalar angle
$\mathbf{x}$	Pose vector parametrizing a kinematic chain
δ	Gradient or perturbation of a pose vector
$\mathcal{C}$	A kinematic chain
$\mathrm{Pa}_{\mathcal{C}}(b)$	Parent joints of segment b in $\mathcal{C}$

$\mathbf{R}$	A rotation matrix $\mathbf{R} \in SO(3)$
$\mathbf{M}$	A matrix representing a rigid body motion $\mathbf{M} \in SE(3)$
$\pi(\mathbf{a})$	Projection of a 3D point $\mathbf{a}$ to homogeneous pixel coordinates
$\mathbf{K}$	Matrix of camera intrinsics
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance Σ
$\oplus$	Perturbation-operator ($\oplus \colon G \times \mathfrak{g} \to G$)

Exponential Mapping

G	A Lie group
$\mathfrak{g}$	A Lie algebra
$\mathbf{G}_i$	Generator matrix associated to dimension i of a Lie algebra
$\widehat{\mathbf{a}}$ or $\mathbf{a}^\wedge$	Wedge-operator to construct a Lie algebra element from a coordinate vector $\mathbf{a}$
$\mathbf{A}^\vee$	Vee-operator to obtain coordinate vector from an element of a Lie algebra
exp	Matrix exponential to map from a Lie Algebra element to a Lie Group element
log	Logarithm to map from a Lie Group element to a Lie Algebra element

Sets and Graphs

$\mathbb{R}$	The set of real numbers
$\mathcal{G}$	A graph
$v \in \mathcal{V}$	A vertex v in a vertex set $\mathcal{V}$
$e \in \mathcal{E}$	An edge e in an edge set $\mathcal{E}$
c	A cost variable
$\mathcal{F}$	Feasibility set
$l \in \mathcal{L}$	A label l in a label set $\mathcal{L}$
x	A binary indicator variable
$\mathcal{H}$	An assignment hypothesis

Abstract

This dissertation explores approaches to capture human motions with a small number of sensors. Conventional methods either use a large number of static cameras, which severely limits the recording space, or a high number of body-worn inertial sensors, which is intrusive and only accurate for short time periods.

The first part of this thesis presents an approach that reconstructs the body pose from only 6 inertial sensors. Conventionally, up to 17 sensors are needed to cover all degrees of freedom of the body. Since fewer sensors inevitably lead to ambiguities, previous approaches estimate the missing information from pre-recorded motion databases. In contrast, in this work a model-based approach is proposed. More specifically, a global optimization problem is solved to maximize the consistency of measurements and model over an entire recording sequence. A key observation is that the kinematic constraints imposed by a statistical human body model constrain the search space significantly. This allows to utilize the acceleration data of inertial sensors to compensate for the missing sensor information. The performance of the method is demonstrated in challenging outdoor scenarios and accuracy is evaluated on two benchmark datasets.

The second part of this thesis deals with a hybrid approach to fuse visual information from a single hand-held camera with inertial sensor data. This approach combines the advantages of both sensor modalities. It enables capturing multiple interacting people and works even if many more people are visible in the camera image. In addition, systematic errors of the inertial sensors can be compensated, leading to a substantial increase in accuracy. In order to fuse the sensor modalities, visual information from the camera has to be associated to inertial sensor data. This is done automatically by formulating a discrete graph labeling problem. Subsequently, all sensor information of an entire tracking sequence is transformed into a global model-based optimization problem, which reconstructs body poses, camera pose and sensor errors. In several experiments accuracy is evaluated quantitatively and qualitatively. The combination of a single hand-held camera and body-worn inertial sensors enables motion capture in new complex settings. Using the approach a variety of motions are recorded, e.g. during shopping in a crowded pedestrian zone or during a bus ride. These recordings are composed into a novel dataset, which was made publicly available for research purposes.

Keywords: Human Pose Estimation, Inertial Sensors, Video, Non-static Camera, Model-based Optimization, Sparse Sensors

Kurzfassung

Diese Dissertation untersucht Ansätze zur Erfassung menschlicher Bewegungen mit wenigen Sensoren. Herkömmliche Verfahren verwenden entweder eine große Anzahl an statischen Kameras, was den Aufnahmebereich stark einschränkt, oder eine hohe Anzahl am Körper getragenen Inertialsensoren, was als unangenehm empfunden wird und nur für kurze Zeiträume präzise funktioniert.

Im ersten Teil dieser Arbeit wird ein Ansatz vorgestellt, der die Körperhaltung aus den Messdaten von nur 6 Inertialsensoren rekonstruiert. Üblicherweise werden bis zu 17 Sensoren benötigt um alle Freiheitsgrade des Körpers abzudecken. Da weniger Sensoren zwangsläufig zu Uneindeutigkeiten führen, werden in bisherigen Ansätzen die fehlenden Informationen aus zuvor aufgenommenen Bewegungsdatenbanken geschätzt. Im Gegensatz dazu wird ein modellbasierter, generativer Ansatz entwickelt. Sämtliche Messwerte eine Aufnahmesequenz werden in ein globales Optimierungsproblem überführt und die Konsistenz von Messdaten und Modell maximiert. Die modellierten kinematischen Einschränkungen des menschlichen Skelettes führen zu einer wesentlichen Eingrenzung des Suchraums und ermöglichen so die Beschleunigungsdaten der Inertialsensoren zur Kompensation der fehlenden Sensorinformationen heranzuziehen. Die Präzision des Ansatzes wird experimentell untersucht und durch Bewegungsrekonstruktionen aus anspruchsvollen Außenaufnahmen demonstriert.

Im zweiten Teil der Arbeit wird der vorhergehende Ansatz erweitert, um visuelle Informationen von einer in der Hand gehaltenen Smartphone-Kamera mit den Daten der Inertialsensoren zu fusionieren. Dieser Ansatz ermöglicht eine mobile Bewegungserfassung von mehreren interagierenden Personen und funktioniert selbst wenn im Kamerabild viele weitere Personen sichtbar sind. Zusätzlich können systematische Fehler der Inertialsensoren geschätzt werden, was zu einer erheblich präziseren Bewegungsschätzung führt. Um die verschiedenen Sensorinformation miteinander zu fusionieren, muss zunächst eine Zuordnung von Bildinformationen und Inertialsensordaten stattfinden. Diese Zuordnung wird zeitlich konsistent durch eine diskrete Optimierung mittels Graph-Labeling gelöst. Anschließend werden sämtliche Sensorinformationen einer gesamten Sequenz in ein globales Optimierungsproblem überführt und neben der Körperhaltung nun auch die relative Entfernung zur Kamera, die Kamerapose und Sensorfehler geschätzt. Die Präzision des Ansatzes wird in zahlreichen Experimenten evaluiert. Zusätzlich werden die im Rahmen der Arbeit aufgenommenen Bewegungssequenzen in Form eines neuartigen Datensatzes vorgestellt und für Forschungszwecke bereitgestellt. Die Kombination von Smartphone-Kamera und Inertialsensoren ermöglicht erstmalig eine mobile Bewegungserfassung von mehreren Personen, die auch für Alltagssituationen wie beispielsweise beim Einkaufen in einer belebten Fußgängerzone geeignet ist.

Schlagwörter: Erfassung menschlicher Bewegungen, Inertialsensoren, Video, bewegliche Kamera, modell-basierte Optimierung, wenige Sensoren