Fortschritt-Berichte VDI

VDI

Reihe 10

Informatik/
Kommunikation

Nr. 869

Bastian Wandt, M. Sc.,
Hannover

Human Pose Estimation from Monocular Images

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

Human Pose Estimation from Monocular Images

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 Herrn

Bastian Wandt, M. Sc.

geboren am 30. September 1984 in Peine

2020

Hauptreferent:	Prof. Dr.-Ing. Bodo Rosenhahn
Korreferent:	Prof. Dr. Ralph Ewerth
Vorsitzender:	Prof. Dr.-Ing. Markus Fidler

Tag der Promotion: 21. April 2020

Fortschritt-Berichte VDI

Reihe 10

Informatik/
Kommunikation

Bastian Wandt, M. Sc.,
Hannover

Nr. 869

Human Pose Estimation from Monocular Images

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

Wandt, Bastian
Human Pose Estimation from Monocular Images
Fortschr.-Ber. VDI Reihe 10 Nr. 869. Düsseldorf: VDI Verlag 2020.
130 Seiten, 47 Bilder, 8 Tabellen.
ISBN 978-3-18-386910-7, ISSN 0178-9627,
€ 52,00/VDI-Mitgliederpreis € 46,80.

Keywords: Human Pose Estimation – 3D Reconstruction – Monocular Cameras – Structure From Motion

This dissertation deals with the problem of capturing human motions and poses using a single camera. The first part of the thesis proposes two closely related approaches for the 3D reconstruction of human motions from image sequences. To resolve inherent ambiguities in monocular 3D reconstruction the main idea of this part is to exploit temporal properties of human motions in combination with a human body model learned from training data. The second part of the thesis tackles the problem of reconstructing a human pose from a single image. A human body model is learned by training a deep neural network that covers non-linearities and anthropometric constraints.

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 978-3-18-386910-7

ACKNOWLEDGEMENT

This thesis was written in the course of my activity as a research assistant at the *Insitut für Informationsverarbeitung* of the Leibniz Universität Hannover.
First, I would like to thank my doctoral advisor Prof. Dr.-Ing. Bodo Rosenhahn for giving me the opportunity to do my studies under his supervision. He always supported me in my research and gave me the freedom I needed to successfully finish my doctorate. I am especially thankful for long discussions about work and non-work related topics, which not only helped me grow as a researcher but also as a person. Also many thanks to him and Prof. Dr.-Ing. Jörn Ostermann for providing an outstanding research environment.
I also like to thank Prof. Dr. Ralph Ewerth for being the second examiner and Prof. Dr.-Ing. Markus Fidler for being the chair of the defense committee. I thank the whole committee for making it possible to defend my thesis during the COVID-19 pandemic.
During my time at the institute, I had many amazing colleagues who made the time at TNT unforgettable. Especially, I like to thank my office mate Petrissa Zell for many academic and private conversations and the fantastic work atmosphere in our office, Roberto Henschel for very detailed discussions and founding our consulting company together, and the TNT Alpine Team for 6 memorable trips to Austrian skiing resorts. Also, many thanks to the administrative staff for their support in technical and organizational tasks.
Finally, my special thanks go to my family for their support and encouragement during my studies.

CONTENTS

ACRONYMS

2D	two-dimensional
3D	three-dimensional
3DPE	3D Positioning Error
AUC	Area Under Curve
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
GT	Ground Truth
KCS	Kinematic Chain Space
NRSfM	Non-Rigid Structure from Motion
MoCap	Motion Capture
MPJPE	Mean Per Joint Positioning Error
PA	Procrustes Alignment
PCA	Principle Component Analysis
PCK	Percentage of Correctly Positioned Keypoints
ReLU	Rectified Linear Units
RepNet	Reprojection Network
SfM	Structure from Motion
SVD	Singular Value Decomposition
SVT	Singular Value Thresholding

Numbers and Arrays

a	A scalar (integer or real)
$\boldsymbol{a}$	A vector
$\boldsymbol{A}$	A matrix
$\boldsymbol{A}^T$	Transpose of matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	Inverse of quadratic matrix $\boldsymbol{A}$
$\boldsymbol{A}^+$	Moore-Penrose Pseudoinverse of matrix $\boldsymbol{A}$
$trace(\boldsymbol{A})$	Trace of matrix $\boldsymbol{A}$
$\|\boldsymbol{a}\|$	Vector norm of $\boldsymbol{a}$
$\|\boldsymbol{A}\|$	Matrix norm of $\boldsymbol{A}$
$\|\cdot\|_F$	Frobenius norm
$\|\cdot\|_*$	Nuclear norm
$\boldsymbol{I}_n$	Identity matrix of dimension $n \times n$
$\mathbf{0}$	Vector of all zeros
$\mathbf{1}$	Vector of all ones

Symbols

$\boldsymbol{X}$	Matrix $\boldsymbol{X} \in \mathbb{R}^{3\times j}$ describing a human pose with j joints
$\boldsymbol{X}_{2D}$	Backprojection of $\boldsymbol{X}$ to image coordinates
j	Number of joints
b	Number of bones
f	Number of frames
$\boldsymbol{K}$	Camera matrix containing intrinsic and extrinsic parameters
$\boldsymbol{R}$	Rotation matrix
$\boldsymbol{t}$	Translation vector
x, y, z	Coordinates in 3D space
u, v	Image coordinates

$\boldsymbol{W}$	Measurement matrix
$\boldsymbol{Q}$	Linear pose basis
$\boldsymbol{B}$	Bone matrix $\boldsymbol{X} \in \mathbb{R}^{3\times b}$ for b bones
$\boldsymbol{C}$	Linear mapping from 3D coordinates into the Kinematic Chain Space
$\boldsymbol{D}$	Linear mapping from the Kinematic Chain Space to 3D coordinates
$\boldsymbol{\Psi}$	Kinematic Chain Space matrix
$\mathcal{N}(\mu, \sigma)$	Gaussian distribution mean μ and standard deviation σ
$\mathcal{L}$	Loss function

ABSTRACT

This dissertation deals with the problem of capturing human motions and poses using a single camera. The constantly growing research field has various applications in medicine, sports, autonomous driving and human-robot interaction. In contrast to traditional multi-sensor solutions, this thesis presents different methods employing only a single consumer camera which opens up a wide variety of new applications.

The first part of the thesis proposes two closely related approaches for the 3D reconstruction of human motions from image sequences. Since images taken by a camera are projections of a 3D scene to a 2D plane, depth information is inevitably lost which gives infinitely many possible 3D reconstructions. To resolve these inherent ambiguities the main idea of this part is to exploit temporal properties of human motions in combination with a human body model learned from training data. The natural assumptions that human motions are smooth and bone lengths of one person do not change are formulated as smoothness constraints and a variance minimization. This approach gives pleasing results on several benchmark datasets. However, it is restricted to the motions used for training the human body model. Therefore, the body model is replaced by a more general kinematic chain model in a later step. This allows for the reconstruction of even subtle motion variations, e. g. limping instead of walking. The first approach accurately reconstructs everyday motions even with very noisy input data and occlusions but struggles to recover small variations in the motion. The second approach complements the first by also reconstructing these small deviations with only minor degradation in robustness to noise and occlusions.

The second part of the thesis tackles the problem of reconstructing a human pose from a single image. As shown in the first part, linear human body models give a strong prior for possible 3D reconstructions. However, the space of human poses is highly nonlinear. To this end, a human body model is learned by training a deep neural network that covers these non-linearities. Similar previous approaches train neural networks in a supervised manner using known 2D to 3D correspondences. Due to the limited amount of diverse training data these models tend to simply memorize specific poses in the training set and ignore rare poses. To avoid this a weakly supervised training scheme is proposed that learns a mapping between distributions of 2D and 3D poses. The consistency with the 2D observations is enforced by a novel reprojection layer which projects the estimated 3D poses back to 2D. The performance is shown on several benchmark datasets and achieves state-of-the-art

results, even compared to supervised approaches. The proposed method shows improved generalization to uncommon human poses and camera angles. Interestingly, applying this single image approach to sequences does not significantly increase the reconstruction errors.

Keywords – Human Motion Capture, Pose Estimation, Camera Estimation, Reprojection Error Optimization.

KURZFASSUNG

Diese Dissertation befasst sich mit der Erfassung menschlicher Bewegungen und Posen mit einer einzigen Kamera. Dieses ständig wachsende Forschungsgebiet hat verschiedene Anwendungen in der Medizin, im Sport, beim autonomen Fahren und bei der Mensch-Roboter Interaktion. Im Gegensatz zu traditionellen Multisensorlösungen werden in dieser Arbeit verschiedene Methoden vorgestellt, die nur eine einzige handelsübliche Kamera verwenden, was eine Vielzahl neuer Anwendungen eröffnet.

Der erste Teil der Arbeit präsentiert zwei eng miteinander verbundene Ansätze zur 3D-Rekonstruktion menschlicher Bewegungen aus Bildsequenzen. Da es sich bei den von einer Kamera aufgenommenen Bildern um Projektionen einer 3D-Szene auf eine 2D-Ebene handelt, gehen zwangsläufig Tiefeninformationen verloren, woraus sich unendlich viele mögliche 3D-Rekonstruktionen ergeben. Um diese inhärenten Mehrdeutigkeiten aufzulösen, besteht die Hauptidee dieses Teils darin, die zeitlichen Eigenschaften menschlicher Bewegungen in Kombination mit einem menschlichen Körpermodell zu nutzen, das aus den Trainingsdaten gelernt wurde. Die physikalisch gegebenen Annahmen, dass menschliche Bewegungen glatt sind und sich die Knochenlängen einer Person nicht ändern, werden als Glattheitsbeschränkungen und als eine Varianzminimierung formuliert. Dieser Ansatz führt zu guten Ergebnissen bei mehreren Benchmark-Datensätzen. Er ist jedoch auf die Bewegungen beschränkt, die für das Training des menschlichen Körpermodells verwendet werden. Daher wird das Körpermodell in einem späteren Schritt durch ein allgemeineres kinematisches Kettenmodell ersetzt. Dies ermöglicht die Rekonstruktion selbst subtiler Bewegungsvariationen, z.B. Humpeln statt Gehen. Der erste Ansatz rekonstruiert die alltäglichen Bewegungen selbst bei stark verrauschten Eingabedaten und Verdeckungen sehr genau, hat aber Schwierigkeiten, kleine Bewegungsvariationen zu rekonstruieren. Der zweite Ansatz ergänzt den ersten, indem er ebenfalls diese kleinen Abweichungen rekonstruiert, wobei die Robustheit gegenüber verrauschten Daten nur geringfügig beeinträchtigt wird.

Der zweite Teil der Arbeit befasst sich mit dem Problem der Rekonstruktion einer menschlichen Pose aus einem einzigen Bild. Wie im ersten Teil gezeigt wurde, schränken lineare Modelle des menschlichen Körpers mögliche 3D-Rekonstruktionen sehr gut ein. Allerdings ist der Raum der menschlichen Posen hochgradig nichtlinear. Zu diesem Zweck wird ein menschliches Körpermodell gelernt, indem ein tiefes neuronales Netzwerk trainiert wird, das diese Nichtlinearitäten abdecken kann. Ähnliche frühere Ansätze trainieren neuronale Netze in einer überwachten Weise unter

Verwendung bekannter 2D-3D-Korrespondenzen. Aufgrund der begrenzten Menge an unterschiedlichen Trainingsdaten neigen diese Modelle dazu, sich häufig vorkommende Posen im Trainingsdatensatz einfach zu merken und selten vorkommende Posen zu ignorieren. Um dies zu vermeiden, wird ein schwach überwachtes Trainingsschema vorgeschlagen, das eine Zuordnung zwischen Verteilungen von 2D- und 3D-Posen lernt. Die Konsistenz mit den 2D-Beobachtungen wird durch eine neuartige Rückprojektionsschicht erzwungen, welche die geschätzten 3D-Posen auf die 2D-Positionen zurückprojiziert. Die Performanz wird auf mehreren Benchmark-Datensätzen gezeigt und erreicht selbst im Vergleich zu überwachten Trainingsansätzen Ergebnisse, die mit dem aktuellen neuesten Stand der Technik konkurrieren. Die vorgeschlagene Lösung zeigt eine verbesserte Verallgemeinerung auf unübliche menschliche Posen und Kamerawinkel. Interessanterweise erhöht die Anwendung dieses Einzelbild-Ansatzes auf Videosequenzen den Rekonstruktionsfehler nicht signifikant.

Stichworte – Erfassung menschlicher Bewegungen, Poseschätzung, Kameraschätzung, Rückprojektionsfehleroptimierung.