Fortschritt-Berichte VDI

VDI

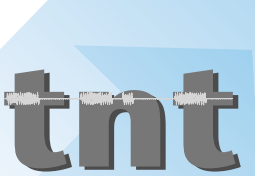
Reihe 10

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.

1 INTRODUCTION

With the continuous improvement of technology intelligent machines more and more influence our daily lives. Nowadays, every smartphone runs several computer programs that assist the owner, robots autonomously clean our homes and self-driving cars will be on the streets in the near future. To be part of the environment these machines need to interpret their surroundings and particularly people around them. On the one hand a machine should assist its owner or other people, but on the other hand it should not negatively influence them. To this end, it is essential to understand and interpret human motion, behavior and intentions. Before a machine is able to achieve this goal it first needs to capture them and translate its sensory input to a numeric representation. This task is commonly known as *Human Motion Capture (MoCap)*.

1.1 APPLICATIONS AND COMMERCIAL SYSTEMS

Human Motion Capture has a wide range of applications throughout the whole society. The entertainment market was one of the first to adapt and develop MoCap technologies. In movie production motions of human-like avatars are animated from motions of a human actor that are recorded in specially equipped MoCap studios. Some of the most remarkable entirely computer-generated avatars created by MoCap technology are King Kong, Gollum (Lord of the Rings) and Davy Jones (Pirates of the Caribbean). The consumer entertainment market also benefits from motions capture systems in the form of gaming devices such as the Microsoft Kinect [110]. They enable the user to interact with the system through natural gestures instead of a game controller. When integrated into virtual reality devices, motion capture systems enable the user to interact more realistically with the virtual environment in the future. In sports applications, the motions of athletes can be recorded, visualized and analyzed which helps to prevent injuries, detect unhealthy motions and optimize the athletes' movements. This not only allows physiotherapists to analyze even subtle motions but also gives the athlete the opportunity to have an external view on his or her movements. For several years the industry incorporates collaborative robots into their manufacturing processes. Instead of being isolated by security fences or walls, collaborative robots can share a workspace with humans by perceiving their environment with

multiple sensors. Nowadays, collaborative industrial robots are able to detect human presence or detect contact with a person by torque sensors. Combining them with MoCap technology to sense or predict human body parts will pave the way to a more collaborative workspace and increased security. Also, mobile platforms in logistics need to perceive humans to avoid collisions and injuries. In medical applications elderly people can be monitored to detect falls, or trembling of Parkinson patients can be quantified to adjust their medication.

For these applications there are several commercial systems available on the market. The most important ones are optical systems, which can be grouped into marker-based and markerless systems. In the first group the three most common professional systems are Vicon [95], Qualisys [67] and OptiTrack [59]. They require the user to wear a tight suit with optical markers attached to it. These products achieve highly accurate 3D estimations of the captured person. However, due to the high requirements for specific clothing, markerless solutions are developed. The most notable ones are available from Simi [77] and The Captury [82]. They achieve similar accuracy as the marker-based systems but can also capture people in everyday clothing. However, they still require a setup of multiple calibrated cameras. Other products without cameras exist using electromagnetic sensors [66], mechanical devices [55] and inertial measurement units [107]. Since these devices are attached to the human body they can affect the natural movements of the person. An ideal system, without the restrictions of all the products mentioned above, will be non-invasive and requires only a minimal uncalibrated sensor setup. Cameras as sensors appear to be a reasonable choice since they are physically non-invasive, produce an information-rich sensor output and are readily available as consumer products. Therefore, this thesis deals with human motion capture using only a single camera.

1.2 IMAGE-BASED MOTION CAPTURE

Human motion capture is defined as the process of recording the movement of a person from sensor measurements[1]. Although every kind of sensor data, e. g. velocities and accelerations from inertial measurement units [49], can be utilized to record human motions, in the following the terms *human motion capture* or *human pose estimation* (for a single pose) refers to 3D reconstructions from data recorded by a camera.

For decades video-based MoCap has been realized by placing optical markers on the human body and capture them with several synchronized cameras. The detected markers in each camera are then matched among them. The 3D positions of the markers are obtained by triangulation or

1 Here, only the major bones in the body are considered. If additionally hands and face are regarded it is commonly referred to as *Performance Capture*.

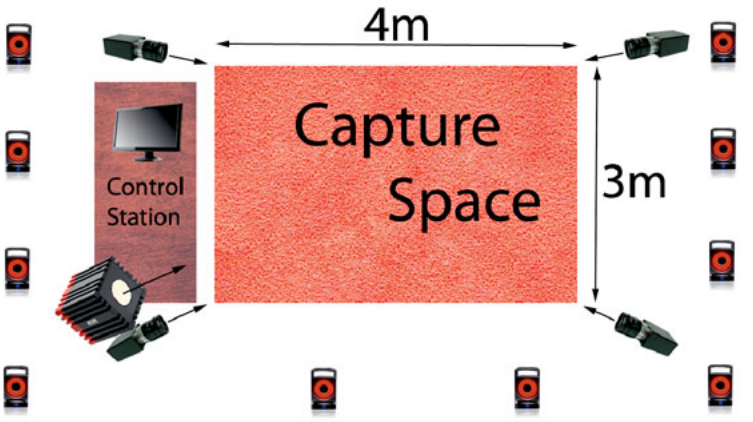


Figure 1.1: An exemplary camera arrangement of a traditional motion capture setting [33]. Ten infrared cameras capture the reflecting markers from different perspectives. Four additional RGB cameras acquire video data with $50Hz$.

similar methods. A standard MoCap setup is shown in Figure 1.1 using 10 infrared cameras and 4 RGB cameras to record a capture space of 3×4 meters. This technique has proven to be effective and is marketed in many commercial products (Section 1.1. However, wearing a marker suit is impractical in many real-world scenarios and a synchronized multi (infrared) camera setup is expensive, which limits its applicability to laboratory setups. Moreover, in scenarios where for instance a mobile device needs to interpret a human, multiple cameras are impractical or even infeasible. Therefore, this thesis focuses on the special case of **markerless MoCap from a single monocular camera** which can even be a consumer-grade mobile device. In contrast to other measurement devices, a consumer camera is an inexpensive device that produces an information-rich output. Moreover, a camera is non-invasive which means it has no physical impact on the movements of the recorded person. Due to the amount of data a camera produces, sophisticated computer vision solutions are required to extract the relevant data from the images. The extraction of only the meaningful information (i. e. a numerical description of the human pose) remains a challenge in current MoCap research. In recent years machine learning approaches have shown great success in accomplishing this task which is the reason why this thesis develops and improves state-of-the-art machine learning methods to solve the 3D pose estimation problem.

From Image to 3D. Reconstructing 3D poses of a human from an image or video is typically divided into two steps: first detecting the 2D

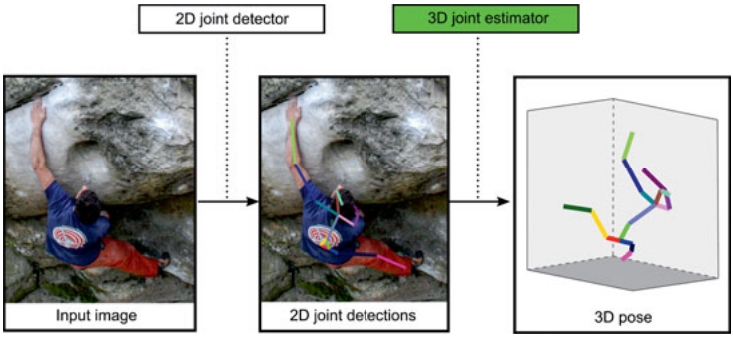


Figure 1.2: The reconstruction of a 3D human pose from an image or image sequence is usually divided into two steps: 1. detecting 2D keypoints in the image and 2. lifting these detections to 3D. The main part of this thesis, the 3D joint estimation, is marked in green. For better visibility the joints are connected by lines representing the *bones* of the underlying kinematic chain.

positions of the major joints, second lifting them to 3D. The steps are visualized in Figure 1.2. For the detection step many off the shelf joint detectors are available, including the well known and most used Stacked Hourglass Networks [58] and the real-time detector OpenPose [14]. The detection step is followed by the 3D reconstruction step which is the main focus of this thesis. The goal is to find the correct 3D pose of the person given 2D joint detections. Since the detections are projections of the corresponding 3D joints[2] to the image plane, the depth information is lost. From geometry follows that a candidate for a 3D point can lie anywhere on a straight line from the camera center through the 2D observation. Consequently, the 3D reconstruction from a monocular camera is an ill-posed problem with an infinite number of solutions. Many approaches exist to solve it that can be roughly divided into two categories. One category relies only on the detected 2D coordinates for the 3D reconstruction step and ignores the image data. The major advantage is that the detection and reconstruction steps are modular and can be interchanged. The other category of approaches learns an end-to-end system that directly infers a 3D pose from image data. However, most of them still require an intermediate 2D representation to work properly. When trained on a specific dataset its characteristic visual features (e. g. the background structure of indoor scenes) are learned which leads to exceptional performance on this dataset. Expectedly, the transferability to other image domains (e. g. outdoor scenes) is not ideal for the majority of these approaches. Since all presented methods in this thesis are

2 More precisely, they are only estimates for the projected 3D joints.

supposed to reconstruct humans in any scenario, here only detections are considered. This allows for a generalization to an arbitrary scene as long as a reliable joint detector is available.

3D Reconstruction. Recovering a 3D structure from monocular image sequences or even single images is a heavily under-constrained problem. One possible solution arises if sequences of images are regarded, which allows for the formulation of temporal priors to resolve depth ambiguities. This is known as the *Structure-from-Motion* problem [84] or, in the case of deforming objects, as the *Non-Rigid Structure-from-Motion (NRSfM)* [10] problem. For human motion capture the practical applicability is limited since all solutions to the NRSfM problem require sufficient object or camera motion to work reliably. To this end, two methods are proposed in this thesis to combine knowledge about human motion and poses which significantly improve traditional NRSfM approaches and make them applicable to human motion capture.

If only single images are considered, temporal constraints are no longer possible. Instead, the only knowledge about the scene is the presence of a person. Therefore, it appears reasonable to formulate a mathematical model of the human body that can be used to derive meaningful constraints for the solution space. Most recent approaches either describe the human body using joint angles or 3D coordinates of the major joints:

1. In accordance with the major joint types of the human body, namely ball, saddle and hinge joints, the human pose can be described by one, two or three angles per joint and the respective bone lengths. This representation separates the body configuration from anthropometric properties (e. g. bone lengths) and is therefore well-suited for motion analysis. However, projecting from joint angles to 2D key points requires multiple nonlinear processing steps.

2. A simple and intuitive representation is to describe a joint using its 3D coordinates. It can be easily projected to a 2D plane using simple matrix multiplications. In contrast to joint angles, properties of the human skeleton, e. g. constant bone length and symmetry, have to be enforced implicitly during the 3D reconstruction.

Both variants require a large number of variables to describe a single human pose, which are unknowns in the 3D reconstruction process. To reduce them and simplify the reconstruction a promising approach is to learn a subspace by a principal component analysis (PCA) [19] or similar methods. Most practicable methods are linear mappings from the data space into the subspace and vice versa. However, human motions are highly nonlinear and thus linear mappings appear to be not an optimal choice. For this reason, a method using a *Generative Adversarial Neural*

Network (GAN) [21] to learn a nonlinear subspace of human poses is proposed.

1.3 CONTRIBUTIONS

The goal of this thesis is to reconstruct 3D human poses from monocular images or videos. The presented contributions look at the problem from two perspectives:

1. Human movements are subject to several temporal constraints. For this reason, physically grounded assumptions are made on smooth motions, bone lengths constancy and symmetry. They are formulated to improve traditional NRSfM approaches for human MoCap.

2. Since human motions are highly nonlinear a Generative Adversarial Network is used to learn a space of plausible human poses. It is combined with a novel reprojection layer included into a neural network (called *RepNet*) that enforces consistency with 2D observations.

1.3.1 *Time Consistent Human Motion Reconstruction*

Given a time series of human poses a feasible assumption is that specific properties of the human body only slightly change from frame to frame or do not change at all: for instance, bone lengths remain constant in the 3D domain for the same person throughout a recorded sequence. Even for an unknown 3D skeleton this assumption is still valid although the exact bone lengths are unknown. This thesis formulates this fact as a minimization of the bone length changes over time. Since the changes instead of the absolute lengths are minimized the proposed solution is independent of an anthropometrically predefined skeleton. This led to plausible and temporally stable reconstructions. However, each human pose is defined by a previously learned PCA basis of poses and therefore does not cover all possible human poses.

To generalize, the bone lengths constancy assumption was relaxed and reformulated as a nuclear norm optimization problem in [98]. Here, only a known kinematic structure needs to be enforced which was done by developing the *Kinematic Chain Space (KCS)*. The proposed simple yet effective algorithm can be applied to every kinematic chain and is not restricted to human poses. A variation of the KCS is later applied for single image 3D human pose reconstruction in [99].

Summary of contributions:

- A method to formulate a temporal bone length constancy constraint as a variance minimization problem.

- A novel representation of human poses that efficiently encodes a kinematic chain in a Kinematic Chain Space (KCS).
- A nuclear norm based optimization which is derived by a relaxation of a bone length constancy constraint based on the KCS.

1.3.2 *RepNet*

In order to reconstruct a human pose, it needs to be mathematically represented. There are several representations in the literature. The most common ones are based on 3D coordinates of major joints of the human body or joint angles. The variety of human poses includes strong redundancies, e. g. the possible positions of the left hand are constrained by the position of the left elbow. Therefore, dimensionality reduction techniques, e. g. PCA, are helpful to reduce the number of variables. Although widely used PCA or similar methods only apply linear transformations. However, human poses are highly nonlinear. Therefore, a novel representation based on discriminator networks is proposed in Chapter 5. In contrast to previous approaches that apply neural networks to directly regress 3D coordinates of the skeletal joints from 2D inputs [51], a generative adversarial network (GAN) [21] is trained to achieve a mapping from a distribution of 2D poses to a distribution of feasible 3D poses. To enable the discriminator network to learn anthropometric properties, such as bone lengths and symmetries, it is extended by a layer implementing the mapping into the kinematic chain space which turned out to be very effective. By combining the GAN with a novel layer called *reprojection layer* that projects the 3D pose back to 2D the complete network is trained in a weakly supervised fashion and gives the proposed method its name *RepNet*. In contrast to supervised approaches, this weakly supervised training helps immensely to avoid overfitting.

Summary of contributions:

- A discriminator network that distinguishes predicted 3D coordinates from valid human poses.
- A kinematic chain space layer that enables the discriminator to learn anthropometric properties.
- A neural network motivated by GANs combined with a novel reprojection layer to infer 3D human poses from 2D detection.

1.4 STRUCTURE OF THE THESIS

The remainder of this thesis is structured in the following parts and visualized in Figure 1.3:

Chapter 2: An overview of the related work. To give a concise overview of existing methods for human pose estimation the research in the three fields *non-rigid structure from motion*, *time consistent motion capture* and *single image pose estimation* are presented and discussed.

Chapter 3: The fundamentals of human pose estimation and 3D reconstruction are introduced. Different camera representations are described. Dimensionality reduction techniques are explained and embedded in the context of human pose estimation.

Chapter 4: Two methods are proposed that exploit temporal properties of human movements. The first approach formulates a bone lengths constancy constraint as a variance minimization. It is based on a previously learned PCA basis of human poses. The second approach introduces the Kinematic Chain Space (KCS) which is employed to relax the bone length consistency formulation of the first approach. The nuclear norm optimization derived using the KCS allows for generalization to a larger set of human poses and can recover even subtle changes in human poses.

Chapter 5: This chapter presents *RepNet*, a neural network to directly infer 3D joint coordinates from 2D observations. It combines Generative Adversarial Networks with a novel reprojection layer. This layer ensures that the recovered human pose is not only anthropometrically correct but also satisfies reprojection constraints. The discriminator of the GAN is enriched with a mapping into the KCS from the previous chapter. In contrast to previous works, the complete network is trained in a weakly supervised fashion such that it efficiently avoids overfitting.

Chapter 6: The work is concluded and an outlook to future work is given.

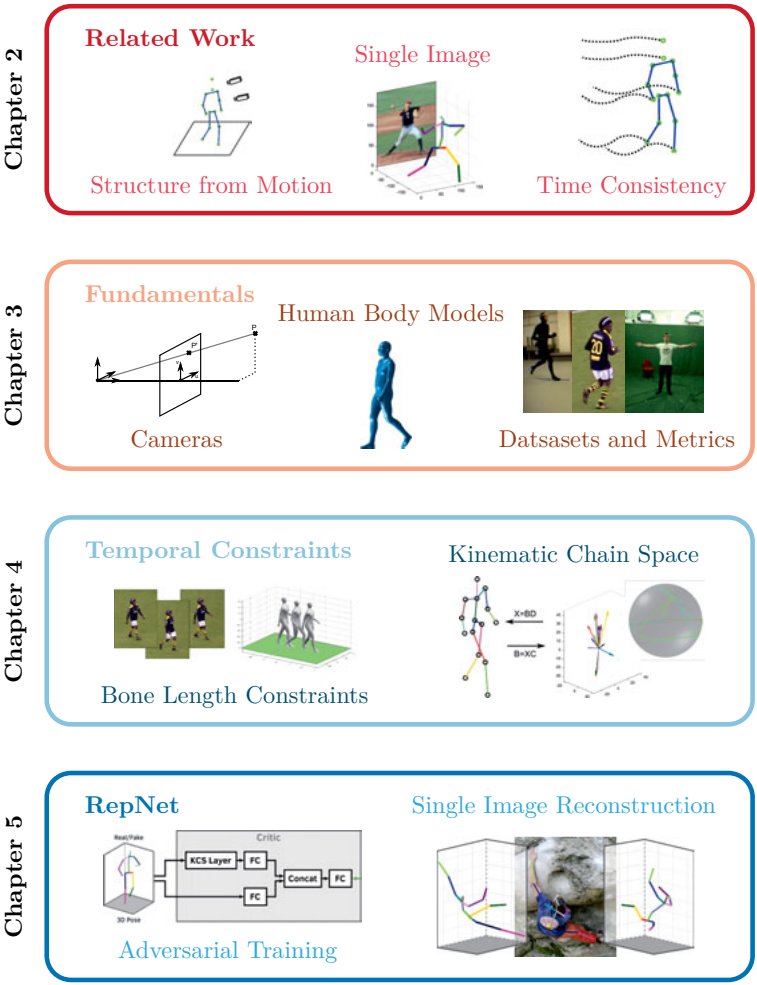


Figure 1.3: Thesis overview.

1.5 LIST OF PUBLICATIONS

This chapter lists the publications written during the time at TNT. Section 1.5.1 contains publications related to this thesis. Parts of this thesis are taken from these publications. Section 1.5.2 lists other publications in the fields of video coding and machine learning.

1.5.1 *Human Motion Capture*

[96] **Bastian Wandt**, Hanno Ackermann, Bodo Rosenhahn. 3D Human Motion Capture from Monocular Image Sequences. *In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015.

This paper tackles the problem of estimating non-rigid human 3D shape and motion from image sequences taken by uncalibrated cameras. Similar to other state-of-the-art solutions we factorize 2D observations in camera parameters, base poses and mixing coefficients. Existing methods require sufficient camera motion during the sequence to achieve a correct 3D reconstruction. To obtain convincing 3D reconstructions from arbitrary camera motion, our method is based on a-priorly trained base poses. We show that strong periodic assumptions on the coefficients can be used to define an efficient and accurate algorithm for estimating periodic motion such as walking patterns. For the extension to non-periodic motion we propose our novel regularization term based on temporal bone length constancy. In contrast to other works, the proposed method does not use a predefined skeleton or anthropometric constraints and can handle arbitrary camera motion. Multiple experiments based on a 3D error metric demonstrate the stability of the proposed method. Compared to other state-of-the-art methods our algorithm shows a significant improvement.

[97] **Bastian Wandt**, Hanno Ackermann, Bodo Rosenhahn. 3D Reconstruction of Human Motion from Monocular Image Sequences. *In: Transactions on Pattern Analysis and Machine Intelligence*, 2016.

This article tackles the problem of estimating non-rigid human 3D shape and motion from image sequences taken by uncalibrated cameras. Similar to other state-of-the-art solutions we factorize 2D observations in camera parameters, base poses and mixing coefficients. Existing methods require sufficient camera motion during the sequence to achieve a correct 3D reconstruction. To obtain convincing 3D reconstructions from arbitrary camera motion, our method is based on a-priorly trained base poses. We show that strong periodic assumptions on the coefficients can be used to define

an efficient and accurate algorithm for estimating periodic motion such as walking patterns. For the extension to non-periodic motion we propose a novel regularization term based on temporal bone length constancy. In contrast to other works, the proposed method does not use a predefined skeleton or anthropometric constraints and can handle arbitrary camera motion. We achieve convincing 3D reconstructions, even under the influence of noise and occlusions. Multiple experiments based on a 3D error metric demonstrate the stability of the proposed method. Compared to other state-of-the-art methods our algorithm shows a significant improvement.

[96] Petrissa Zell, **Bastian Wandt**, Bodo Rosenhahn. Joint 3D Human Motion Capture and Physical Analysis from Monocular Videos. *In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

Motion analysis is often restricted to a laboratory setup with multiple cameras and force sensors which requires expensive equipment and knowledgeable operators. Therefore it lacks in simplicity and flexibility. We propose an algorithm combining monocular 3D pose estimation with physics-based modeling to introduce a statistical framework for fast and robust 3D motion analysis from 2D video-data. We use a factorization approach to learn 3D motion coefficients and join them with physical parameters, that describe the dynamic of a mass-spring-model. Our approach does neither require additional force measurement nor torque optimization and only uses a single camera while allowing to estimate unobservable torques in the human body. We show that our algorithm improves the monocular 3D reconstruction by enforcing plausible human motion and resolving the ambiguity of camera and object motion. The performance is evaluated on different motions and multiple test data sets as well as on challenging outdoor sequences.

[3] Thiemo Alldieck, Marc Kassubeck, **Bastian Wandt**, Bodo Rosenhahn, Marcus Magnor. Optical Flow-based 3D Human Motion Estimation from Monocular Video. *In: Proc. of the German Conference on Pattern Recognition*, 2017.

This paper presents a method to estimate 3D human pose and body shape from monocular videos. While recent approaches infer the 3D pose from silhouettes and landmarks, we exploit properties of optical flow to temporally constrain the reconstructed motion. We estimate human motion by minimizing the difference between computed flow fields and the output of our novel flow renderer. By just using a single semi-automatic initialization step, we are able to reconstruct monocular sequences without joint annotation. Our test

scenarios demonstrate that optical flow effectively regularizes the under-constrained prob- lem of human shape and motion estimation from monocular video.

[109] Petrissa Zell, **Bastian Wandt**, Hanno Ackermann, Bodo Rosenhahn. Physics-based Models for Human Gait Analysis. *In: Springer Handbook of Human Motion*, 2018.

This chapter deals with fundamental methods as well as current research on physics-based human gait analysis. We present valuable concepts that allow efficient modeling of the kinematics and the dynamics of the human body. The resulting physical model can be included in an optimization-based framework. In this context, we show how forward dynamics optimization can be used to determine the producing forces of gait patterns. To present a current subject of research, we provide a description of a 2D physics-based statistical model for human gait analysis that exploits parameter learning to estimate unobservable joint torques and external forces directly from motion input. The robustness of this algorithm with respect to occluded joint trajectories is shown in a short experiment. Furthermore, we present a method that uses the former techniques for video-based gait analysis by combining them with a nonrigid structure from motion approach. To examine the applicability of this method, a brief evaluation of the performance regarding joint torque and ground reaction force estimation is provided.

[97] **Bastian Wandt**, Hanno Ackermann, Bodo Rosenhahn. A Kinematic Chain Space for Monocular Motion Capture. *In: Proc. of the European Conference on Computer Vision Workshops*, 2018.

This paper deals with motion capture of kinematic chains (e. g. human skeletons) from monocular image sequences taken by uncalibrated cameras. We present a method based on projecting an observation onto a kinematic chain space (KCS). An optimization of the nuclear norm is proposed that implicitly enforces structural properties of the kinematic chain. Unlike other approaches our method is not relying on training data or previously determined constraints such as particular body lengths. The proposed algorithm is able to reconstruct scenes with little or no camera motion and previously unseen motions. It is not only applicable to human skeletons but also to other kinematic chains for instance animals or industrial robots. We achieve state-of-the-art results on different benchmark databases and real world scenes.

[99] **Bastian Wandt**, Bodo Rosenhahn. RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human

Pose Estimation. *In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

This paper addresses the problem of 3D human pose estimation from single images. While for a long time human skeletons were parameterized and fitted to the observation by satisfying a reprojection error, nowadays researchers directly use neural networks to infer the 3D pose from the observations. However, most of these approaches ignore the fact that a reprojection constraint has to be satisfied and are sensitive to overfitting. We tackle the overfitting problem by ignoring 2D to 3D correspondences. This efficiently avoids a simple memorization of the training data and allows for a weakly supervised training. One part of the proposed reprojection network (RepNet) learns a mapping from a distribution of 2D poses to a distribution of 3D poses using an adversarial training approach. Another part of the network estimates the camera. This allows for the definition of a network layer that performs the reprojection of the estimated 3D pose back to 2D which results in a reprojection loss function.

Our experiments show that RepNet generalizes well to unknown data and outperforms state-of-the-art methods when applied to unseen data. Moreover, our implementation runs in real-time on a standard desktop PC.

1.5.2 *Other Publications*

[100] **Bastian Wandt**, Thorsten Laude, Yiqun Liu, Bodo Rosenhahn, Jörn Ostermann. Extending HEVC Using Texture Synthesis. *In: Proc. of the IEEE Visual Communications and Image Processing*, 2017.

The High Efficiency Video Coding (HEVC) standard provides superior coding efficiency compared to its predecessors. Nevertheless, the encoding of complex and thus hardly to predict textures either requires high bit rates or results in low quality of the reconstructed signal. To compensate for this limitation of HEVC, we propose a sophisticated texture synthesis framework which solves multiple lacks of previous texture synthesis approaches. By easing the bit rate cost for synthesizable regions and reallocating the freed bit rate resources to non-synthesizable regions, we are able to achieve average BD-rate gains of 21.9% for all-intra, 17.6% for low delay, and 16.3% for random access, respectively, while maintaining the same objective quality for the latter. Subjective tests for the

synthesizable regions confirm the objectively measured convincing results.

[101] **Bastian Wandt**, Thorsten Laude, Bodo Rosenhahn, Jörn Ostermann. Detail-aware image decomposition for an HEVC-based texture synthesis framework. *In: Proc. of the Data Compression Conference*, 2018.

Modern video coding standards like High Efficiency Video Coding (HEVC) provide superior coding efficiency. However, this does not state true for complex and hard to predict textures which require high bit rates to achieve a high quality. To overcome this limitation of HEVC, texture synthesis frameworks were proposed in previous works. However, these frameworks only result in good reconstruction quality if the decomposition into synthesizable and non-synthesizable regions is either known or trivial. The frameworks fail for more challenging content, e. g. for content with fine non-synthesizable details within synthesizable regions. To enable texture synthesis-based video coding with high quality for this content, we propose sophisticated detail-aware decomposition techniques in this paper. These techniques are based on an initial coarse segmentation step followed by a refinement step that detects even small differences in the previously segmented region. With this new approach, we are able to achieve average luma BD-rate gains of 13.77 % over HEVC and 3.03 % over the closest related work from the literature. Furthermore, the considerably improved visual quality in addition to the bit rate savings is confirmed by comprehensive subjective tests.

[102] **Bastian Wandt**, Thorsten Laude, Bodo Rosenhahn, Jörn Ostermann. Extending HEVC with a Texture Synthesis Framework using Detail-aware Image Decomposition. *In: Proc. of the Picture Coding Symposium*, 2018.

In recent years, there has been a tremendous improvement in video coding algorithms. This improvement resulted in 2013 in the standardization of the first version of High Efficiency Video Coding (HEVC) which now forms the state-of-the-art with superior coding efficiency. Nevertheless, the development of video coding algorithms did not stop as HEVC still has its limitations. Especially for complex textures HEVC reveals one of its limitations. As these textures are hard to predict, very high bit rates are required to achieve a high quality. Texture synthesis was proposed as solution for this limitation in previous works. However, previous texture synthesis frameworks only prevailed if the decomposition into synthesizable and non-synthesizable regions was either known or very easy. In

this paper, we address this scenario with a texture synthesis framework based on detail-aware image decomposition techniques. Our techniques are based on a multiple-steps coarse-to-fine approach in which an initial decomposition is refined with awareness for small details. The efficiency of our approach is evaluated objectively and subjectively: BD-rate gains of up to 28.81% over HEVC and up to 12.75% over the closest related work were achieved. Our subjective tests indicate an improved visual quality in addition to the bit rate savings.

[37] Florian Kluger, Christoph Reinders, Kevin Raetz, Philipp Schelske, **Bastian Wandt**, Hanno Ackermann, Bodo Rosenhahn. Region-based Cycle-Consistent Data Augmentation for Object Detection. *In: Proc. of the IEEE International Conference on Big Data Workshops*, 2018.

Roads constitute a major part of the lives of everybody. Heavy use, for instance by cars and especially trucks, and even soil movement lead to visible damages. While major roads are regularly inspected, smaller roads often lack attention. It is therefore of great interest to have camera-based systems which can automatically detect and even classify damages. This report presents a system developed by the authors as part of the Road Damage Detection and Classification Challenge at the 2018 IEEE Big Data Cup [47]. Further contributions made here are techniques to augment the small set of training data. As a major contribution we also propose refinements to the dataset and evaluation metric to improve the challenge.

[73] Marco Rudolph, **Bastian Wandt**, Bodo Rosenhahn. Structuring Autoencoders. *In: Proc. of the IEEE International Conference on Computer vision Workshops*, 2019.

In this paper we propose *Structuring AutoEncoders (SAE)*. SAEs are neural networks which learn a low dimensional representation of data and are additionally enriched with a desired structure in this low dimensional space. While traditional Autoencoders have proven to structure data naturally they fail to discover semantic structure that is hard to recognize in the raw data. The SAE solves the problem by enhancing a traditional Autoencoder using weak supervision to form a structured latent space. In the experiments we demonstrate, that the structured latent space allows for a much more efficient data representation for further tasks such as classification for sparsely labeled data, an efficient choice of data to label, and morphing between classes. To demonstrate the general applicability of our method, we show experiments on the benchmark

image datasets MNIST, Fashion-MNIST, DeepFashion2 and on a dataset of 3D human shapes.

2

RELATED WORK

The following chapter presents an overview over the related work. Section 2.1 briefly reviews the most important works in the field of non-rigid structure from motion which is closely related to the first part of the thesis, see Chapter 4. Section 2.2 discusses several methods for single image 3D human pose estimation. In Section 2.3 the rather small amount of publications in the field of time consistent human motion capture is reviewed. Commonly used error metrics and MoCap datasets are discussed in the subsequent chapter in Section 3.4 and 3.5.

2.1 NON-RIGID STRUCTURE-FROM-MOTION

In 1992 Tomasi and Kanade [84] presented the first factorization based approach for a set of 2D points tracked over a sequence. The presented factorization allowed for the reconstruction of the underlying rigid 3D scene. They decomposed the input data via a Singular Value Decomposition (SVD) into two sets of variables, one of which is associated with the motion parameters, the other with the coordinates of the rigid 3D structure. An extension to Tomasi and Kanade's approach was proposed in 2000 by Bregler et al. [10] which generalizes it to deforming shapes. They expressed the 3D shape in any particular frame as a linear combination of multiple rigid basis shapes. Different priors such as Gaussian assumptions or rank constraints were used by Torresani et al. [85–87] to avoid the troublesome step of non-rigid self-calibration. The basis shapes of Bregler et al. [10] are ambiguous as shown by Xiao et al. [106]. They proposed to employ constraints on the basis shapes to resolve the ambiguity. Later, Akhter et al. [2] showed that these basis constraints are still not sufficient to resolve the ambiguity. They exploit the duality of an object-independent trajectory basis and the formerly used shape basis. By employing a Discrete Cosine Transform (DCT) basis as the trajectory basis the number of unknown parameters is significantly reduced. The idea of [2] was taken over by Gotardo and Martinez [23] who applied the DCT representation to enforce a smooth 3D shape trajectory. Since the DCT basis restricts the reconstructions to specific predefined frequencies Gotardo and Martinez [22] proposed another solution that uses the kernel trick to model the nonlinear deformations. The kernel trick was also applied by Hamsici et al. [26] to learn a mapping between

the 3D shape and the 2D input data. Activity-independent spatial and temporal constraints were introduced by Park et al. [61]. Valmadre et al. [94] took inspiration from [2] and [61] which lead to a dynamic programming approach combined with temporal filtering. Dai et al. [17] impose a sparsity constraint and formulates it as a minimization of the trace norm of the transformation matrix. To avoid the sparsity constraint Lee et al. [41] define additional constraints on motion parameters. Rehan et al. [70] were the first to propose a reconstruction method for rigidly deforming objects, such as human skeletons. This is achieved by factorizing only a small number of consecutive frames.

Although the above approaches are targeted to reconstruct arbitrary deforming objects, they are also suitable for the special case of human motion capture. Indeed, there is a human motion sequence in the NRSfM benchmark dataset of [23]. However, due to the more general formulation of the problem, there are several constraints on the setting that need to be satisfied. The most concerning one regarding human motion capture is the need for a large camera motion. Therefore, it is impractical for most in-the-wild motion capture scenarios. The works [96–98] that are part of this thesis present how to efficiently solve this problem by introducing knowledge about human skeletons to the NRSfM problem. Further information and detailed discussions about other works on NRSfM can be found in [60].

2.2 SINGLE IMAGE APPROACHES

The research field of human pose estimation is constantly growing and there is a vast amount of publications. This section summarizes only the most relevant works that left an impact and guided future research. An exhaustive overview can be found in [74]. The recovery of a 3D human pose from a single image dates back to the work of Lee and Chen [40] in 1985. They use a binary decision tree and a known skeletal model of the person in the image. The more recent approaches can be roughly divided into two categories. The first group contains optimization-based approaches. The basic idea is to deform a predefined or learned 3D human body model such that it satisfies a reprojection error, i. e. the distance of the given 2D points to the inferred 3D points backprojected to 2D[1]. The second category contains fairly new approaches benefiting from the recent rise of neural networks. They try to estimate 3D poses directly from images or keypoints detected by a 2D joint detector.

1 The definition of the backprojection error can be found in Section 3.2.

2.2.1 *Reprojection Error Optimization*

Jiang [34] divides a pose into the upper and lower body and searches a database of more than a million poses for the best fitting pose. Chen and Ramanan [15] follow a similar approach by finding the nearest neighbor from large human pose database which best fits the observations. Since this simple database lookup is very expensive a widely used approach is to compress the knowledge from these databases in an overcomplete dictionary by either using principal component analysis (PCA) or another dictionary learning method. The most common method is to use a linear basis for 3D human poses which is obtained by a PCA basis. Wei et al. [104] define bone symmetry constraints and rigid body constraints to restrict the solution space. Ramakrishna et al. [68] propose a regularization term based on known proportions in the human body. They also develop a method called *Projected Matching Pursuit* for the coordinate descent on the reprojection error objective function. The approach of [68] was extended by Wang et al. [103] with a sparsity regularization of the base pose coefficients. Simo-Serra et al. [78] sample a large amount of 3D pose candidates that satisfy the reprojection constraint and in a second step select the most plausible pose in terms of anthropometric regularity. Akhter et al. [1] calculate joint angle limits for the main body joints to enforce more plausible 3D reconstructions. Zhou et al. [112] developed a convex relaxation for the reprojection error.

2.2.2 *Direct Inference using Neural Networks*

Recently, neural networks are directly applied to regress a 3D human pose either from image data directly or from 2D joint detections in a preprocessing step. Li et al. [42] were the first to learn CNNs to directly regress a 3D pose from image input. By integrating a structured learning framework into CNNs they later improved their work [43]. Tekin et al. [81] introduce a deep learning architecture that relies on an overcomplete autoencoder. Park et al. [63] learn relative 3D positions between joints. To facilitate depth estimation Du et al. [18] integrated height maps into their framework. A transfer learning approach is introduced by Mehta et al. [52] to allow for in-the-wild pose estimation of datasets where no training data is available. They also released a new dataset called *MPI-INF-3DHP* containing more diverse poses than existing datasets at that time. This framework was later extended by Mehta et al. [53] to achieve real-time performance. A voxel-based approach is introduced by Pavlakos et al. [64]. Another real-time approach is proposed by Rogez et al. [72] that is capable of reconstructing several persons in one image. Luo et al. [45] represent a human pose by using limb orientations.

There are other approaches that do not consider the image data directly but use a pretrained 2D joint detector. Moreno-Noguer [56] represents a human pose as a distance matrix and learns a mapping from 2D to 3D distance matrices. Martinez et al. [51] directly train a neural network on 2D detections and 3D ground truth. It achieves an impressive performance on the benchmark dataset Human3.6M [33]. However, this approach is very sensitive to overfitting on a specific dataset since it has significantly more parameters than poses in the training set of Human3.6M. Instead of learning an understanding of the structure of a human pose this indicates a simple memorization of poses. The approach of [51] was extended by Hossain et al. [69] by employing a recurrent neural network for sequences of human poses.

2.3 TIME CONSISTENT HUMAN MOTION CAPTURE

Exploiting temporal coherence in image sequences appears to be a promising approach to resolve the natural ambiguities that occur from projecting 3D points to 2D. Temporal properties are difficult to formulate as an optimization problem that can be solved uniquely. The most important approaches are summarized in the following. An exhaustive overview can be found in [74].

Early works on human pose estimation from image sequences build volumetric body models by approximating each major body joint as a tube or similar geometric shape and project it to the semantically segmented image plane [9, 76]. Wei and Chai [104] apply a parameterized human body model and deform it to fit the 2D observations. They enforce rigidity constraints in several frames and estimate camera and body pose by a nonlinear optimization algorithm. Valmadre and Lucey [93] contradicted some of the statements in [104]. They found that rigidity constraints are not sufficient for a unique solution and should only be enforced on sub-structures. They proposed an approach using deterministic structure from motion based on assumptions of rigidity only in the body's torso. Tekin et al. [80] use convolutional neural networks to estimated a spatio-temporal volume of bounding boxes. These are used to regress the 3D pose in the central frame. In a later work [81] they employ an autoencoder trained on existing human poses to learn a structured latent representation of the human pose in 3D. Other latent variable models are also used by some authors [4, 79, 83, 108]. The most recent works used a learned pose basis and enforce temporal bone length constancy constraints [97, 98][2] or recurrent neural networks [69]. A different approach is taken by Alldieck et al. [3]. They estimate the optical

2 Parts of this thesis are based on these publications.

flow between two consecutive frames and use a differentiable renderer to produce an artificial optical flow that best matches the observation.

3

FUNDAMENTALS

This chapter explains the fundamental mathematical concepts used in this thesis. In Section 3.1 an overview over different camera models and their simplifications is given which is mainly based on [27]. Representations of human poses used in this thesis are explained in Section 3.2. Additionally, linear subspaces (e. g. obtained by a PCA) and nonlinear (e. g. learned by neural networks) of human poses are introduced. The basics of factorization based (non-rigid) structure from motion approaches are explained in Section 3.3. Common error metrics to evaluate the quality of the 3D reconstructions are discussed in Section 3.4. The datasets used for training and evaluation are described in Section 3.5.

3.1 CAMERA MODELS

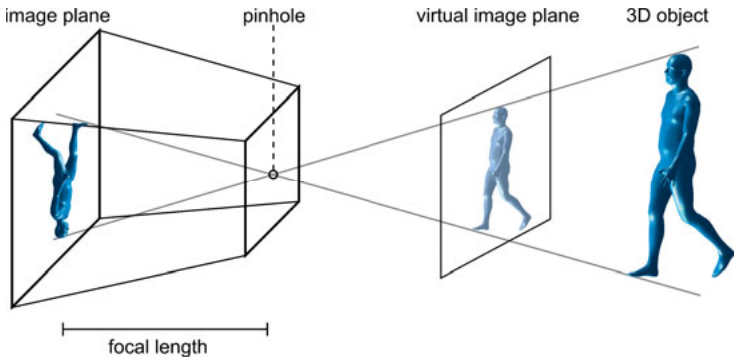


Figure 3.1: The pinhole camera model. The light emitted by a 3D object enters a dark box through a small hole and produces an inverted 2D image.

This section explains and derives the different camera models used in this thesis. A brief discussion on the applicability of the different models in the context of monocular human motion capture is given.

The common pinhole camera model is visualized in Figure 3.1. A 3D object can be seen through a small pinhole in an otherwise opaque plane. If an image plane is placed behind the pinhole the 3D object gets projected to this plane. The distance from the image plane to the pinhole

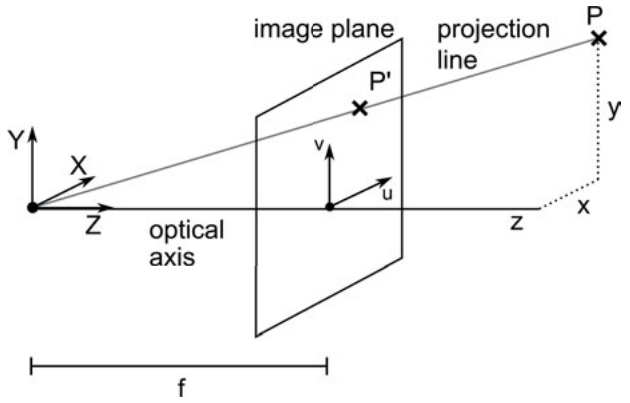


Figure 3.2: Mathematical description of the pinhole camera model.

is called *focal length* f. Since the rays of light from the object through the pinhole are straight the 2D projection appears inverted on the image plane. From symmetry follows that there is an inverted image plane in front of the pinhole, called the *virtual image plane*. The distance of the virtual image plane to the pinhole is again the focal length. Note that this is an idealized model with an infinitely small pinhole.

3.1.1 *Projective Transformations*

Mathematically a camera performs a projective transformation of the points of the observed 3D scene to map them to the image plane. In contrast to Euclidean transformations a projective transformation only maps straight lines to straight lines but does not preserve angles and distances. In Euclidean space two parallel lines do not have an intersection or are sometimes referred to as having an *intersection at infinity*. These points do not exist in Euclidean space. However, 3D points at infinity can get projected to a 2D image plane. For example, if two parallel lines in 3D space are observed on the image plane they meet in a specific point, which is called the *vanishing point*. To describe these projections of points at infinity *homogeneous coordinates* can be used. By definition, a point in 2D Euclidean space is described by the pair (x, y) and is extended to homogeneous coordinates by the triplet $(x, y, 1)$. Moreover, each triplet (kx, ky, k) with $k \in \mathbb{R}$ corresponds to the same point in Euclidean space.

By using homogeneous coordinates cameras can be described as a linear map of a point $P = (x, y, z, 1)^T$ to its projected point $P' = (u, v, w)^T$ on the image plane which is visualized in Figure 3.2. (x, y, z) are the 3D

coordinates of P and $(u/w, v/w)$ are the 2D image coordinates of P'. Let $\boldsymbol{K} \in \mathbb{R}^{4\times4}$ be a linear map that projects P to P' by

$$P' = \boldsymbol{K}P. \tag{3.1}$$

The matrix $\boldsymbol{K}$ is called *camera matrix* in this thesis. The camera matrix can be decomposed into an *intrinsic* and *extrinsic* parameter matrix by

$$P' = \boldsymbol{K}P = \boldsymbol{C}_{in}\boldsymbol{C}_{ex}P, \tag{3.2}$$

where $\boldsymbol{C}_{in} \in \mathbb{R}^{3\times4}$ and $\boldsymbol{C}_{ex} \in \mathbb{R}^{4\times4}$ contain the intrinsic and extrinsic parameters, respectively. The intrinsics model the internal camera properties such as focal length, center point and distortion. They are described in Section 3.1.2. The extrinsics describe the rotation and translation from a global coordinate system to the camera coordinate system and are explained in Section 3.1.3.

3.1.2 *Intrinsic Parameters*

Reconsidering Figure 3.2 the 3D point $P = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$ in the camera coordinate system is projected to the image plane. From symmetry follows

$$\frac{x}{z} = \frac{u}{f} \tag{3.3}$$

and

$$\frac{y}{z} = \frac{v}{f}, \tag{3.4}$$

with u, v as the image coordinates of the projected point P. The projection P' of the point P is obtained by

$$\begin{pmatrix} u \\ v \end{pmatrix} = \frac{f}{z}\begin{pmatrix} x \\ y \end{pmatrix}. \tag{3.5}$$

Using homogeneous coordinates for the 2D projections it can be written as

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \frac{f}{z}\begin{pmatrix} x \\ y \\ \frac{z}{f} \end{pmatrix} = \frac{1}{z}\begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \frac{1}{z}\boldsymbol{C}_{in}\begin{pmatrix} x \\ y \\ z \end{pmatrix}, \tag{3.6}$$

where $\boldsymbol{C}_{in}$ is the matrix containing the intrinsic camera parameters. If the origin of the image plane is not at the center but at the point $\begin{pmatrix} c_u \\ c_v \end{pmatrix}$ the matrix $\boldsymbol{C}_{in}$ in Eq. (3.6) extends to

$$\boldsymbol{C}_{in} = \begin{pmatrix} f & 0 & c_u \\ 0 & f & c_v \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.7}$$

In real-world cameras the pixels are not perfectly square and the projection plane can be slightly skewed which can be modeled using the stretching parameters d, e and the skew parameter α by

$$\boldsymbol{C}_{in} = \begin{pmatrix} df & \alpha & c_u \\ 0 & ef & c_v \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.8}$$

If the 3D point is described by homogeneous coordinates Eq. (3.8) is written as

$$\boldsymbol{C}_{in} = \begin{pmatrix} df & \alpha & c_u & 0 \\ 0 & ef & c_v & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{3.9}$$

Since the methods proposed in this thesis employ no camera intrinsics calibration step, distortions from manufacturing inaccuracies are neglected. Additionally assuming that the 2D coordinate origin is at the center $\boldsymbol{C}_{in}$ simplifies to

$$\boldsymbol{C}_{in} = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{3.10}$$

3.1.3 *Extrinsic Parameters*

The extrinsic parameters describe the rotation and translation to a global coordinate system. Let P_K be the 3D point P in the camera coordinate system. Applying only rotation and translation P_K is calculated by

$$P_K = \boldsymbol{R}P + \boldsymbol{t}, \tag{3.11}$$

where $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ is a rotation matrix and $\boldsymbol{t} \in \mathbb{R}^3$ is a vector containing the translational components. The rotation matrix is orthonormal, i. e. $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}_3$ or $\boldsymbol{R}^{-1} = \boldsymbol{R}^T$. Since reflections of 3D objects are impossible to

achieve by physical manipulation in the real world the second important property is $\det(\boldsymbol{R}) = 1$. Therefore, $\boldsymbol{R}$ belongs to the group known as the *special orthogonal group* $SO(3)$.

To avoid the summation in Eq. (3.11) it is written in homogeneous coordinates

$$\begin{pmatrix} P_K \\ 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0} & 1 \end{pmatrix} \begin{pmatrix} P \\ 1 \end{pmatrix}, \tag{3.12}$$

where $\boldsymbol{0} = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}$. Combined with the intrinsic camera matrix from Eq 3.10 it leads to the final formulation for the projection to the image plane

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \frac{1}{z_K} \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0} & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}, \tag{3.13}$$

where z_K is the z component of the rotated and translated point P.

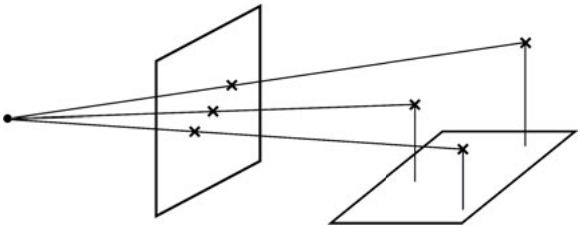
3.1.4 *Simplified Camera Models*

This thesis proposes several approaches for 3D reconstruction of non-rigid objects from multiple keypoints detected in monocular images without a previously calibrated camera and without knowledge about the exact shape of the observed object. According to Eq. (3.6) the 2D position of each point depends on its respective depth. Since the depth of each point is unknown this results in infinitely many solutions for the 3D object given only the 2D points. One possible solution to this problem is a relaxation of Eq. (3.6) to a *weak perspective* (or *scaled orthographic*) projection. For objects with minor deviations in z-direction compared to the distance to the camera it can be assumed that all points on the object lie on a plane parallel to the image plane with distance z_0. This is done by relaxing Eq. (3.6) using only a single scale component $s = \frac{f}{z_0}$ which leads to

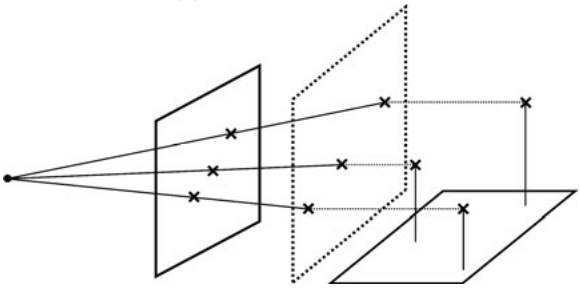
$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}. \tag{3.14}$$

Visually it can be seen as an orthogonal projection of each point to a plane with distance z_0 to the image plane, as seen in Figure 3.3b.

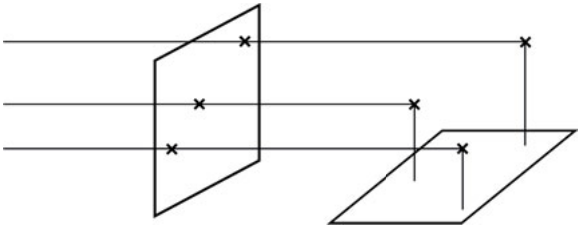
Eq. (3.14) can be relaxed further to an orthographic projection with $s = 1$, i. e. $\boldsymbol{C}_{in} = \boldsymbol{I}_3$. Geometrically this equals to a scene at infinite distance from the camera or equivalently an optical center at infinity. All projection lines are parallel to the optical axis. A visual comparison of the



(a) Perspective camera.



(b) Weak perspective camera.



(c) Orthographic camera.

Figure 3.3: Comparison of perspective, weak perspective and orthographic camera models.

camera models can be seen in Figures 3.3a, 3.3b and 3.3c. Early works on structure from motion e. g. [84] assumed an orthographic projection of the observed object.

In this thesis only weak perspective projections are used. By centralizing the 2D and 3D data points the translation component becomes zero and the homogeneous coordinates can be avoided. Eq. (3.13) can then be written as

$$\begin{pmatrix} u \\ v \end{pmatrix} = s\tilde{\boldsymbol{R}} \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \tag{3.15}$$

where $\tilde{\boldsymbol{R}} \in \mathbb{R}^{2\times 3}$ is the first and second row of a rotation matrix. Note that $\tilde{\boldsymbol{R}}$ is not a rotation matrix since it is not quadratic.

3.2 HUMAN POSE REPRESENTATIONS

Human poses can be mathematically represented in various ways. The most common ones are based on the skeleton of the human body. Depending on the task only the major joints or every single joint of this skeleton may be regarded. This can simply be defined by the 3D coordinates of these joints which is discussed in Section 3.2.1. Human poses can also be described by the angles between joints using exponential maps and twists. Since these descriptions are not considered this thesis the reader is referred to [57]. Particularly with the improvement of 3D scanning technology volumetric body shape models caught recent attention (Section 3.2.2). In contrast to skeletal models, they allow for a precise definition of the human body shape. Since both representations have a large number of variables different subspaces of human poses are discussed in Section 3.2.3. These different representations will be introduced and discussed in this section.

3.2.1 *Coordinate-based Representations*

The simplest and most intuitive representation of a human pose considers only the main joints of the human skeleton. An example skeleton, which is used in most parts of this thesis, can be seen in Figure 3.4. This skeleton is also used in most publications using a *17 joint model*, e. g. [51]. The choice of the joints depends strongly on the task, e. g. if the pose of a running person is reconstructed the fingers are less important than in the task of body and sign language recognition. Although the body shape is ignored the descriptive power for further tasks, such as motion recognition or prediction, remains. There are several possibilities to order the x, y, z-coordinates of the respective joints to describe a single pose.

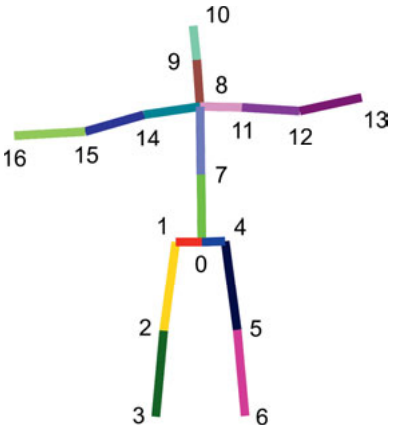


Figure 3.4: The 17 joint skeleton model used for most parts of this thesis. The body is oriented forward, i. e. the joint numbered 13 is the right hand and the joint numbered 16 is the left hand.

The most important one, which was already used in the seminal work of Bregler et al. [10], is explained in the following.

Let

$$\boldsymbol{X} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_J) \tag{3.16}$$

represent a human pose with J joints, where every $\boldsymbol{x}_i \in \mathbb{R}^{3\times 1}$ with $i = 1, 2, \ldots, J$ contains the x, y, z-coordinate of the joint i. In contrast to simply stacking the coordinate values as a vector of the form $(x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_J, y_J, z_J)^T$, Eq. (3.16) benefits from being easily projected to 2D by

$$\boldsymbol{X}_{2d} = \boldsymbol{K}\boldsymbol{X}, \tag{3.17}$$

where $\boldsymbol{K} \in \mathbb{R}^{2\times 3}$ is a projection matrix (different projections are discussed in Section 3.1). The reprojection matrix $\boldsymbol{X}_{2d}$ then has the form

$$\boldsymbol{X}_{2d} = (\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_J), \tag{3.18}$$

where each $\boldsymbol{u}_i \in \mathbb{R}^{2\times 1}$ contains the 2D coordinates projected to an image plane. This allows for the easy definition of a reprojection error

$$e_{rep} = \|\boldsymbol{W} - \boldsymbol{X}_{2d}\| = \|\boldsymbol{W} - \boldsymbol{K}\boldsymbol{X}\|, \tag{3.19}$$

where $\| \bullet \|$ is a matrix norm[1] and $\boldsymbol{W}$ is the measurement matrix which has the same structure as $\boldsymbol{X}_{2d}$. In almost all works on 3D reconstruction this error is minimized either implicitly or explicitly.

1 In this thesis the Frobenius norm is used which is defined by the square root of the sum over all quadratic matrix entries. It is denoted as $\| \bullet \|_F$.

So far Eq. (3.17) ignores translational components. To generalize to translations Eq. (3.17) can be written in homogeneous coordinates

$$\boldsymbol{X}_{2d} = \begin{pmatrix} \boldsymbol{K} & \boldsymbol{t} \end{pmatrix} \begin{pmatrix} \boldsymbol{X} \\ \mathbf{1} \end{pmatrix}, \tag{3.20}$$

where $\mathbf{1}$ denotes a vector filling the respective row with ones and $\boldsymbol{t} \in \mathbb{R}^{2\times 1}$ describes the translation in the image plane. If $\boldsymbol{K}$ represents a (scaled) orthographic projection (cf. Section 3.1) then the translational component can be removed by subtracting the mean from the respective x, y, z-coordinates of $\boldsymbol{W}$ and $\boldsymbol{X}$, respectively. In geometric terms this equals to moving the pose to the origin. Since the translational component $\boldsymbol{t}$ is zero Eq. (3.20) then can again be written as 3.17.

3.2.2 *Surface Mesh-based Representations*

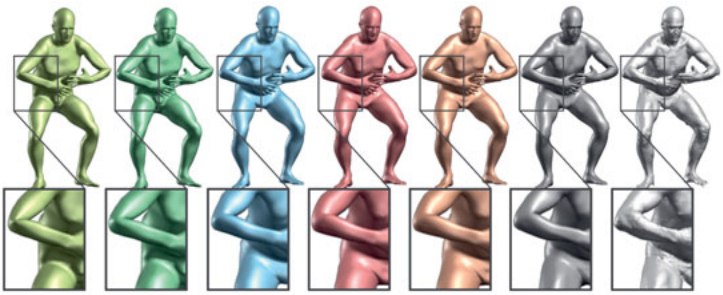


Figure 3.5: A comparison between SCAPE, BlendSCAPE [30] and SMPL. From left to right: (light green) Linear blend skinning (LBS), (dark green) Dual-quaternion blend skinning (DQBS), (blue) BlendSCAPE, (red) SMPL-LBS, (orange) SMPL-DQBS. The main differences can be seen in the highlighted regions around the elbow and hip. Image is taken from [44].

If not only the pose but also the shape of the body is of interest it can be represented as a watertight polygon mesh consisting of vertices, faces and edges (in the following shortly denoted as *mesh*). In practice a 3D point cloud is obtained by a 3D body scanner, a Microsoft Kinect or a similar device. Subsequently, a mesh with known topology is fitted to the measured point cloud. Since even small structures such as fingers should be modeled the mesh obviously must have a large number of vertices. Considering fitted meshes approximating several people, there is a strong redundancy in the data. Moreover, a user who wants to modify these meshes should have an easier deformation method than changing every single vertex.

There are two common state-of-the-art human mesh representation in the literature, namely SCAPE [5] and SMPL [44]. Both models divide the problem into body shape (e. g. small vs. tall) and pose (e. g. hanging vs. raised hands) estimation. SCAPE and SMPL both apply a PCA to registered meshes to account for shape deformation. The main difference between the models lies in the shape deformation model. In SCAPE it is based on triangle deformations using twists [57] and a computationally expensive refinement step. SMPL uses a Blend Skinning [39] based approach to model deformations. The SMPL algorithm is independent of the applied skinning technique. The initial blend weights are set manually. The model parameters are learned such that the complete model reduces to a function depending only on two variable vectors which define body shape and pose. Due to its simplicity SMPL is the most used model for human mesh representation today. A subjective comparison between the two models can be seen in Figure 3.5.

In this thesis the SMPL model is used only for visualization.

3.2.3 *Subspaces of Human Poses*

The coordinate-based human body representation in Section 3.2.1 requires a large set of variables to define a specific pose. Since every joint position is described by its x,y,z coordinates a pose constructed from J joints is described by $3J$ values. However, not all combinations of variables yield a physically plausible human pose. It can be easily seen that only some values in the coordinate-based representation give a meaningful human pose. This gives rise to the idea that there exists a subspace which contains all physically valid human poses. This section discusses methods to estimate such subspaces.

3.2.3.1 *Linear Subspaces*

An overcomplete dictionary can be learned from a dataset of human poses such that a single pose $\boldsymbol{x} \in \mathbb{R}^{3J}$ in vectorial form can be reconstructed from a dictionary $\boldsymbol{D} \in \mathbb{R}^{3J \times d}$ with d bases such that

$$\boldsymbol{x} \approx \boldsymbol{D}\boldsymbol{y}, \tag{3.21}$$

where each row in $\boldsymbol{D}$ is one basis vector that is weighted by the mixing coefficients $\boldsymbol{y} \in \mathbb{R}^d$. An optimal dictionary can be obtained by minimizing the optimization problem

$$\min_{\boldsymbol{D},y} \|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{y}\| \tag{3.22}$$

for all poses in the training dataset. Principle Components Analysis (PCA) [19] is one of the most used techniques to solve this problem.

The principal components (PC), i. e. the rows in $\boldsymbol{D}$, are calculated in such a way that the first PC covers the largest variance in the data. Per definition, each succeeding PC accounts for the highest variance under the constraint that it is orthogonal to the previous PC's.

In [88] it was shown that the first principal component obtained by a PCA covers 84% of the variance in a gait motion. Only 4 principal components of a gait motion are required to account for more than 98% of the variance. This idea was adapted by many 3D reconstruction approaches from single images 2.2.1 and also is the basis for the periodic and non-periodic reconstruction method proposed in Chapter 4 of this thesis. With a given PCA basis $\boldsymbol{Q} \in \mathbb{R}^{3d \times J}$ (obtained by stacking the mean pose and all principal components converted into the shape of a pose, as defined in Section 3.2.2) the pose estimation problem defined by the reprojection error (cf. Section 3.2.1) is simplified to

$$\min_{\boldsymbol{K},\boldsymbol{\theta}} \|\boldsymbol{X} - \boldsymbol{K}\boldsymbol{\theta}\boldsymbol{Q}\|, \tag{3.23}$$

where

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{I}_3 & \vartheta_1 \boldsymbol{I}_3 & \vartheta_2 \boldsymbol{I}_3 & \dots & \vartheta_d \boldsymbol{I}_3 \end{pmatrix} \tag{3.24}$$

with ϑ_i as the value of the i-th mixing coefficient. Now only the mixing coefficients ϑ instead of the complete pose need to be estimated which significantly reduces the number of variables.

3.2.3.2 *Discriminator Networks as Subspace Constraint*

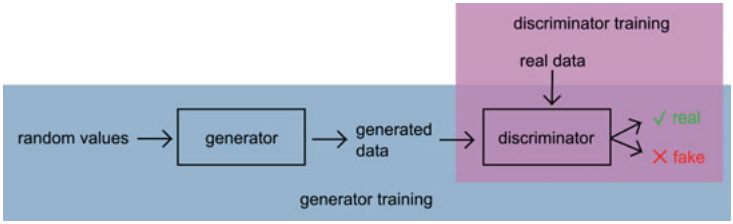


Figure 3.6: Structure of a GAN. The training process splits into two alternating steps: generator and discriminator training. During generator training the weights in the discriminator are fixed.

The subspace learning methods in Sec 3.2.3 learn a linear mapping from a data space to a latent space. For the application in human pose estimation it is important that the inverse mapping from the latent space to the data space exists. This restricts the choice to invertible subspace mapping methods. Most of the eligible methods are linear mappings[2] (see

2 except autoencoders with nonlinear activation functions

Section 3.2.3.1). However, human poses are highly nonlinear and therefore seemingly not the ideal choice. As later shown in Chapter 5 discriminator networks, as used in *Generative Adversarial Networks* (GAN) [21], are an excellent tool to constrain the human pose space. The key idea is to avoid the explicit mapping into the subspace by constraining the output of the generator to plausible human poses. Thus, after successful training the generator can only output human poses that lie in the learned subspace. The same argument holds for any type of data a GAN outputs. Since the original GAN paper [21] a vast amount of modifications were made to improve the performance. For this reason, this section gives an overview of the key ideas behind generative adversarial networks and avoids going into the mathematical details of the original implementation.

The original GAN structure of [21] is shown in Figure 3.6. A generator network learns a mapping from randomly distributed variables (in most cases a Gaussian or equal distribution) to a target distribution. From the target distribution only samples from the training data are known. There exist no correspondences between the input samples to a sample in the target distribution. The training splits into two alternating steps: 1. training the generator to output a sample labeled as real by the discriminator and 2. training the discriminator. In the first step the target distribution is learned by the discriminator network which is trained to classify data from the training set as *real* and data produced by the generator as *fake*. In the second step the complete adversarial model (generator connected to the discriminator, as shown in Figure 3.6) is trained while the discriminator weights are fixed. The target output of the adversarial model is the class *real*. In every iteration the generator gets better to generate data that the discriminator cannot distinguish from the real data, and simultaneously the discriminator gets better in classifying data into real and fake. When the training is converged the generator outputs data that is close to the distribution of the real data. Consequently, the generator has learned to map from a latent distribution to the constrained space of the real data. The generator network can be adapted to the task and cope with nonlinearities in the data. Therefore, when applied to human poses, it expectedly achieves similar or better performance than a PCA in terms of reconstruction error. The properties explained above are exploited in Chapter 5.

In Chapter 5 the Wasserstein GAN [6] and Improved Wasserstein GAN [25] works were used. Details can be found in Chapter 5 and the respective papers.

3.3 NON-RIGID STRUCTURE FROM MOTION

Structure from Motion (SfM) deals with recovering the 3D structure and motion of a rigid object from its 2D projections. The extension to

Non-Rigid Structure from Motion (NRSfM) additionally includes the recovery of deforming objects. This section introduces the basic ideas behind the factorization approaches in the seminal papers of Tomasi and Kanade [84] (SfM) and Bregler et al. [10] (NRSfM). Further information and detailed discussions can be found in [60].

Corresponding to the original works the following part assumes an orthographic projection (cf. Section 3.1.4). A measurement matrix $\boldsymbol{W} \in \mathbb{R}^{2f \times n}$ that contains the coordinates $u_{i,g}, v_{i,g}$ of the tracked feature point i at frame g can be written as

$$\boldsymbol{W} = \begin{pmatrix} u_{1,1} & u_{2,1} & \dots & u_{n,1} \\ v_{1,1} & v_{2,1} & \dots & v_{n,1} \\ u_{1,2} & u_{2,2} & \dots & u_{n,2} \\ v_{1,2} & v_{2,2} & \dots & v_{n,2} \\ \vdots & & & \vdots \\ u_{1,f} & u_{2,f} & \dots & u_{n,f} \\ v_{1,f} & v_{2,f} & \dots & v_{n,f} \end{pmatrix}, \tag{3.25}$$

where f is the number of frames and n is the number of feature points. To ignore translational camera movement from each row its mean is subtracted, i. e. the cameras can be described by only rotational components. Since each single observation is a projection of a rigid 3D structure $\boldsymbol{W}$ contains many redundant equations and therefore is strongly rank deficient. Without measurement noise $rank(\boldsymbol{W}) \leq 3$. The reason is that $\boldsymbol{W}$ can be factorized into a matrix $\tilde{\boldsymbol{R}} \in \mathbb{R}^{2f \times 3}$ representing the camera and a shape matrix $\boldsymbol{S} \in \mathbb{R}^{3 \times n}$ by

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{R}_1 \\ \boldsymbol{R}_2 \\ \vdots \\ \boldsymbol{R}_f \end{pmatrix} \begin{pmatrix} x_1 & x_2 & & x_n \\ y_1 & y_2 & \dots & y_n \\ z_1 & z_2 & & z_n \end{pmatrix} = \tilde{\boldsymbol{R}}\boldsymbol{S}. \tag{3.26}$$

The shape matrix $\boldsymbol{S}$ contains the 3D coordinates of the n feature points. The matrix $\tilde{\boldsymbol{R}}$ consists of two rows of a rotation matrix and can be interpreted as the orientation of the vertical and horizontal camera axis. Assuming noisy measurements the actual rank of the measurement matrix can be larger than 3. However, since $\boldsymbol{S}$ represents a 3D shape its rank should be constrained to 3. Thus, the best possible rank-3 approximation is calculated by singular value decomposition

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T. \tag{3.27}$$

Setting all but the first three singular values to 0 gives the matrix $\hat{\boldsymbol{\Sigma}}$ and the candidates $\hat{\boldsymbol{R}}, \hat{\boldsymbol{S}}$ for $\tilde{\boldsymbol{R}}$ and $\boldsymbol{S}$ by

$$\hat{\boldsymbol{R}} = \boldsymbol{U}\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}} \tag{3.28}$$

$$\hat{\boldsymbol{S}} = \hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{V}^T. \tag{3.29}$$

That means $\boldsymbol{W} = \hat{\boldsymbol{R}}\hat{\boldsymbol{S}}$ is a valid best rank 3 solution. However, it is not unique since multiplying with any invertible matrix $\boldsymbol{A} \in \mathbb{R}^{3\times 3}$ also gives a valid solution in the form

$$\boldsymbol{W} = (\hat{\boldsymbol{R}}\boldsymbol{A})(\boldsymbol{A}^{-1}\hat{\boldsymbol{S}}) = \hat{\boldsymbol{R}}(\boldsymbol{A}\boldsymbol{A}^{-1})\hat{\boldsymbol{S}} = \hat{\boldsymbol{R}}\hat{\boldsymbol{S}}. \tag{3.30}$$

An estimate for the true $\tilde{\boldsymbol{R}}$ and $\boldsymbol{S}$ can be found by applying the linear transformation $\boldsymbol{A}$

$$\tilde{\boldsymbol{R}} = \hat{\boldsymbol{R}}\boldsymbol{A} \tag{3.31}$$

$$\boldsymbol{S} = \boldsymbol{A}^{-1}\hat{\boldsymbol{S}}. \tag{3.32}$$

This can be solved up to a rotation of the complete system (cameras and 3D shape) by enforcing orthonormality constraints for $\tilde{\boldsymbol{R}}$, i. e.

$$\boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}_2. \tag{3.33}$$

The extension to NRSfM by Bregler et al. [10] follows a similar factorization approach. The main idea is that a specific configuration of a non-rigid shape can be described by a linear combination of k basis shapes

$$\boldsymbol{S} = \sum_{i=1}^{k} \alpha_i \boldsymbol{S}_i, \tag{3.34}$$

With this factorization Eq. (3.26) is extended to

$$\boldsymbol{W} = \begin{pmatrix} \alpha_1\boldsymbol{R}_1 & \alpha_1\boldsymbol{R}_1 & \dots & \alpha_1\boldsymbol{R}_1 \\ \alpha_2\boldsymbol{R}_2 & \alpha_2\boldsymbol{R}_2 & \dots & \alpha_2\boldsymbol{R}_2 \\ \vdots & & & \vdots \\ \alpha_f\boldsymbol{R}_f & \alpha_f\boldsymbol{R}_f & \dots & \alpha_f\boldsymbol{R}_f \end{pmatrix} \begin{pmatrix} \boldsymbol{S}_1 \\ \boldsymbol{S}_2 \\ \vdots \\ \boldsymbol{S}_k \end{pmatrix}. \tag{3.35}$$

Choosing the number of shape bases as k and following the rank deficiency argumentation of [84] directly determines $rank(\boldsymbol{W}) = 3k$. The shape, rotation and weights are obtained by singular value decomposition and optimization for orthonormality constraints as in [84].

3.4 ERROR METRICS

Given an estimated human pose $\boldsymbol{X}$ and a ground truth pose $\boldsymbol{X}_{gt}$ in a coordinate-based representation (cf. Section 3.2.1) several evaluation criteria for the quality of the reconstruction can be defined. The commonly used metrics in the literature are introduced and reviewed in this section. For more details and other metrics that are not used in this thesis the reader is referred to [74].

The most intuitive error metric was first used in [75] to evaluate reconstructions of the HumanEva dataset. It is the mean euclidean distance between all joints of one pose known as *Mean Per Joint Positioning Error (MPJPE)* which is defined as

$$E_{MPJPE} = \frac{1}{j} \sum_{i=1}^{J} \|\boldsymbol{x}_i - \boldsymbol{x}_{i,gt}\|_2, \tag{3.36}$$

where j denotes the number of joints and $\boldsymbol{x}_i, \boldsymbol{x}_{i,gt}$ are the i-th joints from the predicted and ground truth poses, respectively. In some cases the global orientation of the predicted pose is irrelevant. To focus only on the pose Simo-Serra et al. [78] introduced a rigid alignment step before calculating the MPJPE. They align a predicted pose to the ground truth by finding the rotation, translation and scale which minimizes Eq. (3.36). They refer to it as *3D Pose Error (3DPE)*. Since the alignment is done by a Procrustes Analysis (PA) [24] it is also known as MPJPE-PA. This thesis uses the original naming 3DPE. The MPJPE and 3DPE give a good hint on the quality of the reconstruction. However, they average over all distances which can be misleading when only a single joint position is wrong. This is not reflected in the MPJPE. To deal with this problem the *Percentage of Correct Keypoints (PCK)* is defined as the percentage of matching keypoints that are inside a unit sphere of radius $150mm$ around each ground truth keypoint. Different radii can be regarded by calculating the *Area Under Curve (AUC)* for radii in the range of $0mm$ to $150mm$.

For other relatively uncommon metrics the reader is referred to [74].

3.5 DATASETS

Human motion capture is a very active research area for many years. Hence there exists several benchmarking datasets which are described in the following.

In 2010 Sigal et al. [75] published the first version of HumanEva which was the first dataset that contains several synchronized cameras and corresponding 3D data captured by a marker-based MoCap system. It contains 7 calibrated video sequences (4 grayscale and 3 color) that

are synchronized with 3D human body poses. Four subjects perform 6 everyday activities. The dataset is split into training, validation and test set. The subjects are wearing marker suits, the cameras are static and the background remains unchanged during all sequences. Machine learning methods trained on the images from HumanEva tend to overfit to the constrained setting and therefore struggle to transfer to real world scenarios.

The Human3.6M [33] dataset published in 2014 contains a larger number of subjects and activities. The subjects wear optical markers attached to normal clothing which makes there appearance more realistic. It contains 3.6 million 3D human poses and corresponding images. 11 professional actors (6 male, 5 female) perform 17 scenarios (e. g. *discussion*, *smoking*, *taking photo*, *talking on the phone* etc.) captured by 4 RGB cameras and a time-of-flight depth sensor. 3D Laser scans of all subjects are available. Standard evaluation protocols have prevailed in the MoCap community: the subjects 1,5,6,7,8 are used as training set and the subjects 9 and 11 as evaluation set. The two most common ones calculate the average MPJPE and 3DPE (cf. Section 3.4) for the subjects 9 and 11 per activity. Although the number of frames has significantly increased over the previous state-of-the-art the dataset is still restricted to everyday activities and a laboratory setup.

Another dataset captured in a laboratory setup is the still growing CMU dataset [11]. It contains many different activities including very uncommon motions (e. g. gymnastics) and interacting persons. Unfortunately, it contains only very low resolution grayscale images which does not allow to perform 2D keypoint detections directly on the images.

The KTH Multiview Football II dataset is captured by 3 synchronized videos of 4 sequences from a football match. Although it only contains 800 different 3D poses it is recorded in-the-wild, i. e. no manipulations such as optical markers has been done to the subjects.

MPI-INF-3DHP [52] was recorded in a MoCap studio with several synchronized time-of-flight depth sensors. This avoids the optical markers attached to the body which are required for previous approaches. 8 actors (4 male and 4 female), perform 8 activities of different complexity which leads to more diverse motions compared to Human3.6M. The actors wear 2 different sets of clothing. One set is casual everyday apparel and the other is plain-colored. The subjects were recorded in front of a green screen which enables to synthesize various backgrounds to augment images. Additionally, the appearance of the subjects is changed augmenting the plain-colored apparel with other textures.

Figure 3.7 shows exemplary images from the datasets discussed above. All available datasets to date have major limitations since they all are restricted to laboratory setups. An interesting opportunity to build a new and realistic in-the-wild dataset comes from the recent advances

in 3D reconstruction from inertial measurement devices (IMUs) [48–50]. These devices are attached to the human body and have only a negligible influence on the subjects movements. Moreover, they can be hidden under the clothing to avoid influencing the subjects appearance in the image.

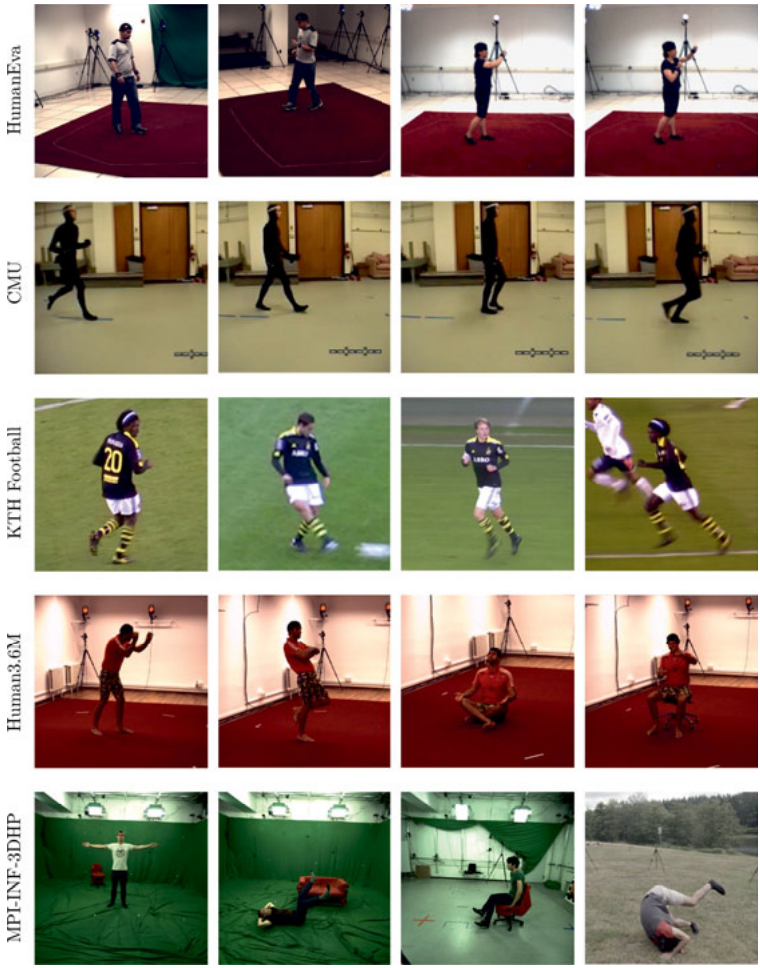


Figure 3.7: Example images from the human MoCap datasets.

4

EXPLOITING TEMPORAL PROPERTIES

Markerless human motion capture has improved constantly over the last years and is marketed in commercial products, e. g. from *Simi Reality Motion Systems* [77] and *The Captury* [82]. These systems use several calibrated cameras arranged around a predefined area in which a person can be captured with great detail. However, for outdoor or mobile applications the practicability is very limited and the costs for such systems are high. A desirable practical solution for these scenarios has a minimal number of mobile sensors. In this thesis only a single camera is used as a sensor. Reconstructing a 3D scene from a video recorded by a single camera is very challenging since the camera projects the 3D scene to a 2D plane which results in the inevitable loss of one dimension. That means, without knowledge about the scene or setting, it is impossible to recover the 3D information completely. Several solutions exist that impose priors on the camera (Section 2.1). They, however, require a large amount of camera motion which, in most cases, is not available in the scenarios mentioned above. Fortunately, several natural constraints are given by human poses (e. g. the 3D position of the elbow defines the possible positions for the hand on the same arm) and motions (e. g. movements are smooth). Identifying and formulating these constraints as an optimization problem is the main contribution of this chapter.

The following chapter proposes two methods to integrate temporal constraints into a factorization approach that is motivated by the NRSfM formulation. Sequences of human poses obey several temporal constraints. Since the velocity of human body parts is naturally restricted there is only a minor difference of poses in consecutive frames. It follows that the 3D motion over several frames is smooth. The same holds true for the camera motion. Another natural constraint is given by the fact that bone lengths of one person do not change during a sequence. These constraints are exploited in Section 4.1 using a periodic prior for periodic motions and a variance minimization for non-periodic motions. Since this approach uses a pretrained pose basis it is restricted to motions in the training dataset. To generalize to other motions knowledge about the human kinematic chain is employed in Section 4.2 by deriving a so-called *Kinematic Chain Space*. This enables the reconstruction of arbitrary skeletons as long as their structure is known. Parts of the following sections are taken from the publications [96–99].

4.1 PERIODIC AND NON-PERIODIC CONSTRAINTS

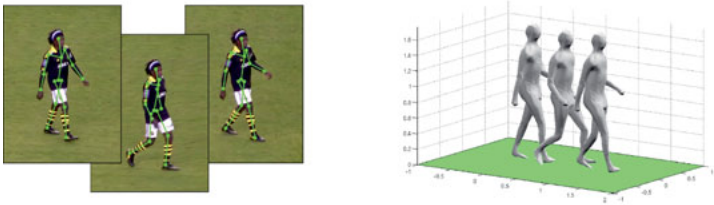


Figure 4.1: Real world scenario of KTH database [36]. Left: frames 115, 136 and 143 of Sequence 1 from Football Dataset II. Right: 3D reconstruction using our proposed method

The recovery of 3D human poses in monocular image sequences is an inherently ill-posed problem, since the observed projection on a 2D image can be explained by multiple 3D poses and camera positions. Nevertheless, experience allows a human observer to estimate the pose of a human body, even with a single eye. The purpose of this approach is to achieve a correct 3D reconstruction of human motions from monocular image sequences as shown in Figure 4.1.

The recovery of 3D structure of an object is a well studied problem in computer vision. In 1992 Tomasi and Kanade [84] proposed the first factorization approach to solve this problem for rigid objects which is well-known as the *Structure from Motion (SfM)* problem. It was later extended by Bregler et al. [10] to the non-rigid case and consequently named *Non-Rigid Structure from Motion (NRSfM)* (see Section 3.3). Since a human skeleton can be considered as an non-rigidly deforming object it is an obvious choice to solve the MoCap problem by NRSfM. Recent works considering NRSfM (e. g. [22, 23, 26]) work well as long as there is a camera rotation around the observed object. However, due to ambiguities in camera placement and 3D shape deformation they fail in realistic scenes, e. g. a static camera filming a person walking by as shown in Figure 4.3. Since there is no frontal view of the person during the whole sequence severe depth estimation errors occur. To solve this we propose to employ a pretrained human pose basis combined with smoothness regularization and bone length constancy constraints. A trilinear factorization approach similar to [23, 62, 68, 103] is used. We assume that a set of feature points on the skeleton of the person is tracked throughout the sequence. Our goal is to decompose it into three factors for camera motion, base poses and mixing coefficients. Different to [62] and [23], the second factor is kept fixed which corresponds to 3D

structure, similar to [103] and [68]. Furthermore, we propose to regularize the third factor, commonly interpreted as the mixing coefficients: Firstly, a prior well suited for periodic motion is imposed. Secondly, constraints on the limb lengths are applied. As opposed to [103] and [68] where lengths or relations of particular limbs need to be *a-priorly* known, we *constrain* the limbs lengths to be invariant.

We demonstrate that our algorithm works on motion capture data (CMU MoCap [11], HumanEva [75]) as well as on challenging real world data as for example the KTH Football Dataset [36] shown in Figure 4.25. Additionally we analyze the influence of the number of base poses and the regularization factor on the reconstruction result. Furthermore we demonstrate that our algorithm is robust to noise and also able to handle occlusions and reconstruct the occluded body parts correctly. We show that it can also be used for motion classification tasks.

The proposed method allows to correctly reconstruct 3D human motion from feature tracks in monocular image sequences with arbitrary camera motion[1]. It does not use a predefined skeleton or anthropometric constraints. Additionally it can handle occlusions and noisy data. Summarizing, the contributions are:

- A periodic model for the mixing coefficients for periodic and quasi-periodic motions such as walking is introduced.
- A novel regularization term for non-periodic motions is proposed.

Our approach consists of three main steps (see Figure 4.2). First we assume, that every 2D motion sequence can be factorized into a camera model and a series of 3D poses (Section 4.1.1), like in standard structure from motion approaches (Section 3.3). The 3D poses are composed of a linear combination of base poses, that are retrieved by a PCA on different motion databases (Section 4.1.6.1). To model periodic motion (eg. walking and running), we show that it is possible, to assume a periodic weight for the base poses to significantly reduce the number of variables, that have to be calculated (Section 4.1.3). The proposed algorithm in Section 4.1.5 is alternatingly recovering the camera matrices (Section 4.1.2) and the 3D poses. Our extension to non-periodic motion calculates the weights for the base poses for each frame. We handle the large number of variables by using a regularization term enforcing bone length constancy over time. This leads to a highly realistic 3D reconstruction of different types of non-periodic motion (Section 4.1.4).

1 Arbitrary camera motion also includes non-moving (i. e. static) cameras.

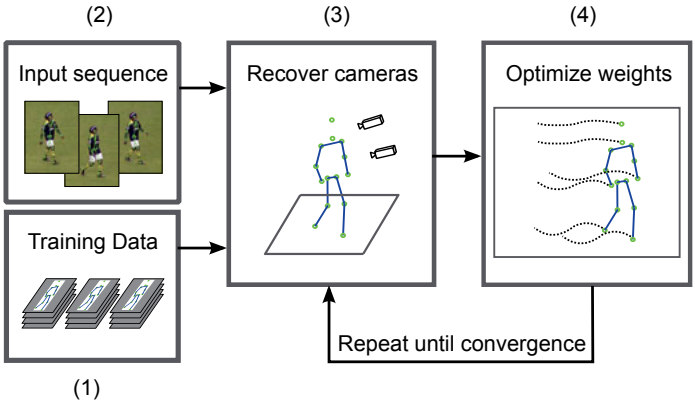


Figure 4.2: Our method. (1) 3D base poses are learned from training data. (2) Input sequence. (3) Cameras are recovered from estimated 3D poses and 2D poses. (4) Weights for base poses are calculated by minimizing the reprojection error. Steps (3) and (4) are alternated until convergence.

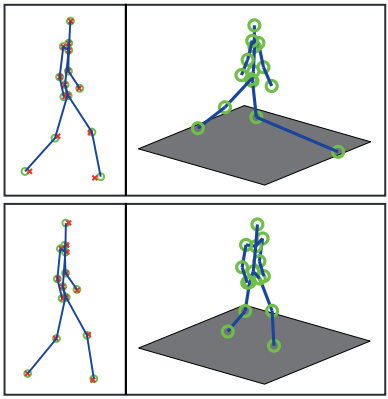


Figure 4.3: 3D reconstruction (green circles, blue lines) and ground truth data (red crosses) of a walking sequence from the CMU dataset. Top: Using a traditional NRSfM approach [26]. Most non-rigid structure from motion approaches with no rotation and unknown base poses fail, although they produce a small reprojection error (left). From other perspectives (right) a wrong reconstruction can be observed. Bottom: The proposed approach produces correct reconstructions in all views.

4.1.1 *Factorization model*

A single 3-dimensional pose $\boldsymbol{X} \in \mathbb{R}^{4\times a}$ with a joints in homogeneous coordinates can be written as a linear combination of k previously learned base poses $\boldsymbol{Q}_l \in \mathbb{R}^{4\times a}$

$$\boldsymbol{X} = \boldsymbol{Q}_0 + \sum_{l=1}^{k} \boldsymbol{\theta}_l \boldsymbol{Q}_l, \tag{4.1}$$

where $\boldsymbol{Q}_0$ is the mean pose of all poses used for training and $\boldsymbol{\theta}_l \in \mathbb{R}^{4\times 4}$ is the weight matrix for the base pose $\boldsymbol{Q}_l$. With ϑ_l as the scalar weight for the l-th base pose each $\boldsymbol{\theta}_l$ has the form

$$\boldsymbol{\theta}_l = \begin{pmatrix} \vartheta_l \boldsymbol{I}_3 & \\ & 0 \end{pmatrix}, \tag{4.2}$$

where $\boldsymbol{I}_3$ is the 3×3 identity matrix. Note that only the coordinates in the mean pose $\boldsymbol{Q}_0$ are describing a point in homogeneous coordinates, while $\boldsymbol{Q}_{1,\dots,k}$ are directions that define *deformations*. By stacking poses we can write a 3D sequence as $\boldsymbol{W} \in \mathbb{R}^{4f\times a}$ of f images, with $\boldsymbol{X}_{1,\dots,f}$ as the poses in frames $1, \dots, f$

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_f \end{pmatrix}. \tag{4.3}$$

With Eq. (4.1) we can do a factorization

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{Q}_0 + \sum_{l=1}^{k} \boldsymbol{\theta}_{l,1} \boldsymbol{Q}_l \\ \vdots \\ \boldsymbol{Q}_0 + \sum_{l=1}^{k} \boldsymbol{\theta}_{l,f} \boldsymbol{Q}_l \end{pmatrix} = \boldsymbol{\Theta} \begin{pmatrix} \boldsymbol{Q}_0 \\ \boldsymbol{Q}_1 \\ \vdots \\ \boldsymbol{Q}_k \end{pmatrix} = \boldsymbol{\Theta}\boldsymbol{Q}, \tag{4.4}$$

where $\boldsymbol{\Theta} \in \mathbb{R}^{4f\times 4k}$ contains the weight matrices $\boldsymbol{\theta}_l$.

The projection of a 3D pose $\boldsymbol{X}_i$ in the i-th frame to a 2D pose $\boldsymbol{X}_{i,2D} \in \mathbb{R}^{2\times a}$ is done by the camera matrix $\boldsymbol{K}_i \in \mathbb{R}^{2\times 4}$

$$\boldsymbol{X}_{i,2D} = \boldsymbol{K}_i \boldsymbol{X}_i. \tag{4.5}$$

To project the whole 3D sequence described by the matrix $\boldsymbol{W}$, the camera matrix $\boldsymbol{K} \in \mathbb{R}^{2f \times 4f}$ is used. Let $\boldsymbol{K}$ be a sparse block diagonal matrix

$$\boldsymbol{K} = \begin{pmatrix} \boldsymbol{K}_1 & & \\ & \ddots & \\ & & \boldsymbol{K}_f \end{pmatrix}. \tag{4.6}$$

The factorization of a 2D sequence given by the matrix $\boldsymbol{W}_{2D} \in \mathbb{R}^{2f \times a}$ can now be written as

$$\boldsymbol{W}_{2D} = \boldsymbol{K}\boldsymbol{\Theta}\boldsymbol{Q}. \tag{4.7}$$

When dealing with missing feature points (for example caused by partly occluded body parts) the equations corresponding to these feature points can be excluded from the optimization. This is further explained and evaluated in Section 4.1.6.7. This model is very similar to the models proposed by [87], [62] and [23]. While they are fixing $\boldsymbol{\Theta}$ and optimize for $\boldsymbol{K}$ and $\boldsymbol{Q}$, our approach is using a previously learned $\boldsymbol{Q}$ and optimize for the weights $\boldsymbol{\Theta}$ like [68] and [103] did for single images.

4.1.2 *Camera Parameter Estimation*

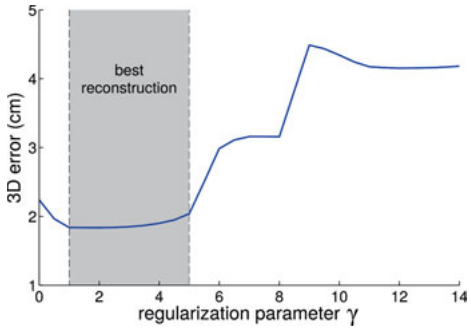


Figure 4.4: Influence of the camera path regularization on the reconstruction result. A low value for the regularization parameter γ avoids flips while a high value enforces a static camera. The best results are obtained for values between 1 and 5.

To reconstruct the camera parameters we are assuming a weak perspective camera. The pose in the i-th frame $\boldsymbol{W}_{2D}^i$ can be factorized with the above notation as

$$\boldsymbol{W}_{2D}^i = \boldsymbol{K}_i\boldsymbol{\Theta}_i\boldsymbol{Q}, \tag{4.8}$$

where $\boldsymbol{\Theta}_i \in \mathbb{R}^{4\times 4k}$ denotes the weight matrix for this frame. For the estimation of the camera parameters we assume the 3D pose described by $\boldsymbol{\Theta}_i\boldsymbol{Q}$ to be known. The solution for the camera matrices for each frame can be obtained by least squares minimization of the reprojection error

$$\min_{\boldsymbol{K}_i} \left|\left| \boldsymbol{W}_{2D}^i - \boldsymbol{K}_i\boldsymbol{\Theta}_i\boldsymbol{Q} \right|\right|_F . \tag{4.9}$$

In our model each $\boldsymbol{K}_i$ describes a weak perspective camera. Therefore we give the optimization algorithm used to solve Eq. (4.8) correct starting values for $\boldsymbol{K}_i$ which satisfy the constraints for a weak perspective camera. We rewrite Eq. (4.8) with $(\boldsymbol{\Theta}_i\boldsymbol{Q})^+$ as the right-inverse of $\boldsymbol{\Theta}_i\boldsymbol{Q}$

$$\boldsymbol{K}_i = \boldsymbol{W}_{2D}^i(\boldsymbol{\Theta}_i\boldsymbol{Q})^+. \tag{4.10}$$

The scale parameter s of the weak perspective camera can be determined by

$$s = \frac{1}{2}\sqrt{||\boldsymbol{K}_{i,1}||^2 + ||\boldsymbol{K}_{i,2}||^2}, \tag{4.11}$$

with $\boldsymbol{K}_{i,1}$ as the first row and $\boldsymbol{K}_{i,2}$ as the second row of $\boldsymbol{K}_i$. We receive an unscaled camera matrix by dividing $\boldsymbol{K}_i$ by s. Next we orthonormalize the first 2×3 block of the unscaled matrix with the help of a singular value decomposition, where all singular values are set to 1. Recombining the orthonormalized block with the scale s and the last column of the unscaled camera matrix gives a good estimation for the starting values.

If we reconstruct the cameras for each frame separately the camera orientations can *flip*. i. e. the camera matrix of the flipped camera not only describes a weak perspective projection but also a reflection at the origin of the coordinate system. As this effect rarely occurs it can be easily avoided by penalizing rapid changes in the camera path. Therefore, we propose a regularization term that calculates the difference between the current camera matrix $\boldsymbol{K}_i$ and the previous camera matrix $\boldsymbol{K}_{i-1}$

$$r_{K,i} = \gamma \left|\left| \boldsymbol{K}_i - \boldsymbol{K}_{i-1} \right|\right|_F , \tag{4.12}$$

with γ as regularization parameter.

The whole minimization problem can now be written as

$$\min_{\boldsymbol{K}_i} \left|\left| \boldsymbol{W}_{2D}^i - \boldsymbol{K}_i\boldsymbol{\Theta}_i\boldsymbol{Q} \right|\right|_F + r_{K,i}. \tag{4.13}$$

While the regularization term also allows smoothing of the camera path, its sole purpose is to avoid camera flips. The regularization is not necessary in most cases as the flips only occur very rarely. Setting the parameter γ to a high value would result in a static camera. Therefore we set γ to a very low value where it avoids flips and only slightly effects the camera path as shown in a small experiment in Figure 4.4. Although the

reconstruction error without the camera regularization ($\gamma = 0$) seems low there are three flips in the camera path causing wrong 3D reconstructions. In contrast to Zhu et al. [115] who solved the problem by using keyframes, we do not assume any prior camera positions or poses.

Considering the entries in

$$\boldsymbol{K}_i = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \end{pmatrix} \tag{4.14}$$

we can enforce a weak perspective camera by exploiting the fact that the first 2×3 block in $\boldsymbol{K}_i$ consists of two rows of a rotation matrix scaled by s, as described in Section 3.1.4. The property

$$\boldsymbol{K}_i \boldsymbol{K}_i^T = s^2 \boldsymbol{I}_2 \tag{4.15}$$

gives the constraints

$$m_{11}^2 + m_{12}^2 + m_{13}^2 - (m_{21}^2 + m_{22}^2 + m_{23}^2) = 0 \tag{4.16}$$

and

$$m_{11} m_{21} + m_{12} m_{22} + m_{13} m_{23} = 0. \tag{4.17}$$

4.1.3 *Periodic Motion*

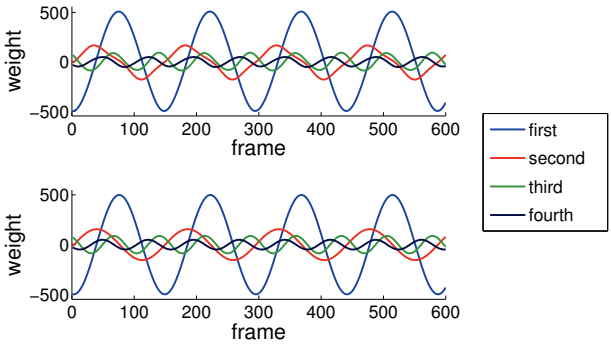


Figure 4.5: Comparison of ground truth coefficients of the first four base poses (top) with fitted periodic function (bottom) using the dataset of Troje [88].

With the camera matrix $\boldsymbol{K}$ calculated as described in Section 4.1.2 the weights $\boldsymbol{\Theta}$ for the base poses can be reconstructed. Trying to optimize the reprojection error for all variables in $\boldsymbol{\Theta}$ fails, as there are to many degrees of freedom. For periodic motion the number of unknowns can be

reduced by using a sine function to model the temporal behavior of the weights in $\boldsymbol{\Theta}$.

Figure 4.5 shows the weights of the first four base poses of a gait sequence and the corresponding fitted sine functions. For this specific sequence the mean absolute error of the periodic reconstruction to the ground truth data is $2.05mm$. It verifies the results obtained by N. Troje in [88, 89]. They used the same periodic assumption to describe human gait patterns and did an extensive research on a large set of persons. These observations can be made with running motions as well. So the periodic assumption appears to be appropriate for periodic motion.

As shown in Section 4.1.1 the number of unknowns in $\boldsymbol{\Theta}$ equals fk. By modelling the temporal behavior of ϑ as

$$\vartheta(t) = \alpha \sin(\omega t + \varphi) \tag{4.18}$$

the number of unknowns can be decreased to $3k$. Note that the number of variables does not depend on the number of frames anymore yet only on the number of base poses. We can thus minimize the 2D reprojection error

$$\min_{\alpha,\omega,\varphi} ||\boldsymbol{W}_{2D} - \boldsymbol{K}\boldsymbol{\Theta}\boldsymbol{Q}||_F \,. \tag{4.19}$$

Note, that the objective function in Eq. (4.19) is nonlinear and non-convex.

The use of sine functions to approximate human motion was firstly proposed by Troje et al. [88, 89]. We use a similar representation in Eq. (4.18) which can be motivated from [2], since a sine function can be represented by a linear combination of DCT bases. Modeling a structure from motion problem in *trajectory space* using DCT bases, requires a manually set or estimated number of DCT bases which mostly results in too many degrees of freedom. In Figure 4.3 we show that 3D reconstructions of approaches derived from [2] (e. g. Gotardo and Martinez [26]) fail when there is no sufficient camera motion in the sequence (i. e. low reconstructibility as defined by Park et al. [62]). Combining the use of a single sine function as weight as proposed by Troje in [88, 89] with trained base poses results in a low number of variables and plausible 3D reconstructions.

4.1.4 *Non-Periodic Motion*

To model non-periodic motion, periodic functions for the weights of the base poses are not applicable anymore. Trying to optimize all weights at once without constraints gives good results for the 2D reprojection, but does not ensure a realistic 3D reconstruction. Figure 4.6 shows the temporal behavior of the bone lengths using the unconstrained opti-

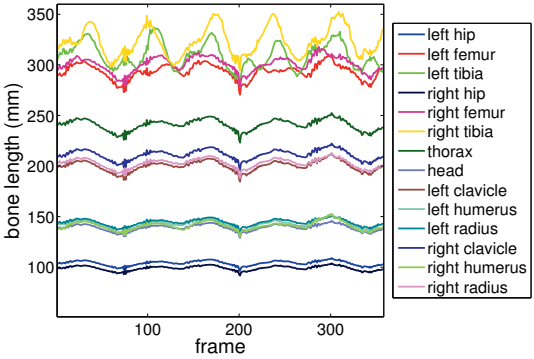


Figure 4.6: Temporal behavior of bone lengths obtained by unconstrained optimization. The maximal variation is about 40mm. Computed on CMU MoCap (subject7/walk1).

mization. There are variations in lengths up to $40mm$. This is caused by a slightly wrong initial camera position, which the optimizer later tries to compensate by weighting base poses wrongly. It results in a 3D reconstruction where unrealistic bone length changes occur. To compensate this we propose a regularization term, which holds the bone lengths constant over time. Different to [68] and [103] we are not using bone length constraints. Such a constraint would restrict the model to a particular person.

The length of a bone is defined by the euclidean distance between the 3D joint coordinates of that bone. These can be directly obtained from the 3D reconstruction described by $\boldsymbol{\Theta Q}$. We denote the length of bone s as

$$b_s = ||\boldsymbol{j}_{s,2} - \boldsymbol{j}_{s,1}||_2 \,, \tag{4.20}$$

where $\boldsymbol{j}_{s,1}$ and $\boldsymbol{j}_{s,2}$ are the coordinates of the endpoints of that bone. We want to hold the bone lengths nearly constant over time to ensure a realistic reconstructed skeleton, but do not want to be too restrictive to the optimizer. In other words the bone lengths should not change much. In the optimal case they are not changing at all. We are using the variance of the length changes over time of each bone as a measure. To build the regularization term r_B, we sum the variances $\mathrm{Var}(\bullet)$ of all bone lengths over time

$$r_B = \beta \sum_i \mathrm{Var}(b_i), \tag{4.21}$$

with β as the regularization parameter. This regularizer holds the bone length constant but is not fixing it to a specific value. Note, that the same variance for a short bone allows larger relative changes in length than for longer bones. Using the relative variance, i. e. normalizing $\mathrm{Var}(b_i)$ by the mean of the bone length avoids this effect. However, as experimentally shown in Figure 4.11 there is no significant difference in using the variance or the relative variance. Due to this finding and to keep computational effort as low as possible, all experiments are using Eq. (4.21) as regularizer.

The optimization problem can be written as

$$\min_{\boldsymbol{\Theta}} ||\boldsymbol{W}_{2D} - \boldsymbol{K}\boldsymbol{\Theta}\boldsymbol{Q}||_F + r_B. \tag{4.22}$$

For the minimization of Eq. (4.19), the parameters α, ω and φ of the functions defined by Eq. (4.18) are estimated. Here, for minimizing the nonlinear and nonconvex objective function in Eq. (4.22) we can estimate the coefficients $\boldsymbol{\Theta}$ of the linear combination $\boldsymbol{\Theta}\boldsymbol{Q}$ subject to the constraints defined by Eq. (4.21) since $\boldsymbol{Q}$ defines the prior knowledge on the possible deformations of human shapes.

The number of variables equals fk, i. e. it linearly depends on the number of frames f. Using a skeleton with 15 joints gives the same number of 2D/3D point correspondences per frame. By keeping the number of used base poses k low there are more equations than unknowns.

4.1.5 *Algorithm*

To estimate the $f+1$ sets of variables $\boldsymbol{K}_1, \ldots, \boldsymbol{K}_f$ and $\boldsymbol{\Theta}$ we alternatingly optimize for each of the sets while keeping the others fixed. The optimization of each camera matrix $\boldsymbol{K}_j$, $j = 2, \ldots, f$, requires the regularization terms $r_{K,j}$ and $r_{K,j+1}$. If we use central differences in Eq. (4.12), we need to optimize all the sets $\boldsymbol{K}_j$, $j = 1, \ldots, f$, simultaneously. Using the proposed forward differences allows to sequentially estimate them, i. e. given $\boldsymbol{K}_1$ we estimate $\boldsymbol{K}_2$, then $\boldsymbol{K}_3$ etc. The precision of the estimated solution is hardly affected while the computation time in our experiments reduces by the factor 5. Shape parameters are estimated by minimizing Eq. (4.19) in the case of periodic motion and Eq. (4.22) in the case of non-periodic motion, respectively. These constrained nonlinear and nonconvex problems are optimized using a second-order gradient descent algorithm.

In the first iteration we use the mean pose as initialization. This means setting all values in $\boldsymbol{\Theta}$ to zero except the ones weighting the mean pose $\boldsymbol{Q}_0$. With that the initial cameras are estimated framewise as described in Section 4.1.2. The optimization for the weights of the base poses follows. This step is depending on whether we are using the periodic (Section

4.1.3) or the non-periodic model (Section 4.1.4). The last two steps are repeated until the reprojection error is not changing anymore.

Alternating the estimation of the parameter sets can be seen as a variant of a block-coordinate descent by formulating one objective function for all parameters:

$$f(\boldsymbol{K}_1, \ldots, \boldsymbol{K}_f, \boldsymbol{\Theta}) = f(\boldsymbol{\Theta}) + \sum_{i=2}^{f} g_i(\boldsymbol{K}_i), \tag{4.23}$$

where

$$f(\boldsymbol{\Theta}) = \quad ||\boldsymbol{W}_{2D} - \boldsymbol{K}\boldsymbol{\Theta}\boldsymbol{Q}||_F + r_B \tag{4.24}$$

$$g_i(\boldsymbol{K}_i) = \quad r_{K,i}. \tag{4.25}$$

The objective function for the periodic reconstruction can be formulated in the same way. Convergence of coordinate gradient descent is guaranteed if the joint objective function is strongly-convex [46]. More recently, results on convergence were established if at least one of the terms is convex (see, e. g. [90]). Since neither of the terms in Eq. (4.23) is convex and they are optimized alternatingly, convergence cannot be guaranteed. However, we will experimentally show that the proposed algorithm converges to a reasonable local minimum in Section 4.1.6.5.

Algorithm 1 Recover camera and shape

```
Q ← base shapes
while no convergence do
  for t = 1 → f do
    calculate starting values for K_t
    optimize ||w^i_2D − K_t Θ_i Q||_F + r_K,i
    insert K_t in K
  end for
  optimize ‖W − KΘQ‖_F + r_B
end while
```

4.1.6 *Experimental Results*

To evaluate our method, we were using three different databases: CMU MoCap [11], HumanEva [75] and KTH Football [36]. We trained base poses (see Section 4.1.6.1) of different motion categories, for example walking, jogging, running and jumping to demonstrate the generality of our method. The motions and datasets used for training vary for the different experiments and will be named in the respective sections.

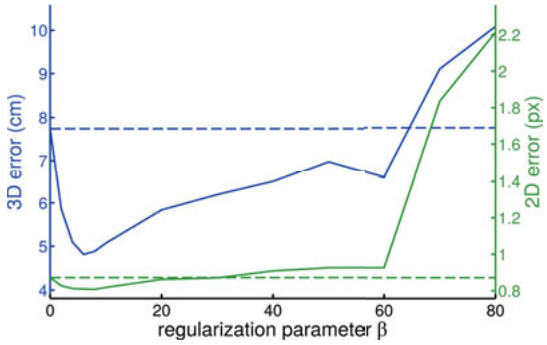


Figure 4.7: 2D reprojection error and 3D reconstruction error with different regularization parameter β. While the 2D error is not changing much or getting worse, the 3D error gets significantly better at most parameter values. Computed on CMU MoCap (subject35/walk1). Qualitatively there is no difference between different motion categories.

Instead of the 2D reprojection error a 3D error e as evaluation criterion is defined

$$e = \frac{1}{f} ||\boldsymbol{W}_{in} - \boldsymbol{W}_{rec}||_F \,, \tag{4.26}$$

with $\boldsymbol{W}_{in}$ as the ground truth 3D data and $\boldsymbol{W}_{rec}$ as the reconstruction. To compare sequences of different lengths, we are dividing the error by the number of frames f. Note that this error is not the MPJPE and 3DPE as described in Section 3.4. Since the proposed algorithm is similar to traditional NRSfM approaches it uses the same evaluation criterion (Eq. (4.26)) for comparability. Evaluation in terms of MPJPE and 3DPE is done in Section 4.2.6. Although the reprojection error is a common metric for the quality of the results produced by many SfM approaches it is a bad criterion for judging a 3D reconstruction. Therefore, it is important to use the 3D error instead of the reprojection error when evaluating 3D reconstructions. For example with our bone length regularizer we achieve a worse reprojection error but a significantly better 3D reconstruction (see Figure 4.7). While the reprojection error remains nearly constant for values of the regularization parameters up to 60, the 3D error is getting better. Only for very high values both errors are getting worse. This is further evaluated in Section 4.1.6.4.

4.1.6.1 *Learning base poses*

For learning the base poses we were using different databases: the well-known CMU Motion Capture Database [11], the HumanEva dataset [75] and as a real world example the KTH Football Dataset II [36]. These

three databases are using slightly different joint annotations, so it is important to learn the base poses for each database separately.

We are learning the base poses by stacking pose vectors of all frames and executing a PCA on this matrix. For each of the used motion categories a linear combination of the first ten eigenvectors obtained by the PCA is enough to cover more than 99% of the variance in the dataset. It is also possible to learn base poses for multiple motions at once. If doing so, the number of base poses should be increased to be able to fully cover all possible motions. The influence of the used number of base poses on the reconstruction result is evaluated later in Section 4.1.6.3.

4.1.6.2 *Periodic Motion*

As shown in Section 4.1.3, the number of unknowns can be reduced when using periodic base functions. This results in a much faster solving of the optimization problem. Figure 4.21 shows some frames of a reconstruction of a gait sequence by just using four base poses. Even with only 12 unknowns to optimize the reconstruction is close to the real 3D data. Note that the number of variables does not depend on the number of frames. That means that the computational effort does not increase much if longer sequences are used as long as the motion does not change. The reconstruction of the shown sequence of 450 frames took about 15 seconds, which is about two magnitudes faster than the non-periodic reconstruction on the same sequence. For periodic motion this method is a fast and efficient way for the 3D reconstruction. Comprehensive results of the periodic reconstruction on different periodic motions can be seen in Section 4.1.6.6.

If bone length constancy is used to additionally regularize the reconstructions we observed no improvement. The reason is that the periodic assumption is such a strong prior that an additional regularization term has no effect. Setting the weight of the bone length regularizer too high results in a local minimum where the skeleton is not moving at all and stays in the mean pose.

4.1.6.3 *Number of base poses*

One of the main questions is how many base poses should be used to achieve a good reconstruction. More base poses can model more deformation but using too many can cause unnatural deformation.

It is important to notice that all motions used for training lie in the space spanned by the base poses. However, not every linear combination of the base poses defines a correct human pose. In fact, every base pose allows for some non-human deformations. Thus the more base poses are used for the reconstruction, the more distorted the reconstruction gets.

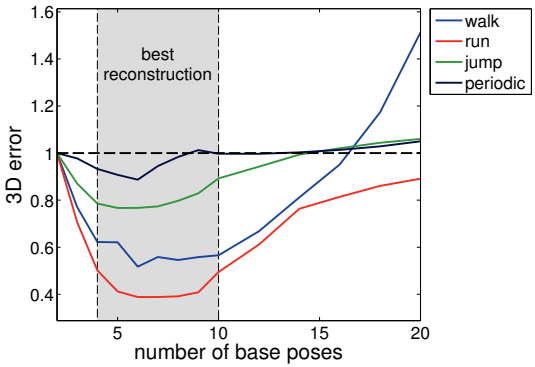


Figure 4.8: Influence on the number of used base poses on the 3D error using the non-periodic reconstruction (labels: walk, run, jump) and the periodic reconstruction (label: periodic). The number of used base poses is crucial for a good 3D reconstruction. Using more than 10 base poses for each motion category worsens the reconstruction error. For better visibility, the errors are normalized on the 3D error when using 2 base poses. The periodic reconstruction is done on the same walking sequence as the non-periodic reconstruction. Computed on CMU MoCap (subject35/walk2/run1, subject13/jump1).

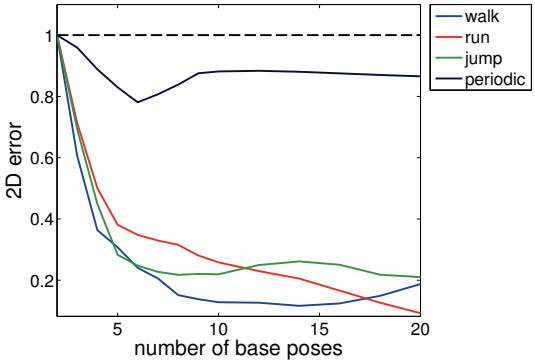


Figure 4.9: Influence on the number of used base poses on the 2D error using the non-periodic reconstruction (labels: walk, run, jump) and the periodic reconstruction (label: periodic). The 2D error decreases when more base poses are used. For better visibility, the errors are normalized on the 2D error when using 2 base poses. The periodic reconstruction is done on the same walking sequence as the non-periodic reconstruction. Computed on CMU MoCap (subject35/walk2/run1, subject13/jump1).

As shown in Figure 4.8 using 4 to 10 base poses results in the best reconstructions for periodic and non-periodic motions. On the test datasets six base poses appear to be the optimum with respect to the 3D error. If too many base poses are used the reconstruction deteriorates, whereas the reprojection error reduces. Comparing Figure 4.8 to Figure 4.9 shows the correlation between the 2D error and 3D error for the same sequences.

4.1.6.4 *Influence of regularization*

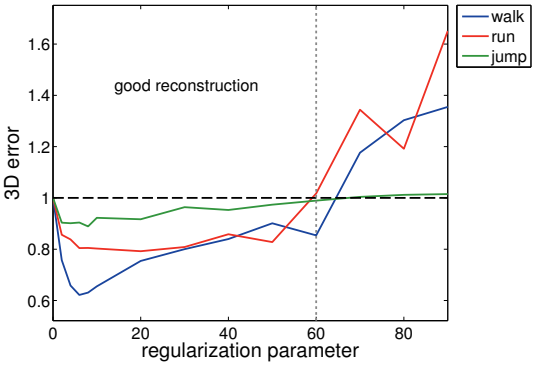


Figure 4.10: Influence of the regularization parameter β on the normalized 3D error. In a wide range, the reconstruction improves (left of dotted line) if the regularizer is used as compared to optimization without it ($\beta = 0$). Computed on CMU MoCap (subject7/walk1/run1, subject13/jump2).

Figure 4.10 shows the influence of the regularizer on the 3D reconstruction for the motion categories walk, run and jump. For better comparability the error is normalized for each motion class on the error value without regularization. Even a small value for the parameter causes a significant improvement of the 3D reconstruction. In a wide range of parameter settings the reconstruction is much better with the regularizer than without it. The selection of values for the regularization factor is crucial. If the value is too high, the reconstruction is getting worse. Using a too strong factor causes the reconstruction to not move at all over time. This is an expectable behavior in the sense of constant bone lengths, but unwanted for a realistic 3D reconstruction.

A comparison of the temporal behavior of the bone lengths of the same sequence with different values for the regularization factor is shown in Figure 4.11. The bone lengths of the periodic reconstruction (first image) are fluctuating heavily. The second image shows the best non-periodic reconstruction in terms of the 3D error. The fluctuation is less than the

one of the periodic reconstruction. The maximal difference in bone length is about 8mm. Considering possible noisy measurements, this should be an acceptable value. On the third image the bone lengths are not changing much, but the 3D error is larger than in the second.

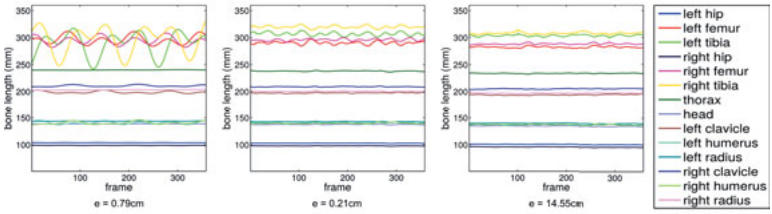


Figure 4.11: Comparison of the temporal behavior of the bone lengths with different regularization factors. First: periodic reconstruction with 3D error of 0.791cm. Second: Non-periodic reconstruction with best 3D error of 0.213cm. Third: Non-periodic reconstruction with very high regularization factor. Bone lengths are nearly constant over time but the 3D error of 14.553cm is larger.Computed on CMU MoCap (subject35/walk1).

4.1.6.5 *Convergence and stability*

As stated in Section 4.1.5 the alternatingly optimized objective functions are nonlinear and nonconvex for the periodic and non-periodic case, respectively. Thus we cannot prove convergence of the proposed algorithm. Instead we demonstrate it experimentally. Figure 4.12 shows the mean and standard deviation of the 2D error during the first 10 iterations of 5 different subjects of the CMU MoCap database. An odd step refers to camera estimation while an even step refers to pose estimation. All experiments done during the evaluation (including those in Figure 4.12) are converging to a plausible local minimum and the value of the 2D error decreases in every step.

As all nonconvex optimization algorithms the proposed algorithm is sensitive to initialization. When initialized with bad starting values it converges to a bad local minimum. As described in Section 4.1.5, initialization is done by the mean pose of the corresponding motion category which is an appropriate assumption. However, it is reasonable to evaluate the stability of the algorithm with bad or noisy initializations. Figure 4.13 shows the mean and standard deviation of the 3D error with Gaussian noise added onto the starting values. Up to a noise level of 10% the reconstructions still look plausible and close to the reconstructions without noise. Above 10% the 3D reconstructions degenerate to unrealistic poses.

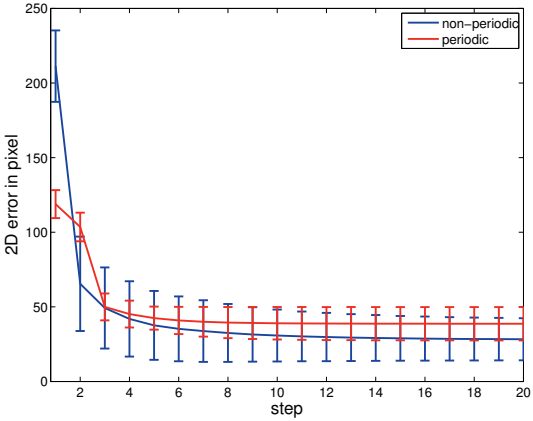


Figure 4.12: Mean 2D error and standard deviation for periodic and non-periodic reconstruction of the CMU dataset (subjects 7,9,13,16,35). Evaluated on 57 different sequences including the motion categories walk, run and jump. Odd steps refer to camera estimation while even steps refer to pose estimation.

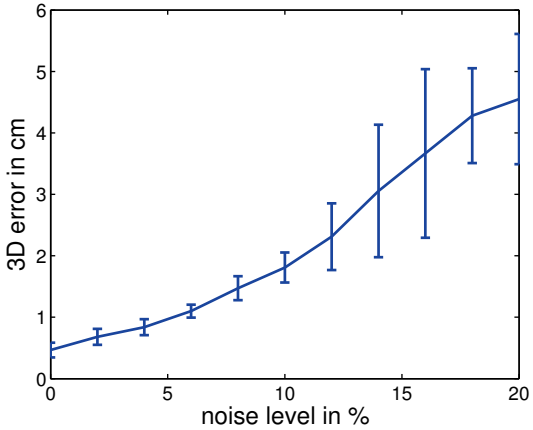


Figure 4.13: Mean 3D error and standard deviation of 57 different sequences of the CMU dataset (subjects 7,9,13,16,35) obtained by optimization with noisy starting values. The noise level is given in percent of body size of the respective subject.

4.1.6.6 *Different Motion classes*

In this section the algorithm is trained on multiple motion classes simultaneously including periodic (walking, running, jogging) and non-periodic motions (jump up/forward). Different datasets are used (CMU MoCap [11], HumanEva [75], KTH Football [36]). The ground truth for the CMU Mocap and the HumanEva datasets are generated from marker based motion capture data of humans performing different actions. The KTH Football dataset contains video sequences with manually labeled joints. The 3D reconstruction which we use as ground truth data was computed using a multi camera system. Overall this dataset is more noisy than the other two datasets and offers a real world scenario. Table 4.1 shows the 3D reconstruction error of our different methods on some of the used motion sequences compared to the results of Gotardo and Martinez [26] and Bregler et al. [10]. It is noticeable that the reconstruction results of the jumping sequences are worse compared to the other sequences. The reason is that the variance between jumping motions of different persons is much larger than between walking motions. So a new (not trained) jumping motion is insufficiently explained by the base poses, while every new walking pattern is very similar to those in the training data. Nevertheless the reconstructions appear realistic (Figure 4.23). All results except the row labeled "np all" are obtained by training on the specific motion categories. When training all motions at once (here we are using walk, run, jog, jump up, jump forward) to get more general base poses, the results are getting worse but stay realistic and are still superior to [10] and [26]. The results of [10] and [26] are obtained with the source code provided by the authors.

Table 4.1: Average 3D reconstruction error in *cm* on the CMU dataset (walk, run, jump), HumanEva walking dataset (HE) and KTH Football dataset. First row: reconstruction with periodic constraints. Second row: non periodic reconstruction without bone length regularizer. Third row: Best reconstruction result achieved with bone length regularizer. Fourth row: best result when using all motions for training simultaneously. Fifth and Sixth row: comparison to other approaches.

Method	walk	run	jump	HE	KTH
periodic	0.784	0.968	-	1.200	0.357
np ($\beta = 0$)	0.295	0.661	1.226	0.564	0.292
best	**0.183**	**0.523**	**1.090**	**0.423**	**0.187**
np all	0.334	2.805	1.313	-	-
[10]	4.557	10.821	8.531	17.824	4.427
[26]	16.359	11.395	17.139	5.714	14.673

Our 3D reconstructions are highly realistic, which was shown by surveying the 3D error. Figures 4.21, 4.22, 4.23, 4.24 show reconstructed motions taken from the CMU MoCap database. Figure 4.21 uses the periodic reconstruction with only 4 base poses. Figure 4.22, 4.23 and 4.24 are using the non-periodic approach.

4.1.6.7 *Occlusions*

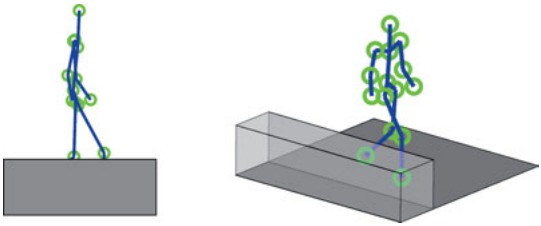


Figure 4.14: Left: Observation data of a person walking behind a box. The legs are partly occluded. Right: 3D reconstruction of occluded body parts using our method.

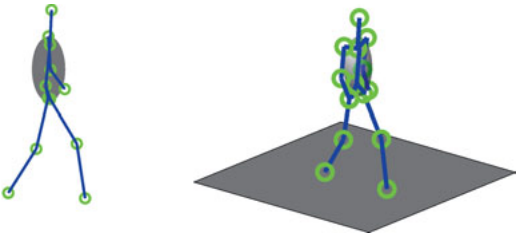


Figure 4.15: Left: During the whole sequence the left hand is occluded by the body. Right: 3D reconstruction of occluded body parts using our method.

In realistic scenes, body parts can be occluded. This happens for example if parts of the observed person are behind an object, for instance as shown in Figure 4.14. Another common case is self-occlusion where one body part occludes another body part. The integration of occlusions in our algorithm is simple. Since we are using the Frobenius norm of the reprojection error it is possible to set occluded values to zero in the observation matrix $\boldsymbol{W}_{2D}$ and the reprojection $\boldsymbol{K\Theta Q}$ while using the same objective function (Eq. (4.19) or Eq. (4.22)) as in the non occluded case. This equals to canceling the corresponding equations in the objective function.

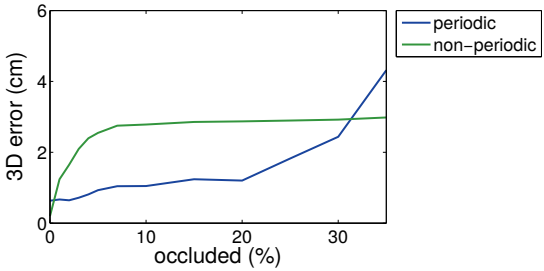


Figure 4.16: Comparison of the 3D error of periodic and non-periodic reconstruction with randomly occluded data points. The periodic reconstruction appears to be more stable as it puts a smoothness constraint on the reconstruction. Computed on CMU MoCap (subject35/walk1)

Figure 4.14 shows a person walking (CMU MoCap, subject7/walk2) behind an artificial box so that the legs cannot be seen in the input data. Our algorithm is able to reconstruct a realistic leg motion that is very close to the original motion. Figure 4.15 shows the problem of self occlusion. In the whole sequence, the back arm (shoulder, elbow and hand) is fully occluded, i. e. 20% of the input data is unknown. On the right of Figure 4.15 the back arm is correctly reconstructed by our method.

For further evaluation of the occlusion handling we randomly delete data points in the input data. Figure 4.16 shows the 3D error of the periodic and non-periodic reconstruction. While the non-periodic reconstruction produces a high 3D error for occlusions higher than 3%, the periodic reconstruction benefits from the smoothness constraint it puts on the reconstruction and remains stable for occlusions up to 20%. For visualization purposes only a single sequence is chosen in Figure 4.16. Other sequences give qualitatively very similar results.

4.1.6.8 *Noise stability*

To evaluate the stability of our method we put additional noise on the 2D input data. Figure 4.17 shows the 3D reconstruction error with respect to the noise level for the periodic and non-periodic reconstruction. In this case 5% noise means Gaussian noise with a standard deviation of 5% of the maximal range of motion of the most moving 2D point. With a very high noise level the reconstruction is still good. Apparently the periodic reconstruction appears to be more stable than the non-periodic reconstruction, because it puts a strong smoothness constraint on the weights $\boldsymbol{\Theta}$ of the base poses. The result is still a smooth motion as shown in Figure 4.18. While the non-periodic reconstruction (center) is getting

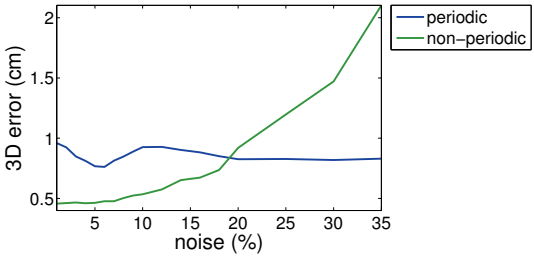


Figure 4.17: Influence of additional noise on periodic and non-periodic reconstruction. While the 3D error of the non-periodic reconstruction raises, the error for the periodic reconstruction remains nearly constant.

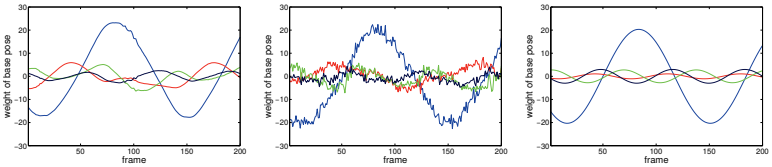


Figure 4.18: Comparison of the weights for the base poses. Left: Ground truth weights. Center: Non-periodic reconstruction with 20% noise (3D error: 1.09cm). Right: Periodic reconstruction with 20% noise (3D error: 1.02cm). Computed on the first 200 frames of CMU MoCap (subject7/walk1).

unstable the periodic reconstruction (right) still achieves a realistic output compared to the ground truth data (left). As in Figure 4.16 only a single sequence is chosen in Figure 4.17. Other sequences give qualitatively very similar results.

4.1.6.9 *Classification*

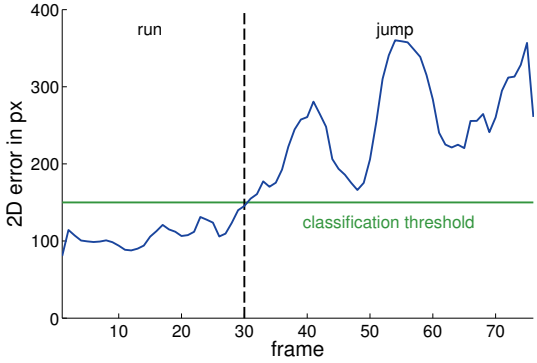


Figure 4.19: 2D error using the periodic reconstruction. The base poses are trained from the CMU running sequences (35/17-26). Reconstructing poses belonging to the jumping motion results in a large 2D error.

Figure 4.20: Combined running and jumping sequence from [28]. The first two frames are reconstructed using the periodic reconstruction, the others are using the non-periodic reconstruction. Although the base poses are trained on another dataset that does not contain this specific motion, the reconstruction is not perfect but realistic.

We also used our proposed method for classification of a mixed motion. In this example we reconstruct the outdoor sequence from [28] of a person running and jumping over an obstacle (Figure 4.20). For the

classification the reconstruction is done for 10 frames wide sections over the whole sequence. Figure 4.19 shows the corresponding 2D error when using the periodic reconstruction with base poses trained from the CMU running sequences (35/17-26). The 2D error increases for non-trained motions, as these can not be reconstructed with the used base poses. In this example the jump over the obstacle around frame 30 can be clearly seen. By setting a threshold for the 2D error a classification in running and non-running motion is possible. For the jumping part of the sequence we may therefore switch from the periodic reconstruction (cf. Section 4.1.3) to the less constrained non-periodic algorithm (cf. Section 4.1.4). Since there is no similar jumping motion in the other datasets (only jumping with legs closed or on one leg), we use base poses trained on the motions walk, run and jump simultaneously as mentioned in Section 4.1.6.6. Although the example sequence is manually labeled and the base poses are trained on another dataset our method achieves realistic results as shown in Figure 4.20.

4.1.7 *Conclusion*

This section presented a new method for the 3D reconstruction of human motion from monocular image sequences. Using periodic functions to model the weights of the base poses turned out to be very effective and stable for periodic motions. Reconstruction of non-periodic motion was successfully done with the new regularization term. In contrast to state-of-the-art methods for estimation of nonrigid shapes from monocular image sequences (e. g. [10, 26]), the proposed regularizations enable to reconstruct plausible human motion even under low reconstructibility. Generalization is shown on multiple benchmark datasets with different motion types. It even performs well under occlusions, noise and on the real world data of the KTH dataset as well as on the outdoor obstacle jump sequence. The main reason for the robustness are the learned base poses. Even if strong noise or occlusions distort the 2D detections the closest matching point from the base pose subspace is a valid human pose. However, this is also a drawback of the proposed approach. Motions with a slight deviation from the learned motion can only be reconstructed as the learned motion, e. g. walking with hands above the head, will always be reconstructed as a standard walking pattern as long as it is not in the training set. A possible solution to this problem is given in the next section.

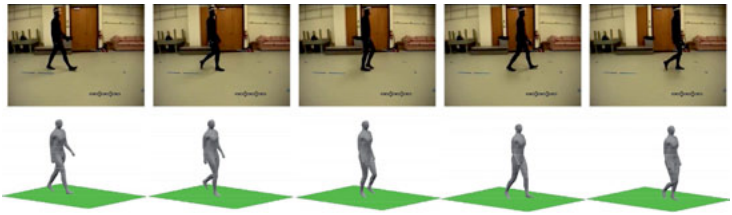


Figure 4.21: Walking sequence 35/02 of the CMU MoCap dataset reconstructed with the periodic reconstruction using only 4 base poses

Figure 4.22: Running sequence 35/17 of the CMU MoCap dataset reconstructed with the non-periodic reconstruction

Figure 4.23: Jumping sequence 13/11 of the CMU MoCap dataset with the non-periodic reconstruction

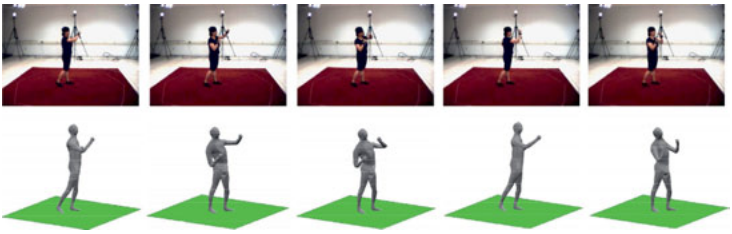


Figure 4.24: Boxing sequence of the HumanEva dataset with the non-periodic reconstruction.

4.2 A NOVEL KINEMATIC CHAIN SPACE

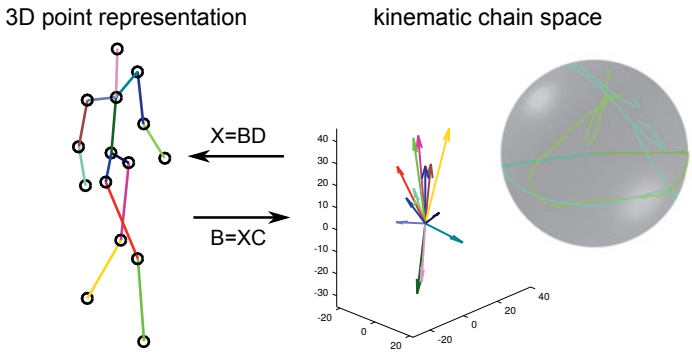


Figure 4.25: Mapping from a 3D point representation to the kinematic chain space. The vectors in the KCS equal to directional vectors in the 3D point representation. The sphere shows the trajectories of left and right lower arm in KCS. Since both bones have the same length their trajectories lie on the same sphere.

The previous section presented a learning-based approach to constrain the reconstructed human poses. Similar linear subspace training approaches have been proposed in [1, 68, 103, 112]. They can efficiently represent human poses, even for 3D reconstruction from single images. However, they require extensive training on known motions which restricts them to reconstructions of the same motion category. Furthermore, learning-based approaches cannot recover individual subtleties in the motion (e. g. limping instead of walking) sufficiently well. Thus, this section presents an approach to recover 3D human motions from image sequences without the need for a previously learned model.

The approach presented in this section closes the gap between non-rigid structure from motion and subspace-based human modeling. Similar to other approaches that depend on the work of Bregler et al. [10] and the previous section, we decompose an observation matrix into three matrices corresponding to camera motion, transformation and basis shapes. Unlike other works that find a transformation that enforces properties of the camera matrices, we develop an algorithm that optimizes the transformation with respect to structural properties of the observed object. This reduces the amount of camera motion necessary for a good reconstruction. We experimentally found that even sequences without camera motion can be reconstructed. Unlike other works in the field of human modeling we propose to first project the observations into a *kinematic chain space (KCS)* before optimizing a reprojection error

with respect to the kinematic model. Figure 4.25 shows the mapping between the KCS and the representation based on 2D or 3D feature points. It is done by multiplication with matrices which implicitly encode a kinematic chain (cf. Section 4.2.2). This representation enables the derivation of a nuclear norm optimization problem which can be solved efficiently by off-the-shelf solvers. Imposing a low rank constraint on a Gram matrix has shown to improve 3D reconstructions [17] which follows a similar idea to the proposed method. However, the method of Dai et al. [17] is only based on constraining the camera motion. Therefore, it requires sufficient camera motion. In contrast, the KCS allows for using a geometric constraint which is based on the topology of the underlying kinematic chain. Thus, the required amount of camera motion is much lower.

The proposed method is evaluated on different benchmark databases (CMU MoCap [11], KTH [36], HumanEva [75], Human3.6M [33]) as well as on our own databases qualitatively and quantitatively. The proposed algorithm achieves state-of-the-art results and can handle problems like motion transfers and unseen motions. Due to the noise robustness of our method we can apply a 2D joint detector [32, 65] which allows us to directly reconstruct human poses from unlabeled videos. Although this method is developed for human motion capture it is applicable to other kinematic chains such as animals or industrial robots as shown in the experiments in Section 4.2.6.3.

Summarizing, our contributions are:

- A method for 3D reconstruction of kinematic chains from monocular image sequences is proposed.
- An objective function based on structural properties of kinematic chains is derived that not only imposes a low-rank assumption on the shape basis but also has a physical interpretation.
- A nuclear norm optimization in a *kinematic chain space* is applied.
- In contrast to other works the proposed method is not limited to previously learned motion patterns and does not use strong anthropometric constraints such a-priorly determined bone lengths.

4.2.1 *Estimating Camera and Shape*

The i-th joint of a kinematic chain is defined by a vector $\boldsymbol{x}_i \in \mathbb{R}^3$ containing the x,y,z-coordinates of the location of this joint. By concatenating j joint vectors we build a matrix representing the pose $\boldsymbol{X}$ of the kinematic chain

$$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_j). \tag{4.27}$$

The pose $\boldsymbol{X}_k$ in frame k can be projected into the image plane by

$$\boldsymbol{X}'_k = \boldsymbol{K}_k \boldsymbol{X}_k, \tag{4.28}$$

where $\boldsymbol{K}_k$ is the projection matrix corresponding to a weak perspective camera. For a sequence of f frames, the pose matrices are stacked such that $\boldsymbol{W} = (\boldsymbol{X}'_1, \boldsymbol{X}'_2, \ldots, \boldsymbol{X}'_f)^T$ and $\hat{\boldsymbol{X}} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_f)^T$. This implies

$$\boldsymbol{W} = \boldsymbol{K}\hat{\boldsymbol{X}}, \tag{4.29}$$

where $\boldsymbol{K}$ is a block diagonal matrix containing the camera matrices $\boldsymbol{K}_{1,\ldots,f}$ for the corresponding frame. After an initial camera estimation we subtract a matrix $\boldsymbol{X}_0$ from the measurement matrix by

$$\hat{\boldsymbol{W}} = \boldsymbol{W} - \boldsymbol{K}\hat{\boldsymbol{X}}_0, \tag{4.30}$$

where $\hat{\boldsymbol{X}}_0$ is obtained by stacking $\boldsymbol{X}_0$ multiple times to obtain the same size as $\boldsymbol{W}$. Here, we take $\boldsymbol{X}_0$ to be a mean pose. We will provide experimental evidence that the algorithm proposed in the following is insensitive w.r.t. the choice of $\boldsymbol{X}_0$ as long as it represents a reasonable configuration of the kinematic chain. In all the experiments dealing with kinematic chains of humans, we take $\boldsymbol{X}_0$ to be the average of all poses in the CMU dataset.

Following the approach of Bregler et al. [10] we decompose $\hat{\boldsymbol{W}}$ by Singular Value Decomposition (SVD) to obtain a rank-$3K$ pose basis $\boldsymbol{Q} \in \mathbb{R}^{3K \times j}$. While [10] and similar works then optimize a transformation matrix with respect to orthogonality constraints of camera matrices, we optimize the transformation matrix with respect to constraints based on a physical interpretation of the underlying structure. With $\boldsymbol{A}$ as transformation matrix for the pose basis we may write

$$\boldsymbol{W} = \boldsymbol{K}(\hat{\boldsymbol{X}}_0 + \boldsymbol{A}\boldsymbol{Q}). \tag{4.31}$$

In the following sections we will present how poses can be projected into the kinematic chain space (Section 4.2.2) and how we derive an optimization problem from it (Section 4.2.3). Combined with the camera estimation (Section 4.2.4) an alternating algorithm is presented in Section 4.2.5.

4.2.2 *Kinematic Chain Space*

To define a bone $\boldsymbol{b}_k$ in the kinematic chain, a vector between the r-th and t-th joint is computed by

$$\boldsymbol{b}_k = \boldsymbol{K}_r - \boldsymbol{K}_t = \boldsymbol{X}\boldsymbol{c}, \tag{4.32}$$

where

$$\boldsymbol{c} = (0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)^T, \tag{4.33}$$

with 1 at position r and -1 at position t. The vector $\boldsymbol{b}_k$ has the same direction and length as the corresponding bone. Similarly to Eq. (4.27), a matrix $\boldsymbol{B} \in \mathbb{R}^{3\times b}$ can be defined containing all b bones

$$\boldsymbol{B} = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_b). \tag{4.34}$$

The matrix $\boldsymbol{B}$ is calculated by

$$\boldsymbol{B} = \boldsymbol{X}\boldsymbol{C}, \tag{4.35}$$

where $\boldsymbol{C} \in \mathbb{R}^{j\times b}$ is built by concatenating multiple vectors $\boldsymbol{c}$. Analogously to $\boldsymbol{C}$, a matrix $\boldsymbol{D} \in \mathbb{R}^{b\times j}$ can be defined that maps $\boldsymbol{B}$ back to $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{D}. \tag{4.36}$$

$\boldsymbol{D}$ is constructed similar to $\boldsymbol{C}$. Each column adds vectors in $\boldsymbol{B}$ to reconstruct the corresponding point coordinates. Note that $\boldsymbol{C}$ and $\boldsymbol{D}$ are a direct result of the underlying kinematic chain. Therefore, the matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ perform the mapping from point representation into the *kinematic chain space* and vice versa.

4.2.3 *Trace Norm Constraint*

One of the main properties of human skeletons is the fact that bone lengths do not change over time.

Let

$$\boldsymbol{\Psi} = \boldsymbol{B}^T\boldsymbol{B} = \begin{pmatrix} l_1^2 & \cdot & \cdot & \cdot \\ \cdot & l_2^2 & \cdot & \cdot \\ & & \ddots & \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & l_b^2 \end{pmatrix}. \tag{4.37}$$

be a matrix with the squared bone lengths on its diagonal. From $\boldsymbol{B} \in \mathbb{R}^{3\times b}$ follows $rank(\boldsymbol{B}) = 3$. Thus, $\boldsymbol{\Psi}$ has rank 3. Note that if $\boldsymbol{\Psi}$ is computed for every frame we can define a stronger constraint on $\boldsymbol{\Psi}$. Namely, as bone lengths do not change for the same person the diagonal of $\boldsymbol{\Psi}$ remains constant.

Proposition 1. *The nuclear norm of $\boldsymbol{B}$ is invariant for any bone configuration of the same person.*

Proof. The trace of $\boldsymbol{\Psi}$ equals the sum of squared bone lengths (Eq. (4.37))

$$trace(\boldsymbol{\Psi}) = \sum_{i=1}^{b} l_i^2. \tag{4.38}$$

From the assumption that bone lengths of humans are invariant during a captured image sequence the trace of $\boldsymbol{\Psi}$ is constant. The same argument holds for $trace(\sqrt{\boldsymbol{\Psi}})$. Therefore, we have

$$\|\boldsymbol{B}\|_* = trace(\sqrt{\boldsymbol{\Psi}}) = const. \tag{4.39}$$

□

Since this constancy constraint is non-convex we will relax it to derive an easy to solve optimization problem. Using Eq. (4.35) we project Eq. (4.31) into the KCS which gives

$$\boldsymbol{WC} = \boldsymbol{K}(\hat{\boldsymbol{X}}_0\boldsymbol{C} + \boldsymbol{AQC}) \tag{4.40}$$

The unknown is the transformation matrix $\boldsymbol{A}$. For better readability we define $\boldsymbol{B}_0 = \boldsymbol{X}_0\boldsymbol{C}$ and $\boldsymbol{S} = \boldsymbol{QC}$.

Proposition 2. *The nuclear norm of the transformation matrix $\boldsymbol{A}$ for each frame has to be greater than some scalar c, which is constant for each frame.*

Proof. Let $\boldsymbol{B} = \boldsymbol{B}_1 + \boldsymbol{B}_0$ be a decomposition of $\boldsymbol{B}$ into the initial bone configuration $\boldsymbol{B}_0$ and a difference to the observed pose $\boldsymbol{B}_1$. It follows that

$$\|\boldsymbol{B}\|_* = \|\boldsymbol{B}_1 + \boldsymbol{B}_0\|_* = c_1, \tag{4.41}$$

where c_1 is a constant. The triangle inequality for matrix norms gives

$$\|\boldsymbol{B}_1\|_* + \|\boldsymbol{B}_0\|_* \geq \|\boldsymbol{B}_1 + \boldsymbol{B}_0\|_* = c_1. \tag{4.42}$$

Since $\boldsymbol{B}_0$ is known, it follows

$$\|\boldsymbol{B}_1\|_* \geq c_1 - \|\boldsymbol{B}_0\|_* = c, \tag{4.43}$$

where c is constant. $\boldsymbol{B}_1$ can be represented in the shape basis $\boldsymbol{S}$ (cf. Section 4.2.1) by multiplying it with the transformation matrix $\boldsymbol{A}$

$$\boldsymbol{B}_1 = \boldsymbol{AS}. \tag{4.44}$$

Since the shape base matrix $\boldsymbol{S}$ is a unitary matrix the nuclear norm of $\boldsymbol{B}_1$ equals

$$\|\boldsymbol{B}_1\|_* = \|\boldsymbol{A}\|_*. \tag{4.45}$$

By Eq. (4.43) follows that

$$\|\boldsymbol{A}\|_* \geq c. \tag{4.46}$$

□

Proposition 2 also holds for a sequence of frames. Let $\hat{\boldsymbol{A}}$ be a matrix built by stacking $\boldsymbol{A}$ for each frame and $\hat{\boldsymbol{B}}_0$ be defined similarly, we relax Eq. (4.46) and obtain the final formulation for our optimization problem

$$\begin{aligned} \min_{\hat{\boldsymbol{A}}} \quad & \|\hat{\boldsymbol{A}}\|_* \\ \text{s.t.} \quad & \|\boldsymbol{W}\boldsymbol{C} - \boldsymbol{K}(\hat{\boldsymbol{A}}\boldsymbol{S} + \hat{\boldsymbol{B}}_0)\|_F = 0. \end{aligned} \tag{4.47}$$

Eq. (4.47) does not only define a low rank assumption on the transformation matrix. By the derivation above, we showed that the nuclear norm is reasonable because it has a concise physical interpretation. More intuitively, the minimization of the nuclear norm will give solutions close to a mean configuration $\boldsymbol{B}_0$ of the bones in terms of rotation of the bones. The constraint in Eq. (4.47) which represents the reprojection error prevents the optimization from converging to the trivial solution $\|A\|_* = 0$. This allows for a reconstruction of arbitrary poses and skeletons. Moreover, Eq. (4.47) is a well studied problem which can be efficiently solved by common optimization methods such as Singular Value Thresholding (SVT) [12].

The following paragraph briefly introduces SVT and its application to Equation (4.47). For further details the reader is referred to [12]. SVT was originally proposed to solve the matrix completion problem of a measurement matrix $\boldsymbol{M}$ with the known entries $(i, k) \in \Omega$, where Ω is a subset of all entries in $\boldsymbol{M}$, by a low rank approximation of the form

$$\begin{aligned} \min \quad & rank(\boldsymbol{Y}) \\ \text{s.t.} \quad & \boldsymbol{Y}_{ik} = \boldsymbol{M}_{ik}. \end{aligned} \tag{4.48}$$

Since this is an ill-posed problem it is relaxed to a nuclear norm optimization [13] by

$$\begin{aligned} \min \quad & \|\boldsymbol{Y}\|_* \\ \text{s.t.} \quad & \boldsymbol{Y}_{ik} = \boldsymbol{M}_{ik}. \end{aligned} \tag{4.49}$$

With a starting value $\boldsymbol{Z}^0$ which has the same dimension as $\boldsymbol{M}$ the SVT algorithm is defined by two alternating steps:

1. Perform a Singular Value Decomposition (SVD) $\boldsymbol{Z}^{t-1} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ and set all singular values in $\boldsymbol{\Sigma}$ to zero that are below a predefined threshold. The updated $\boldsymbol{Y}^t$ is given by $\boldsymbol{Y}^t = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ using the updated $\boldsymbol{\Sigma}$.

2. Update $\boldsymbol{Z}^t = \boldsymbol{Z}^{t-1} + \delta(\boldsymbol{M} - \boldsymbol{Y}^t)$ with δ as the step size which relates the speed of convergence.

In [12] also constraints in the form $f(\boldsymbol{Y})$ are considered, where f is as a convex function. This is exactly the form of Equation 4.47. That means the SVT algorithm can be applied directly to the pose estimation problem.

4.2.4 *Camera*

The objective function in Eq. (4.47) can also be optimized for the camera matrix $\boldsymbol{K}$. Since $\boldsymbol{K}$ is a block diagonal matrix, Eq. (4.47) can be solved block-wise for each frame. With $\boldsymbol{X}_i'$ and $\boldsymbol{K}_i$ corresponding to the observation and camera at frame i the optimization problem can be written as

$$\min_{\boldsymbol{K}_i} \|\boldsymbol{X}_i'\boldsymbol{C} - \boldsymbol{K}_i(\boldsymbol{A}\boldsymbol{S} + \boldsymbol{B}_0)\|_F. \tag{4.50}$$

Similar to Section 4.1.2 considering the entries in

$$\boldsymbol{K}_i = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \end{pmatrix} \tag{4.51}$$

we can enforce a weak perspective camera by the constraints

$$m_{11}^2 + m_{12}^2 + m_{13}^2 - (m_{21}^2 + m_{22}^2 + m_{23}^2) = 0 \tag{4.52}$$

and

$$m_{11}m_{21} + m_{12}m_{22} + m_{13}m_{23} = 0. \tag{4.53}$$

4.2.5 *Algorithm*

In the previous sections we derived an optimization problem that can be solved for the camera matrix $\boldsymbol{K}$ and transformation matrix $\boldsymbol{A}$ respectively. As both are unknown we propose algorithm 2 which alternatingly solves for both matrices. Initialization is done by setting all entries in the transformation matrix $\boldsymbol{A}$ to zero. Additionally, an initial bone configuration $\boldsymbol{B}_0$ is required. It has to roughly model a human skeleton but does not need to be the mean of the sequence.

4.2.6 *Experiments*

For the evaluation of our algorithm different benchmark datasets (CMU MoCap [11], HumanEva [75], KTH [36], Human3.6M [33]) were used. The quality of the 3D reconstructions is evaluated in terms of the MPJPE and 3DPE as described in Section 3.4. To compare sequences of different

Algorithm 2 Factorization algorithm for kinematic chains

% **Input:**
$\boldsymbol{B}_0 \leftarrow$ initial bone configuration
$\boldsymbol{C} \leftarrow$ kinematic chain matrix
$\boldsymbol{W} \leftarrow$ observation
$f \leftarrow$ number of frames
$\boldsymbol{A} \leftarrow \boldsymbol{0}$

while no convergence **do**
  for $t = 1 \rightarrow f$ **do**
    optimize $\|\boldsymbol{X}_t\boldsymbol{C} - \boldsymbol{K}_t(\boldsymbol{A}\boldsymbol{S} + \boldsymbol{B}_0)\|_F$
    insert $\boldsymbol{K}_t$ in $\boldsymbol{K}$
  end for
  perform SVT on
    $\min \|\hat{\boldsymbol{A}}\|_*$ s.t. $\|\boldsymbol{W}\boldsymbol{C} - \boldsymbol{K}(\hat{\boldsymbol{A}}\boldsymbol{S} + \hat{\boldsymbol{B}}_0)\|_F = 0$
end while

% **Output:**
$\boldsymbol{K}$: camera matrices
$(\hat{\boldsymbol{A}}\boldsymbol{S} + \hat{\boldsymbol{B}}_0)\boldsymbol{D}$: 3D poses

lengths the mean of the 3DPE over all frames is used. In the following it is referred to as *3D error*.

Additional to this quantitative evaluation we perform reconstructions of different kinematic chains in Section 4.2.6.3 and on unlabeled image sequences in Section 4.2.6.4. All animated meshes in this section are created using SMPL [44]. The SMPL model is fitted to the reconstructed skeleton and is used solely for visualization.

4.2.6.1 *Evaluation on Benchmark Databases*

To qualitatively show the drawbacks of learning-based approaches we reconstructed a sequence of a limping person. We use the method of Section 4.1 trained on walking patterns to reconstruct the 3D scene. Although the motions are very similar, the algorithm of Section 4.1 is not able to reconstruct the subtle motions of the limping leg. Figure 4.28 shows the knee angle of the respective leg. The learning-based method reconstructs a periodic walking motion and cannot recover the unknown asymmetric motion which makes it unusable for gait analysis applications. The proposed algorithm is able to recover the motion in more detail.

We compare our method with the unsupervised works [2, 26] and the learning-based approach of Section 4.1. The codes of [2] and [26] are freely available. Although there are slightly newer works, these two

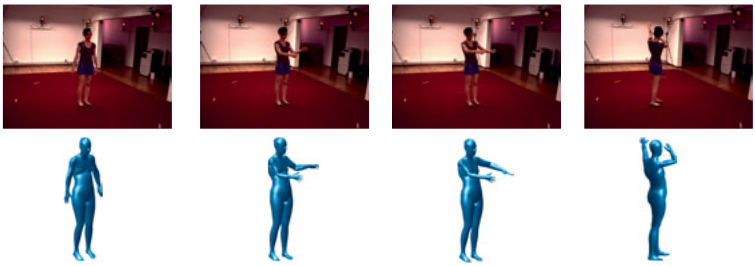


Figure 4.26: Reconstruction of the highly articulated *directions* sequence from the Human3.6M dataset subject 1.

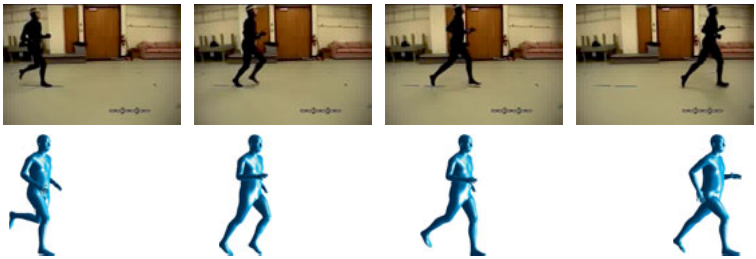


Figure 4.27: Reconstruction of a running motion from the CMU database subject 35/17.

approaches show the inherent problem of these unsupervised methods (as also shown in [70]). We are not aware of any works that are able to reconstruct scenes with very limited or no camera motion without a model of the underlying structure. Rehan et al. [70] assume a local rigidity that allows for defining a kinematic chain model. This reduced the amount of necessary camera motion to 2 degrees per frame. However, due to their assumption that the observed object is approximately rigid in a small time window they are limited to a constantly moving camera.

For each sequence we created 20 random camera paths with little or no camera motion and compared our 3D reconstruction results with the other methods. Table 4.2 shows the 3D error in *mm* for different sequences and datasets. For the entry *walk35* we calculated the mean overall 3D errors of all 23 walking sequences from subject 35 in the CMU database. The columns *jump* and *limp* show the 3D error of a single jumping and limping sequence. *KTH* means the football sequence of the KTH dataset [36] and *HE* the walking sequence of the HumanEva dataset [75]. The last four columns are average errors over all subjects

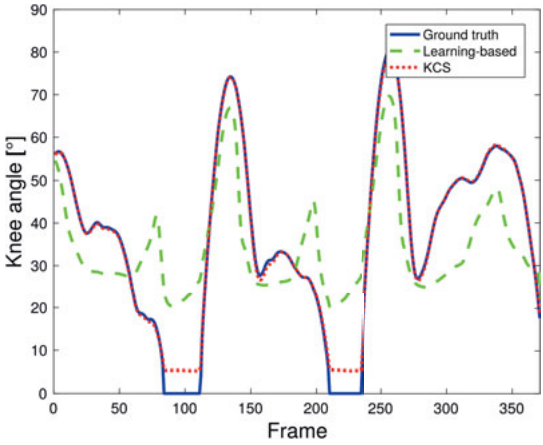


Figure 4.28: Knee angle of reconstructions of a limping motion. The learning-based method (Section 4.1) struggles to reconstruct minor differences from the motion patterns used for training whereas the proposed learning-free approach recovers the knee angle in more detail.

performing the respective motions of the Human3.6M dataset [33]. Note that the highly articulated motions from Human3.6M dataset vary a lot in the same category and therefore are harder to learn by approaches like in Section 4.1. All these sequences are captured with little or no camera motion. The unsupervised methods of [2] and [26] require more camera motion and completely fail in these scenarios. The learning-based approach of Section 4.1 reconstructs plausible poses for all sequences. They even achieve a better result for the walking motions. However, motions with larger variations between persons and sequences (e. g. jumping and limping) are harder to reconstruct from the learned pose basis. Although the results look like plausible human motions, they lack the ability to reconstruct subtle motion variations. In contrast, the proposed method is able to reconstruct these variations and achieves a better result. Some of our reconstructions are shown in Figs. 4.26 and 4.27 for sequences of the Human3.6M and CMU dataset, respectively.

4.2.6.2 *Convergence*

We alternatingly optimize the camera matrices (Eq. (4.47)) and transformation matrix (Eq. (4.50)). Since convergence of the algorithm cannot be guaranteed we show it by experiment. Figure 4.29 shows the convergence of the reprojection error in pixel for a sequence from the CMU MoCap database. However, the reprojection error only shows the convergence of

Table 4.2: 3D error in *mm* for different sequences and datasets. The column *walk35* shows the mean 3D error of all sequences containing walking motion from subject 35 in the CMU database. *jump* refers to the jumping motion of subject 13/11 of the CMU database and *limp* to the limping motion of subject 91/16. *KTH* means the football sequence of the KTH dataset [36]. The column *HE* shows the 3D error for the HumanEva walking sequence [75]. The last four columns are average errors over all subjects performing the respective motions of the Human3.6M dataset [33]. The row *4.1* shows the results obtained by the method presented in this Section 4.1 and the row *KCS* shows the results obtained by the method presented in this section.

	walk35	jump	limp	KTH	HE
[2]	228.68	210.14	99.37	108.91	106.92
[26]	264.75	186.70	112.92	114.03	102.99
4.1	**11.22**	45.49	64.46	68.88	58.62
KCS	18.94	**36.50**	**19.24**	**53.10**	**44.36**
	3.6M walk	3.6M dir.	3.6M pose	3.6M photo	3.6M mean
[2]	86.76	130.43	121.33	145.44	120.99
[26]	**66.70**	121.40	120.56	136.30	111.24
4.1	71.54	110.36	135.87	124.52	110.57
KCS	74.44	**80.83**	**109.28**	**101.76**	**91.58**

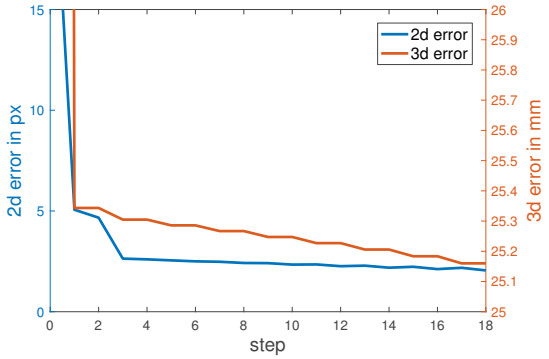


Figure 4.29: Reprojection error and 3D error with respect to number of iterations for subject35/sequence1 from the CMU MoCap dataset. Even steps refer to camera estimation while odd steps correspond to shape estimation.

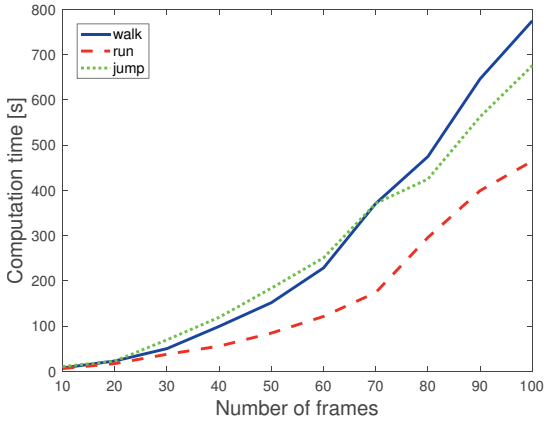


Figure 4.30: Computation time for walking, running and jumping sequences of the CMU dataset using unoptimized Matlab code. It mostly depends on the number of frames and less on the observed motion.

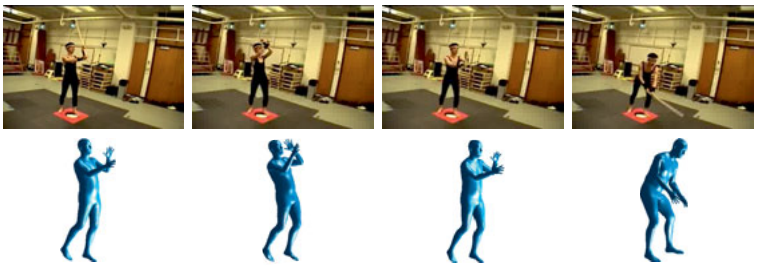


Figure 4.31: Reconstruction of the sword play sequence of the CMU database. The kinematic chain is extended such that the hands are rigidly connected.

the proposed algorithm but cannot prove that the 3D reconstructions will improve every iteration. We additionally estimated the convergence of the 3D error in Figure 4.29. In most cases our algorithm converges to a good minimum in less than 3 iterations. Further iterations do not improve the visual quality and only deform the 3D reconstruction less than $1mm$. The 3D error remains constant during camera estimation which causes the *steps* in the error plot.

Figure 4.30 shows the computation time over the number of frames for three different sequences. The computation time mostly depends on the number frames and less on the observed motion. We use unoptimized Matlab code on a desktop PC for all computations.

4.2.6.3 *Other Kinematic chains*

Figure 4.32: Reconstruction of a sequence of an industrial robot moving along a path. The reconstruction is shown as an augmented overlay over the images.

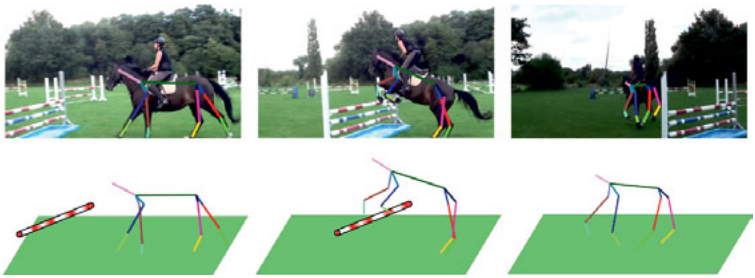


Figure 4.33: Reconstruction of a horse riding sequence. Although a very rough model for the skeleton of the horse is used plausible reconstructions are obtained.

Although our method was developed for the reconstruction of human motion, it generalizes to all kinematic chains that do not include translational joints. In this section we show reconstructions of other kinematic chains such as people holding objects, animals and industrial robots.

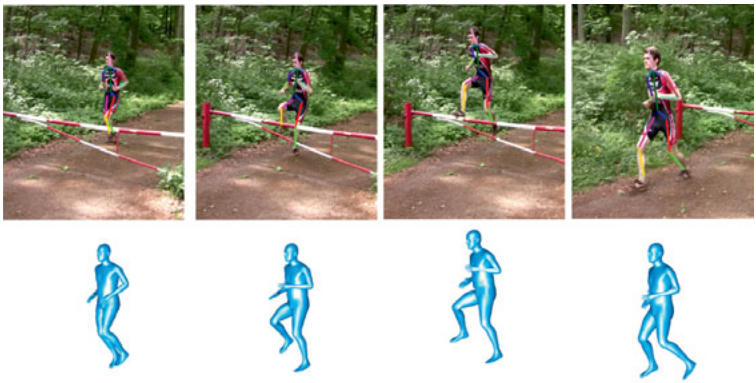


Figure 4.34: Reconstruction of a running and jumping sequence from [28] automatically labeled by *deeperCut* [32, 65].

In situations where people hold objects with both hands the kinematic chain of the body can be extended by another rigid connection between the two hands. Figure 4.31 shows the reconstruction of the sword fighting sequence of the CMU dataset. By simply adding another column to the kinematic chain space matrix $\boldsymbol{C}$ (cf. Section 4.2.2) the distance between the two hands is enforced to remain constant. The exact distance does not need to be known, however.

Figure 4.32 shows a robot used for precision milling and the reconstructed 3D model as overlay. The proposed method is able to correctly reconstruct the robots motion. In Figure 4.33 we reconstructed a more complex motion of a horse during show jumping. We used a simplified model of the bone structure of a horse. Also in reality the shoulder joint is not completely rigid. Despite these limitations the algorithm achieves plausible results.

4.2.6.4 *Image Sequences*

The proposed method is designed to reconstruct a 3D object from labeled feature points. In the former sections this was done by setting and tracking them semi-interactively. In this section we will show that our method is also able to use the noisy output of a human joint detector. We use *deeperCut* [32, 65] to estimate the joints in the outdoor run and jump sequence from [28]. Figure 4.34 shows the joints estimated by *deeperCut* and our 3D reconstruction. As can be seen in Figure 4.34 we achieve plausible 3D reconstructions even with automatically labeled noisy input data.

4.2.7 *Conclusion*

This section presented a method for the 3D reconstruction of kinematic chains from monocular image sequences. By projecting into the kinematic chain space a constraint is derived which is based on the assumption that bone lengths are constant over time. This results in the formulation of an easy to solve nuclear norm optimization problem. It allows for the reconstruction of scenes with little camera motion where other non-rigid structure from motion approaches fail. The presented method does not rely on previous training or predefined body measures such as known limb lengths. It generalizes to the reconstruction of other kinematic chains and achieves state-of-the-art results on benchmark datasets. In comparison to Section 4.1 it also reconstructs subtle motions such as limping in Figure 4.28. However, the robustness to occlusions is limited. To summarize, the method proposed in the previous Section 4.1 is a good choice for 2D detections with strong noise and occlusions, whereas the algorithm described in this section is preferred to reconstruct small deviations from everyday motions.

5

SINGLE IMAGE RECONSTRUCTION USING ADVERSARIAL TRAINING

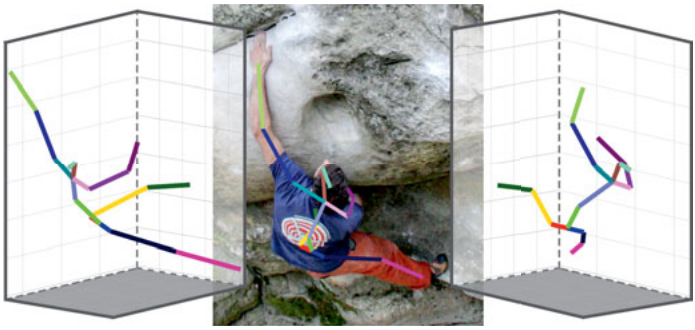


Figure 5.1: The proposed network predicts 3D human poses from noisy 2D joint detections. We use weakly supervised adversarial training without 2D to 3D point correspondences. The critic networks enforces a plausible 3D pose while a reprojection layer projects the 3D pose back to 2D. Even strong deformations and unusual camera poses can be reconstructed.

Parts of this chapter are based on a previous publication [99]. This chapter presents *RepNet*, a neural network that infers 3D joint coordinates directly from 2D observations. In contrast to Chapter 4, RepNet produces 3D reconstructions from single images, and therefore cannot employ temporal priors. It builds upon the findings of Section 4.1 that a learned space of human poses gives reasonable constraints. Since the presented linear model contains many implausible poses it needs to be regularized by temporal smoothness and bone lengths constancy priors. Since single images do not allow for temporal priors a nonlinear model is learned using a generative adversarial network (Section 3.2.3.2) that only contains plausible human poses. Additionally, the kinematic chain space, which was successfully applied to image sequences in Section 4.2, is integrated into a neural network layer to improve the network's capability of learning meaningful anthropometric constraints.

Comparable recent approaches are able to infer 3D human poses from monocular images in good quality (Section 2.2). However, most of them use neural networks that are straightforwardly trained with a

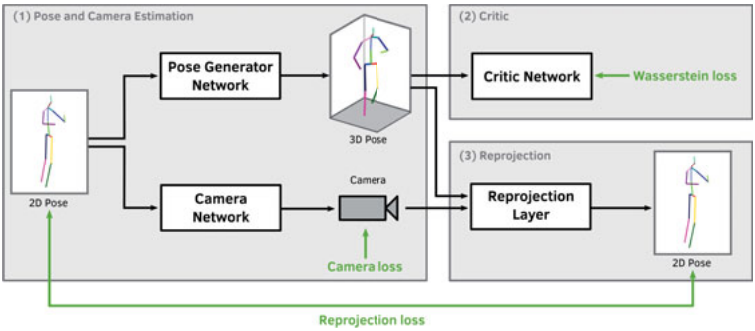


Figure 5.2: The proposed adversarial training structure for RepNet consists of three parts: a pose and camera estimation network (1), a critic network (2) and a reprojection network (3). There are losses (green) for the critic, the camera, and the reprojection. Similar to the training of GAN's the shown network is alternatingly trained with the discriminator network (Section 3.2.3.2).

strict assignment from input to output data, which are also mentioned in Section 2.2.2. This leads to surprisingly impressive results on similar data, but usually, the generalization to unknown motions and camera positions is problematic. This chapter presents a method to overcome this problem by using a neural network trained with a weakly supervised adversarial learning approach. We relax the assumption that a specific 3D pose is given for every image in the training data by training a discriminator network –widely used in generative adversarial networks (GAN) [21]– to learn a distribution of 3D human poses. A second neural network learns a mapping from the distribution of detected 2D keypoints (obtained by [58]) to the distribution of 3D keypoints which are valid 3D human poses according to the discriminator network. From the generative adversarial network point-of-view this can be seen as the generator network. To force the generator network to generate matching 3D poses to the 2D observations we propose to add a third neural network that predicts camera parameters from the input data. The inferred camera parameters are used to reproject the estimated 3D pose back to 2D which gives this framework its name: **Rep**rojection **Net**work (*RepNet*). Note that neither 2D-3D pairs nor known cameras are required, which enables the network to be trained with 2D poses from datasets without 3D annotations. Figure 5.2 shows an overview of the proposed network. Additionally, to further enforce kinematic constraints we propose to employ an easy to calculate and implement descriptor for joint lengths and angles inspired by the kinematic chain space (KCS) presented in Section 4.2.

In contrast to other works the proposed method is very robust against overfitting to a specific dataset. This claim is reinforced by our ex-

periments where the network can even infer human poses and camera positions that are not in the training set. Even if there are strong deformations or unusual camera poses our network achieves good results as can be seen in the rock climbing image in Figure 5.1. This leads to our conclusion that the discriminator network does not *memorize* all poses from the training set but learns a meaningful manifold of feasible human poses. As we will show the inclusion of the KCS as a layer in the discriminator network plays an important role for the quality of the discriminator.

We evaluate our method on the three datasets Human3.6M [33], MPI-INF-3DHP [52] and Leeds Sports Pose (LSP) [35]. On all the datasets RepNet achieves state-of-the-art results and even outperforms most supervised approaches. Furthermore, the proposed network can predict a human pose in less than 0.1 milliseconds on standard hardware which allows to build a real-time pose estimation system when combining it with state-of-the-art 2D joint detectors, such as OpenPose [14].

Summarizing, the contributions in this chapter are:

- An adversarial training method for a 3D human pose estimation neural network (RepNet) based on a 2D reprojection.
- Weakly supervised training without 2D-3D correspondences and unknown cameras.
- Simultaneous 3D skeletal keypoints and camera pose estimation.
- A layer encoding a kinematic chain representation that includes bone lengths and joint angle informations.
- A pose regression network that generalizes well to unknown human poses and cameras.

5.1 METHOD

The basic idea behind the proposed method is that 3D poses are regressed from 2D observations by learning a mapping from the input distribution (2D poses) to the output distribution (3D poses).

In standard generative adversarial network (GAN) training [21] a generator network learns a mapping from an input distribution to an output distribution which is rated by another neural network, called discriminator network. The discriminator is trained to distinguish between real samples from a database and samples created from the generator network. When training the generator to create samples that the discriminator predicts as real samples the discriminator parameters are fixed. The generator and the discriminator are trained alternatingly and therefore compete with each other until they both converge to a minimum.

In standard GAN training the input is sampled from a gaussian or uniform distribution. Here, we assume that the input is sampled from a distribution of 2D observations of human poses. Adopting the Wasserstein GAN [6] naming we call the discriminator *critic* in the following. Without knowledge about camera projections the network produces random, yet feasible human 3D poses. However, these 3D poses are very likely the incorrect 3D reconstructions of the input 2D observations. To obtain matching 2D and 3D poses we propose a camera estimation network followed by a reprojection layer. As shown in Figure 5.2 the proposed network consists of three parts: The pose and camera estimation network (1), the critic used in the adversarial training (2) and the reprojection part (3). The critic and the complete adversarial model are trained alternatingly as described above.

5.2 POSE AND CAMERA ESTIMATION

The pose and camera estimation network splits into two branches, one for regression of the pose and the other for the camera estimation. In the following $\boldsymbol{X} \in \mathbb{R}^{3\times n}$ denotes a 3D human pose where each column contains the xyz-coordinates of a body joint. In the neural network this matrix is written as a $3n$ dimensional vector. Correspondingly, if n joints are reconstructed the input of the pose and camera estimation network is a $2n$ dimensional vector containing the coordinates of the detected joints in the image.

The pose estimation part consists of two consecutive residual blocks, where each block has two hidden layers of 1000 densely connected neurons. For the activation functions we use leaky ReLUs [29] which produced the best results in our experiments. The last layer outputs a $3n$ dimensional vector which contains the 3D pose and can be reshaped to $\boldsymbol{X}$. The camera estimation branch has a similar structure as the pose estimation branch with the output being a 6 dimensional vector containing the camera parameters. Here, we use a weak perspective camera model that can be defined by only six variables. To obtain the camera matrix the output vector is reshaped to $\boldsymbol{K} \in \mathbb{R}^{2\times 3}$.

5.3 REPROJECTION LAYER

The reprojecting layer takes the output pose $\boldsymbol{X}$ of the 3D generator network and the camera $\boldsymbol{K}$ of the camera estimation network. The reprojecting into 2D coordinate space can then be performed by

$$\boldsymbol{W}' = \boldsymbol{K}\boldsymbol{X}, \tag{5.1}$$

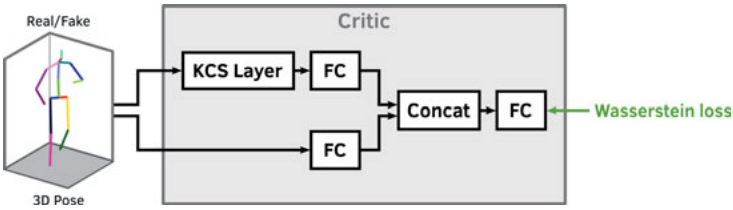


Figure 5.3: Network structure of the critic network. In the upper path the 3D pose is transformed into the KCS matrix and fed into a fully connected (FC) network. The lower path is build from multiple FC layers. The feature vectors of both paths are concatenated and fed into another FC layer which outputs the critic value.

where $\boldsymbol{W}'$ is called the *2D reprojection* in the following. This allows for the definition of a reprojection loss function

$$\mathcal{L}_{rep}(\boldsymbol{X}, \boldsymbol{K}) = \|\boldsymbol{W} - \boldsymbol{K}\boldsymbol{X}\|_F, \tag{5.2}$$

where $\boldsymbol{W}$ is the input 2D pose observation matrix which has the same structure as $\boldsymbol{W}'$. $\|\cdot\|_F$ denotes the Frobenius norm. Note that the reprojection layer is a single layer which only performs the reprojection and does not have any trainable parameters. To deal with occlusions columns in $\boldsymbol{W}$ and $\boldsymbol{X}$ that correspond to not detected joints can be set to zero. This means they will have no influence on the value of the loss function. The missing joints will then be hallucinated by the pose generator network according to the critic network. In fact, the stacked hourglass network that produces the 2D joint detections [58] that we use as the input does not predict the spine joint. We therefore set the corresponding columns in $\boldsymbol{W}$ and $\boldsymbol{X}$ to zero in all our experiments.

5.4 CRITIC NETWORK

The complete network in Figure 5.2 is trained alternatingly with the critic network. The loss on the last layer of the critic is a Wasserstein loss function [6]. The obvious choice of a critic network is a fully connected network with a structure similar to the pose regression network. However, such networks struggle to detect properties of human poses such as kinematic chains, symmetry and joint angle limits. Therefore, the *kinematic chain space (KCS)* introduced in Section 4.2 is integrated into the model. We develop a KCS layer with a successive fully connected network which is added in parallel to the fully connected path. These two paths in the critic network are merged before the output layer. Figure 5.3 shows the network structure of the critic.

The KCS matrix is a representation of a human pose containing joint angles and bone lenghts and can be computed by only two matrix multiplications. A bone $\boldsymbol{b}_k$ is defined as the vector between the r-th and t-th joint

$$\boldsymbol{b}_k = \boldsymbol{p}_r - \boldsymbol{p}_t = \boldsymbol{X}\boldsymbol{c}, \tag{5.3}$$

where

$$\boldsymbol{c} = (0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)^T, \tag{5.4}$$

with 1 at position r and -1 at position t. Note that the length of the vector $\boldsymbol{b}_k$ has the same direction and length as the corresponding bone. By concatenating b bones a matrix $\boldsymbol{B} \in \mathbb{R}^{3\times b}$ can be defined as

$$\boldsymbol{B} = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_b). \tag{5.5}$$

This leads to a matrix $\boldsymbol{C} \in \mathbb{R}^{j\times b}$ The matrix $\boldsymbol{B}$ is calculated by concatenating the corresponding vectors $\boldsymbol{c}$. It follows

$$\boldsymbol{B} = \boldsymbol{X}\boldsymbol{C}. \tag{5.6}$$

Multiplying $\boldsymbol{B}$ with its transpose gives the *KCS matrix* as defined in Section 4.2

$$\boldsymbol{\Psi} = \boldsymbol{B}^T\boldsymbol{B} = \begin{pmatrix} l_1^2 & \cdot & \cdot & \cdot \\ \cdot & l_2^2 & \cdot & \cdot \\ \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & \cdot & l_b^2 \end{pmatrix}. \tag{5.7}$$

Because each entry in $\boldsymbol{\Psi}$ is an inner product of two bone vectors the KCS matrix has the bone lengths on its diagonal and a (scaled) angular representation on the other entries. In contrast to an Euclidean distance matrix [56] the KCS matrix $\boldsymbol{\Psi}$ is easily calculated by two matrix multiplications. This allows for an efficient implementation as an additional layer. By giving the discriminator network an additional feature matrix it does not need to learn joint lengths computation and angular constraints on its own. In fact, in our experiments it was not possible to achieve an acceptable symmetry between the left and right side of the body without the KCS matrix. Section 5.8.1 shows how the 3D reconstruction benefits from adding the additional KCS layer. In our experiments there was no difference between adding convolutional layers or fully connected layers after the KCS layer. In the following we will use two fully connected layers, each containing 100 neurons, after the KCS layer. Combined with the parallel fully connected network this leads to the critic structure in Figure 5.3.

5.5 CAMERA

Since the camera estimation sub-network in Figure 5.2 can produce any 6-dimensional vector we need to force the network to produce matrices describing weak perspective cameras. If the 3D poses and the 2D poses are centered at their root joint the camera matrix $\boldsymbol{K}$ projects $\boldsymbol{X}$ to $\boldsymbol{W}'$ according to Eq. (5.1). A weak perspective projection matrix $\boldsymbol{K}$ has the property

$$\boldsymbol{K}\boldsymbol{K}^T = s^2\boldsymbol{I}_2, \tag{5.8}$$

where s is the scale of the projection and $\boldsymbol{I}_2$ is the 2×2 identity matrix. Since the scale s is unknown we derive a computationally efficient method of calculating s. The scale s equals to the largest singular value (or the ℓ_2-norm) of $\boldsymbol{K}$. Both singular values are equal. Since the trace of $\boldsymbol{K}\boldsymbol{K}^T$ is the sum of the squared singular values

$$s = \sqrt{trace(\boldsymbol{K}\boldsymbol{K}^T)/2}. \tag{5.9}$$

The loss function can now be defined as

$$\mathcal{L}_{cam} = \| \frac{2}{trace(\boldsymbol{K}\boldsymbol{K}^T)} \boldsymbol{K}\boldsymbol{K}^T - \boldsymbol{I}_2 \|_F, \tag{5.10}$$

where $\| \cdot \|_F$ denotes the Frobenius norm. Note that only one matrix multiplication is necessary to compute the quadratic scale.

5.6 DATA PREPROCESSING

The camera estimation network infers the parameters of the weak perspective camera. That means the camera matrix contains a rotational and a scaling component. To avoid ambiguities between the camera and 3D pose rotation all the rotational and scaling components from the 3D poses are removed. This is done by aligning every 3D pose to a template pose. We do this by calculating the ideal rotation and scale for the corresponding shoulder and hip joints via Procrustes alignment. The resulting transformation is applied to all joints.

Depending on the persons size in the image the 2D joint detections can have arbitrary scale. To remove the scale component we divide each 2D pose vector by its standard deviation. Note that using this scaling technique the same person can have different sized 2D pose representations depending on the camera and 3D pose. However, the value for all possible 2D poses is constrained. The remaining scale variations are compensated by the cameras scale component. In contrast to e. g. [51] we do not need to know the mean and standard deviation of the training set. This allows for an easy transfer of our method to a different domain of 2D poses.

5.7 TRAINING

We implemented the Improved Wasserstein GAN training procedure of [25]. In our experience this results in better and faster convergence compared to the traditional Wasserstein GAN [6] and standard GAN training [21] using binary cross entropy or similar loss functions. We use an initial learning rate of 0.001 with exponential decay every 10 epochs.

5.8 RESULTS

We perform experiments on the three datasets Human3.6M [33], MPI-INF-3DHP [52] and LSP [35]. Human3.6M is the largest benchmark dataset containing images temporally aligned to 2D and 3D correspondences. Unless otherwise noted we use the training set of Human3.6M for training our networks. To show quantitative results on unseen data we evaluate our method on MPI-INF-3DHP. For unusual poses and camera angles subjective results are shown on LSP. Matching most comparable methods we use stacked hourglass networks [58] for 2D joint estimations from the input images in most of the experiments.

5.8.1 *Quantitative Evaluation on Human3.6M*

The two main evaluation protocols on the Human3.6M dataset are followed (Section 3.5) by using subjects 1, 5, 6, 7, 8 for training and subject 9, 11 for testing. Both protocols calculate the *mean per joint positioning error* (MPJPE), i. e. the mean Euclidean distance between the reconstructed and the ground truth joint coordinates (Section 3.4). Protocol-I computes the MPJPE directly whereas protocol-II first employs a rigid alignment between the poses. For a sequence the MPJPE's are summed and divided by the number of frames.

Table 5.1 shows the results of protocol-I without a rigid alignment. The rotation of the pose relative to the camera can be directly calculated from the camera matrix estimated by the camera regression network. Rotating the reconstructed pose in the world frame of the dataset gives the final 3D pose. Table 5.2 shows the results of protocol-II using a rigid alignment before calculating the error. The row *RepNet-noKCS* shows the errors without using the KCS layer. It can be seen that the additional KCS layer in the discriminator significantly improves the pose estimation. We are aware of the fact that our method will not be able to outperform supervised methods trained to perform exceptionally well on Human3.6M, such as [51] and [45]. Instead, in this section we show that even if we ignore the 2D-3D correspondences and train weakly supervised our network achieves comparable results to supervised state-of-the-art methods and is even better than most of them. Comparing to weakly

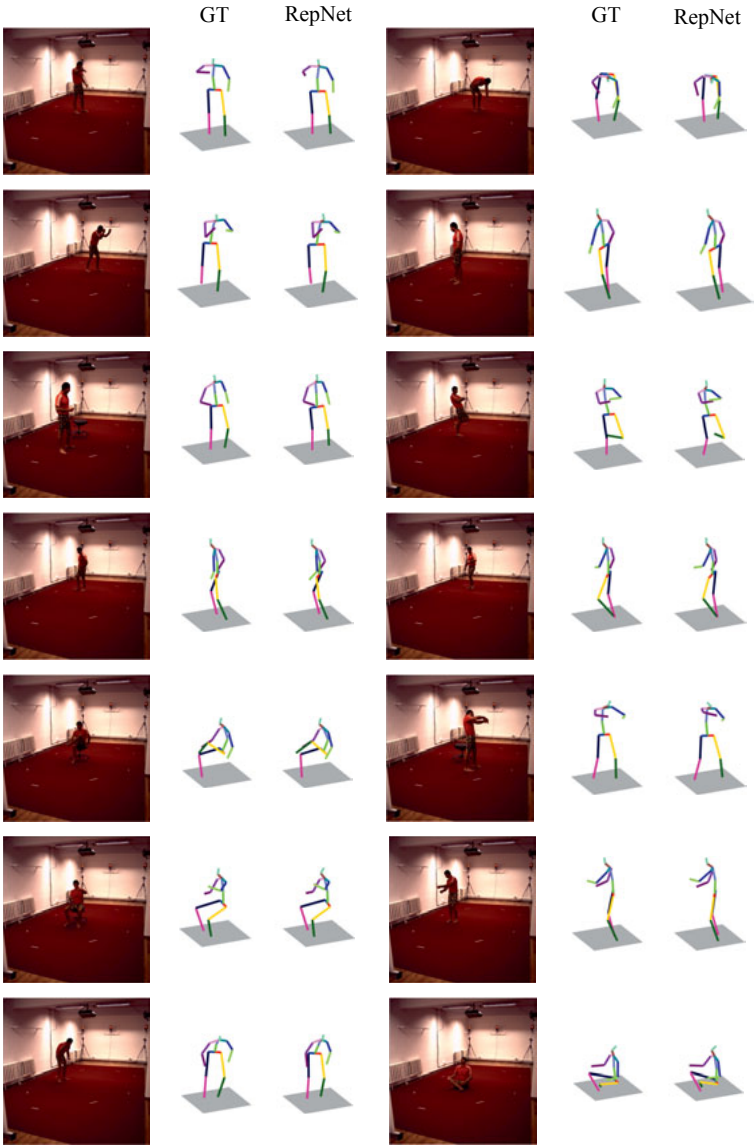


Figure 5.4: One example reconstruction for every motion from the test set of Human3.6M. The left 3D skeleton is the ground truth (GT) and the right shows our reconstruction (RepNet). Even difficult poses such as crossed legs or sitting on the floor are reconstructed well.

Table 5.1: Results for the reconstruction of the Human3.6M dataset compared to other state-of-the-art methods following *Protocol-I* (no rigid alignment). All numbers are taken from the referenced papers. For comparison the row *RepNet+2DGT* shows the error when using the ground truth 2D labels. The column *WS* denotes weakly supervised approaches. Note that there are no results available for other weakly supervised works.

Protocol-I	WS	Direct.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.
LinKDE [33]		132.7	183.6	132.3	164.4	162.1	205.9	150.6	171.3
Tekin et al. [80]		102.4	147.2	88.8	125.3	118.0	182.7	112.4	129.2
Zhou et al. [111]		87.4	109.3	87.1	103.2	116.2	143.3	106.9	99.8
Du et al. [18]		85.1	112.7	104.9	122.1	139.1	135.9	105.9	166.2
Park et al. [63]		100.3	116.2	90.0	116.5	115.3	149.5	117.6	106.9
Zhou et al. [113]		91.8	102.4	96.7	98.8	113.4	125.2	90.0	93.8
Luo et al. [45]		68.4	77.3	70.2	71.4	75.1	86.5	69.0	76.7
Pavlakos et al. [64]		67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3
Zhou et al. [114]		54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6
Martinez et al. [51]		53.3	60.8	62.9	62.7	86.4	82.4	57.8	58.7
RepNet (Ours)	✓	77.5	85.2	82.7	93.8	93.9	101.0	82.9	102.6
RepNet+2DGT (Ours)	✓	50.0	53.5	44.7	51.6	49.0	58.7	48.8	51.3

		Sit	SitD	Smoke	Wait	Walk	WalkD	WalkT	Avg.
LinKDE [33]		151.6	243.0	162.1	170.7	177.1	96.6	127.9	162.1
Tekin et al. [80]		138.9	224.9	118.4	138.8	126.3	55.1	65.8	125.0
Zhou et al. [111]		124.5	199.2	107.4	118.1	114.2	79.4	97.7	113.0
Du et al. [18]		117.5	226.9	120.0	117.7	137.4	99.3	106.5	126.5
Park et al. [63]		137.2	190.8	105.8	125.1	131.9	62.6	96.2	117.3
Zhou et al. [113]		132.2	159.0	107.0	94.4	126.0	79.0	99.0	107.3
Luo et al. [45]		88.2	103.4	73.8	72.1	83.9	58.1	65.4	76.0
Pavlakos et al. [64]		83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Zhou et al. [114]		75.2	111.6	64.2	66.1	63.2	51.4	55.3	64.9
Martinez et al. [51]		81.9	99.8	69.1	63.9	50.9	67.1	54.8	67.5
RepNet (Ours)	✓	100.5	125.8	88.0	84.8	72.6	78.8	79.0	89.9
RepNet+2DGT (Ours)	✓	51.1	66.0	46.6	50.6	42.5	38.8	60.4	50.9

Table 5.2: Results for the reconstruction of the Human3.6M dataset compared to other state-of-the-art methods following *Protocol-II* (rigid alignment). All numbers are taken from the referenced papers, except rows marked with * that are taken from [91]. Although we do not improve over supervised methods on this specific dataset our method clearly outperforms all other weakly supervised approaches (column *WS*). The best results for the weakly supervised methods are marked in bold. The second best approach that is not ours is underlined. For comparison the last row *RepNet+2DGT* shows the error when using the ground truth 2D labels.

Protocol-II	WS	Direct.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.
Akther and Black [1]		199.2	177.6	161.8	197.8	176.2	186.5	195.4	167.3
Ramakrishna et al. [68]		37.4	149.3	141.6	154.3	157.7	158.9	141.8	158.1
Zhou et al. [112]		99.7	95.8	87.9	116.8	108.3	107.3	93.5	95.3
Bogo et al. [8]		62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3
Moreno-Noguer [56]		66.1	61.7	84.5	73.7	65.2	67.2	60.9	67.3
Martinez et al. [51]		44.8	52.0	44.4	50.5	61.7	59.4	45.1	41.9
Luo et al. [45]		40.8	44.6	42.1	45.1	48.3	54.6	41.2	42.9
3Dinterpreter* [105]	✓	78.6	90.8	92.5	89.4	108.9	112.4	77.1	106.7
AIGN [91]	✓	77.6	91.4	89.9	88.0	107.3	110.1	75.9	107.5
RepNet (Ours)	✓	**53.0**	**58.3**	**59.6**	**66.5**	**72.8**	**71.0**	**56.7**	**69.6**
RepNet-noKCS (Ours)	✓	63.1	67.4	71.5	78.5	85.9	82.6	70.8	82.7
RepNet+2DGT (Ours)	✓	33.6	38.8	32.6	37.5	36.0	44.1	37.8	34.9

		Sit	SitD	Smoke	Wait	Walk	WalkD	WalkT	Avg.
Akther and Black [1]		160.7	173.7	177.8	181.9	198.6	176.2	192.7	181.1
Ramakrishna et al. [68]		168.6	175.6	160.4	161.7	174.8	150.0	150.2	157.3
Zhou et al. [112]		109.1	137.5	106.0	102.2	110.4	106.5	115.2	106.7
Bogo et al. [8]		100.3	137.3	83.4	77.3	79.7	86.8	87.7	82.3
Moreno-Noguer [56]		103.5	74.6	92.6	69.6	78.0	71.5	73.2	74.0
Martinez et al. [51]		66.3	77.6	54.0	58.8	35.9	49.0	40.7	52.1
Luo et al. [45]		55.5	69.9	46.7	42.5	36.0	48.0	41.4	46.6
3Dinterpreter* [105]	✓	127.4	139.0	103.4	91.4	79.1	-	-	98.4
AIGN [91]	✓	124.2	137.8	102.2	90.3	78.6	-	-	97.2
RepNet (Ours)	✓	**78.3**	**95.2**	**66.6**	**58.5**	**63.2**	**57.5**	**49.9**	**65.1**
RepNet-noKCS (Ours)	✓	92.2	116.6	77.6	72.2	65.3	73.2	69.6	77.9
RepNet+2DGT (Ours)	✓	39.2	52.0	37.5	39.8	34.1	40.3	34.9	38.2

supervised approaches [91, 105] we outperform the best by about 30% on protocol-II. For subjective evaluation the 1500th frame for every motion can be seen in Figure 5.4. For comparability we show the same frame from every motion sequence from the same viewing angle. Even difficult poses, for instance sitting cross-legged, are reconstructed well.

All approaches in Table 5.1 and Table 5.2 perform 3D reconstructions from single images. For comparison, the average MPJPE over all sequences from the Human3.6M dataset for the approaches in Chapter 4 are 110.6mm (Section 4.1) and 91.6mm (Section 4.2), which is comparable to the error of RepNet under protocol-I (89.9mm). Comparing the MPJPEs of single image approaches and methods employing temporal priors could be misleading since a small number of wrongly estimated joints in single images only have a minor impact on the MPJPE but can strongly distort a human pose. Moreover, oscillation around the ground truth occurs frequently when single image approaches are used for sequences. Although the performance of both approaches in Chapter 4 is not better than RepNet, they produce smoother 3D motions and are more robust to outliers and occlusions in the 2D detections.

In our opinion, although widely used on Human3.6M, the Euclidean distance is not the only metric that should be considered to evaluate the performance of a human pose estimation system. Since there are some single frames that cannot be reconstructed well and can be seen as outliers we also calculate the median of the MPJPE over all frames. Additionally, we calculate the *percentage of correctly positioned keypoints* (PCK3D) as defined by [52] in Table 5.3.

Table 5.3: Performance of our method regarding the median and PCK3D errors for the Human3.6M dataset. For comparison the last row *RepNet+2DGT* shows the error when using the ground truth 2D labels.

	mean	median	PCK3D
RepNet	65.1	60.0	93.0
RepNet+2DGT	38.2	36.0	98.6

In the following section we will show that although we do not improve on all supervised state-of-the-art methods directly trained on Human3.6M our approach outperforms every other known method on MPI-INF-3DHP without additional training.

5.8.2 *Quantitative Evaluation on MPI-INF-3DHP*

Our main contribution is a neural network that infers even unseen human poses while maintaining a meaningful 3D pose. We compare our method against several state-of-the-art approaches. Table 5.4 shows the results

Table 5.4: Results for the MPI-INF-3DHP dataset. All numbers are taken from the referenced papers, except the row marked with * which is taken from [54]. Without training on this dataset the proposed method outperforms every other method. The row *RepNet 3DHP* shows the result when using the training set of MPI-INF-3DHP. The column *WS* denotes weakly supervised approaches. A higher value is better for 3DPCK and AUC while a lower value is better for MPJPE. The best results are marked in bold and the second best approach is underlined.

Method	WS	3DPCK	AUC	MPJPE
Mehta et al. [52]		76.5	40.8	117.6
VNect [53]		76.6	40.4	124.7
LCR-Net[72]*		59.6	27.6	158.4
Zhou et al. [114]		69.2	32.5	137.1
Multi Person [54]		75.2	37.8	122.2
OriNet [45]		<u>81.8</u>	45.2	**89.4**
RepNet H36M (Ours)	✓	<u>81.8</u>	<u>54.8</u>	<u>92.5</u>
RepNet 3DHP (Ours)	✓	**82.5**	**58.5**	97.8

for different metrics. We clearly outperform every other method without having trained our model on this specific dataset. Even approaches trained on the training set of MPI-INF-3DHP perform worse than ours. This shows the generalization capability of our network. The row *RepNet 3DHP* is the result when training on the training set of MPI-INF-3DHP. There is only a minor improvement of the 3DPCK and AUC and even a minor deterioration of the MPJPE compared to the network trained on Human3.6M. This suggests that the critic network converges to a similar distribution of feasible human poses for both training sets.

5.8.3 *Plausibility of the Reconstructions*

Table 5.5: Symmetry error in *mm* of the reconstructed 3D poses on the different datasets with and without the KCS. Adding the KCS layer to the critic networks results in significantly more plausible poses.

Method	mean	std	max
H36M noKCS	31.9	9.3	61.3
H36M KCS	8.2	3.8	20.5
3DHP noKCS	32.9	21.9	143.9
3DHP KCS	11.2	8.0	54.7

The metrics used for evaluation in Section 5.8.1 and 5.8.2 compare the estimated 3D pose to the ground truth. However, a low error in this metrics is not necessarily an indication for a plausible human pose since the reconstructed pose can still violate joint angle limits or symmetries of the human body. For this purpose we introduce a new metric based on bone length symmetry. We calculate bone lengths of the lower and upper arms and legs since there is the largest error per joint. By summing the absolute differences of all matching bones on the right and left side of the body we can calculate a *symmetry error*. The mean symmetry error of the ground truth poses from the test set of Human3.6M and MPI-INF-3DHP for all subjects is $0.7mm \pm 0.8mm$ (max. $2.6mm$) and $2.1mm \pm 1.3mm$ (max. $7.6mm$), respectively. This leads us to the conclusion that an equality between the left and right side and therefore a low symmetry error is one reasonable metric for the plausibility of a human pose. Table 5.5 compares several implementations of our network in terms of the symmetry error. It can be clearly seen that the KCS layer has a significant impact on this metric. The higher values for the MPI-INF-3DHP dataset can be explained by the larger differences in symmetry of the ground truth data.

5.8.4 *Noisy observations*

Table 5.6: Evaluation results for protocol-II (rigid alignment) with different levels of Gaussian noise $\mathcal{N}(0, \sigma)$ (σ is the standard deviation) added to the ground truth 2D positions (GT). The 2D detector noise has large impact on the 3D reconstruction. The right three columns show the mean, standard deviation, and maximal symmetry error in millimeters.

Protocol-II	Direct.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	
GT	33.6	38.8	32.6	37.5	36.0	44.1	37.8	34.9	39.2	
GT + $\mathcal{N}(0, 5)$	54.0	56.8	52.7	56.5	54.4	59.7	55.7	54.1	56.3	
GT + $\mathcal{N}(0, 10)$	70.4	72.2	72.8	75.1	70.2	84.1	68.4	89.3	74.0	
GT + $\mathcal{N}(0, 15)$	86.3	88.0	87.5	89.9	84.0	98.1	84.0	104.2	87.4	
GT + $\mathcal{N}(0, 20)$	101.6	103.0	101.6	104.5	97.5	112.2	99.3	118.1	100.9	

	SitD	Smoke	Wait	Walk	WalkD	WalkT	Avg.	symmetry mean	symmetry std	symmetry max
GT	52.0	37.5	39.8	34.1	40.3	34.9	38.2	6.2	3.7	20.8
GT + $\mathcal{N}(0, 5)$	68.5	56.1	58.7	57.6	56.7	55.3	56.9	9.6	4.0	25.0
GT + $\mathcal{N}(0, 10)$	94.1	68.3	74.3	67.7	73.5	70.0	74.9	13.0	3.8	24.2
GT + $\mathcal{N}(0, 15)$	107.7	82.3	89.3	85.1	89.0	86.0	89.9	17.6	4.2	32.1
GT + $\mathcal{N}(0, 20)$	121.5	95.9	104.0	101.6	104.7	102.3	104.6	22.7	4.5	37.5

Since the performance of our network appears to depend a lot on the detections of the 2D pose detector we evaluate our network on different levels of noise. Following [56] we add Gaussian noise $\mathcal{N}(0, \sigma)$ to

the ground truth 2D joint positions, where σ is the standard deviation in pixel. The results for Human3.6M under protocol-II are shown in Table 5.6. The error scales linearly with the standard deviation. This indicates that the noise of the 2D joint detector has a major impact on the results. Considering Tables 5.1 and 5.2 an improved detector will enhance the results to a level where they outperform current state-of-the-art supervised approaches.

Please note that the maximum person size from head to toe is approximately 200px in the input data. Therefore, Gaussian noise with a standard deviation of $\sigma = 20$px can be considered as extremely large. However, due to the critic network using the KCS layer the output of the pose estimation network is still a plausible human pose. To demonstrate this we additionally investigated the average, standard deviation and maximal symmetry error for the different noise levels which is also shown in Table 5.6. As expected the error increases only slightly since the critic network enforces plausible human poses. Even for noise levels as high as $\mathcal{N}(0, 20)$ we achieve an average symmetry error of only $22.7mm \pm 4.5mm$ which can be considered as very low.

5.8.5 *Qualitative Evaluation*

For a subjective evaluation we use the Leeds Sports Pose dataset (LSP) [35]. This dataset contains 2000 images of different people doing sports. There is a large variety in poses including stretched poses close to the limits of possible joint angles. Some of these poses and camera angles were never seen before by our network. Nevertheless, it is able to predict plausible 3D poses for most of the images. Figure 5.5 shows some of the reconstructions achieved by our method. There are many subjectively well reconstructed poses, even if these are extremely stretched and captured from uncommon camera angles. Note that RepNet was only trained on the camera angles of Human3.6M. This underlines that an understanding of plausible poses and 2D projections is learned. The bottom row in Figure 5.5 shows some failure cases and emphasizes a limitation of this approach: poses or camera angles that are too different from the training data cannot be reconstructed well. However, the reconstructions are still plausible human poses and in most cases at least near to the correct pose.

5.8.6 *Conclusion*

This chapter presented RepNet: a weakly supervised training method for a 3D human pose estimation neural network that infers 3D poses from 2D joint detections in single images. We proposed to use an additional camera estimation network and our novel reprojection layer that projects the

Figure 5.5: Example 3D pose estimations from the LSP dataset. Good reconstructions are in the left columns. The bottom row shows some failure cases with very unusual poses or camera angles. Although not perfect, the poses are still plausible and close to the correct poses.

estimated 3D pose back to 2D. By exploiting state-of-the-art techniques in neural network research, such as improved Wasserstein GANs [25] and kinematic chain spaces [98], we were able to develop a weakly supervised training procedure that does not need 2D to 3D correspondences. This not only outperforms previous weakly supervised methods but also avoids overfitting of the network to a limited amount of training data. We achieved state-of-the-art performance on the benchmark dataset Human3.6M, even compared to most supervised approaches. Using the network trained on Human3.6M to predict 3D poses from the unseen data of the MPI-INF-3DHP dataset showed an improvement over all other methods. We also performed a subjective evaluation on the LSP dataset where we achieved good reconstructions even on images with uncommon poses and perspectives.

6

CONCLUSIONS

This thesis deals with the problem of image-based Human Motion Capture. Several applications can already be found in our everyday lives, e. g. in gaming devices or movies. There is a large amount of other possible applications in the industry, sports, medicine, autonomous driving. However, it still poses a major challenge to integrate existing MoCap technology into a product. The main reason is the need for traditional systems for multiple synchronized cameras and persons wearing optical markers. This is inconvenient and impractical in most scenarios and not suitable for mobile applications. To this end, this thesis focuses on monocular MoCap, i. e. using only a single camera. Since this makes the problem significantly harder compared to multi camera systems it requires sophisticated computer vision and machine learning solutions. Three of them are proposed in this thesis. The first combines a bone length consistency prior with a learned pose basis and temporal priors. Since a pretrained basis is very restrictive in terms of possible reconstructions the second approach avoids training a model for human poses by replacing it by a kinematic chain. The third approach learns a distribution of human poses using an adversarial neural network which is able to even reconstruct poses that are not in the training dataset.

PERIODIC AND NON-PERIODIC CONSTRAINTS

The approach presented in Section 4.1 exploits the facts that human motions are smooth and bone lengths remain constant for one person during an image sequence. Inspired by traditional NRSfM methods we factorize a measurement matrix into three matrices corresponding to the camera, the pose basis and coefficients for the bases. In contrast, we propose to learn the pose basis in advance from training data which gives strong constraints for possible 3D reconstructions. The smoothness of periodic motions is enforced by using periodic functions to model the weights of the base poses which turned out to be very effective and stable for periodic motions such as walking or running. For the reconstruction of non-periodic motions a novel regularization term was proposed. It regularizes the temporal bone length changes over time by a variance minimization. This led to high-quality 3D reconstructions of human motions even under difficult conditions, e. g. low camera motion where

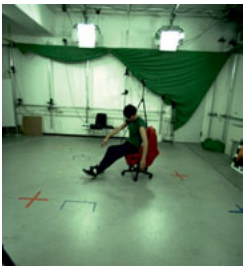
previous NRSfM methods produce degenerated solutions. Moreover, the learned pose basis enables the proposed algorithm to produce good reconstructions even with occlusions, noise and on the real-world data of the KTH dataset as well as on outdoor sequences.

KINEMATIC CHAIN SPACE

This part of the thesis (Section 4.2) introduces a more general method compared to the formerly presented approach. It generalizes to the reconstruction of arbitrary kinematic chains. The only requirement is a known kinematic chain. More specifically, only the connections between the joints are known while the bone lengths can be unknown. The novel kinematic chain space is introduced which allows for the derivation of an easy to solve nuclear norm optimization problem. In contrast to the previous approach it does not require a learned pose basis and is therefore able to reconstruct even subtle motions, for instance stumbling or limping. The proposed algorithm not only achieves state-of-the-art results on benchmark datasets but also generalizes to the reconstruction of other kinematic chains which was shown for industrial robots and horses.

REPNET

The second part of the thesis (Chapter 5) deals with 3D reconstruction of humans from single images. In contrast to the previous approaches that perform a global optimization over several frames the presented approach enables online deployment. It combines the key ideas of both of the former methods, namely learning a meaningful pose basis and representing poses in the kinematic chain space. The contribution of this chapter is to learn a nonlinear space of plausible human poses by applying a GAN. Instead of giving a randomly distributed input vector to the GAN, it receives 2D pose detections and learns a mapping to the 3D pose space that is evaluated by a discriminator network. It turned out that enriching the discriminator with a layer implementing a mapping into the kinematic chain space (introduced in Section 4.2) significantly improves its ability to distinguish plausible from implausible human poses. Additionally, to enforce consistency with the 2D detections a reprojection layer is proposed that reprojects the inferred 3D poses back to 2D which allows for the definition of a reprojection loss during training. Combining these ideas leads to a neural network that is trained with weak supervision and does not require 2D to 3D pose correspondences. This efficiently avoids overfitting to a specific dataset or activity that all previous state-of-the-art methods struggle with. We achieve state-of-the-art performance on the benchmark datasets Human3.6M and MPI-INF-3DHP, even compared


(a) Frame from [52].


(b) Indoor climbing.


(c) Outdoor climbing.

Figure 6.1: Recordings of humans performing different activities. Capturing these activities gets more complicated from (a) to (c). Although the persons in (b) and (c) perform the same activity the outdoor scene in (c) is significantly harder to capture.

to most supervised approaches. The good generalization to other poses is shown in several experiments subjectively and objectively.

FUTURE WORK

This thesis presents approaches to the human pose and motion estimation problem from monocular cameras by either exploiting temporal or structural properties.

One major drawback of every learning-based method is their dependency on the training data since they are only able to reproduce formerly seen poses in the database. Existing databases mostly contain everyday motions, for instance walking, sitting, or different working motion. For some motions, such as rock climbing, parkour running, horse riding or skateboarding recording is extremely challenging or even impossible. Figure 6.1 shows the increasing domain gap between studio-recorded everyday activities in Figure 6.1a, indoor scenes in Figure 6.1b and outdoor scenes in Figure 6.1c. Even if data for the respective activity domains would exist, the inter-domain variability is still large, e. g. if indoor climbing is compared to the more complex setting of outdoor climbing. The presented RepNet is a step in the direction to generalize to a complete human pose space by learning a distribution of human poses instead of simply memorizing them from a database. Although it is able to reconstruct unseen poses it is still restricted to poses and camera views close to the training data. A possible research direction is to adapt the features learned from traditional MoCap datasets (e. g. by RepNet) to the target activity domain that does not exist in the training set. A vast number of domain adaptation techniques for image classification tasks have been proposed in the recent past. The most notable are DANN [20]

and CyCADA [31] that both use adversarial training but are not yet applied to human pose estimation. Another route is taken by [16, 38, 71, 92] who use unlabeled images to train neural networks self-supervised. They, however, require multiple views from the same pose for training.

A step beyond human motion capture is physical motion analysis. Given a 3D sequence of poses, the goal is to estimate the forces and torques acting on the inside and outside of the body. There exists a large amount of biomechanics literature that builds dynamics models of the human body. Even musculoskeletal models exist that model each muscle in the body individually. A rarely considered and very challenging problem is to estimate the torques and forces from monocular image data. One approach [109] was published in a joint work during the time at TNT that estimates the torques in the knees during gait motions and in the back during lifting motions. It was used to detect different gait patterns and unhealthy lifting motions and can give recommendations on how to improve these movements. Including knowledge about external forces into human pose estimation can also help to resolve the ambiguity between depth and person height as shown in [7].

BIBLIOGRAPHY

[1] Ijaz Akhter and Michael J. Black. “Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015, pp. 1446–1455.

[2] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. “Trajectory Space: A Dual Representation for Nonrigid Structure from Motion.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.7 (July 2011), pp. 1442–1456.

[3] Thiemo Alldieck, Marc Kassubeck, Bastian Wandt, Bodo Rosenhahn, and Marcus Magnor. “Optical Flow-based 3D Human Motion Estimation from Monocular Video.” In: *German Conference on Pattern Recognition (GCPR)*. Sept. 2017.

[4] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. “Monocular 3d pose estimation and tracking by detection.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2010, pp. 623–630.

[5] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. “SCAPE: shape completion and animation of people.” In: *ACM Transactions on Graphics* 24 (2005), pp. 408–416.

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou. “Wasserstein Generative Adversarial Networks.” In: *International Conference on Machine Learning (ICML)*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, 2017, pp. 214–223.

[7] Didier Bieler, Semih Gunel, Pascal Fua, and Helge Rhodin. “Gravity as a Reference for Estimating a Person’s Height from Video.” In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 8569–8577.

[8] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. “Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image.” In: *European Conference on Computer Vision (ECCV)*. Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016.

[9] Matthieu Bray, Pushmeet Kohli, and Philip HS Torr. "Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts." In: *European Conference on Computer Vision (ECCV)*. Springer. 2006, pp. 642–655.

[10] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. "Recovering Non-Rigid 3D Shape from Image Streams." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2000, pp. 690–696.

[11] CMU. *Human motion capture database*. 2014. URL: `http://mocap.cs.cmu.edu/`.

[12] Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. "A Singular Value Thresholding Algorithm for Matrix Completion." In: *SIAM Journal on Optimization* 20.4 (Mar. 2010), pp. 1956–1982.

[13] Emmanuel J. Candès and Benjamin Recht. "Exact Matrix Completion via Convex Optimization." In: *CoRR* abs/0805.4471 (2008).

[14] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[15] Ching-Hang Chen and Deva Ramanan. "3D Human Pose Estimation = 2D Pose Estimation + Matching." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5759–5767.

[16] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Stefan Stojanov, and James M Rehg. "Unsupervised 3d pose estimation with geometric self-supervision." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 5714–5724.

[17] Yuchao Dai and Hongdong Li. "A Simple Prior-free Method for Non-rigid Structure-from-motion Factorization." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. CVPR '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 2018–2025.

[18] Yu Du, Yongkang Wong, Yonghao Liu, Feilin Han, Yilin Gui, Zhen Wang, Mohan Kankanhalli, and Weidong Geng. "Marker-less 3D human motion capture with monocular image sequence and height-maps." In: *European Conference on Computer Vision (ECCV)*. Springer. 2016, pp. 20–36.

[19] Karl Pearson F.R.S. “LIII. On lines and planes of closest fit to systems of points in space.” In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572.

[20] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. “Domain-adversarial training of neural networks.” In: *Journal of Machine Learning Research* 17.1 (2016), pp. 2096–2030.

[21] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative Adversarial Nets.” In: *International Conference on Neural Information Processing Systems (NIPS)*. NIPS’14. MIT Press, 2014, pp. 2672–2680.

[22] Paulo Gotardo and Aleix Martinez. “Kernel Non-Rigid Structure from Motion.” In: *International Conference on Computer Vision (ICCV)*. IEEE, 2011.

[23] Paulo Gotardo and Aleix Martinez. “Non-Rigid Structure from Motion with Complementary Rank-3 Spaces.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011.

[24] John C. Gower and Garmt B. Dijksterhuis. *Procrustes problems*. Vol. 30. Oxford Statistical Science Series. Oxford, UK: Oxford University Press, 2004.

[25] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. “Improved Training of Wasserstein GANs.” In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 5767–5777.

[26] Onur Hamsici, Paulo Gotardo, and Aleix Martinez. “Learning Spatially-Smooth Mappings in Non-Rigid Structure from Motion.” In: *European Conference on Computer Vision (ECCV)*. 2011.

[27] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[28] Nils Hasler, Bodo Rosenhahn, Thorsten Thormählen, Michael Wand, and Hans-Peter Seidel. “Markerless Motion Capture with Unsynchronized Moving Cameras.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2009).

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." In: *International Conference on Computer Vision (ICCV)*. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 1026–1034.

[30] David A. Hirshberg, Matthew Loper, Eric Rachlin, and Michael J. Black. "Coregistration: Simultaneous alignment and modeling of articulated 3D shape." In: *European Conference on Computer Vision (ECCV)*. LNCS 7577, Part IV. Springer-Verlag, Oct. 2012, pp. 242–255.

[31] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. "Cycada: Cycle-consistent adversarial domain adaptation." In: *arXiv preprint arXiv:1711.03213* (2017).

[32] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. "DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model." In: *European Conference on Computer Vision (ECCV)*. 2016.

[33] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 36.7 (2014), pp. 1325–1339.

[34] Hao Jiang. "3D Human Pose Reconstruction Using Millions of Exemplars." In: *International Conference on Pattern Recognition (ICPR)* (2010), pp. 1674–1677.

[35] Sam Johnson and Mark Everingham. "Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation." In: *British Machine Vision Conference (BMVC)*. doi:10.5244/C.24.12. 2010.

[36] Vahid Kazemi, Magnus Burenius, Hossein Azizpour, and Josephine Sullivan. "Multi-view Body Part Recognition with Random Forests." In: *British Machine Vision Conference (BMVC)*. 2013.

[37] Florian Kluger, Christoph Reinders, Kevin Raetz, Philipp Schelske, Bastian Wandt, Hanno Ackermann, and Bodo Rosenhahn. "Region-based Cycle-Consistent Data Augmentation for Object Detection." In: *IEEE International Conference on Big Data Workshops*. Dec. 2018.

[38] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. "Self-Supervised Learning of 3D Human Pose using Multi-view Geometry." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[39] Jeff Lander. “Skin them bones: Game programming for the web generation.” In: *Game Developer Magazine* (May 1998), pp. 11–16.

[40] Hsi-Jian Lee and Zen Chen. “Determination of 3D human body postures from a single view.” In: *Computer Vision, Graphics, and Image Processing* 30.2 (1985), pp. 148 –168.

[41] Minsik Lee, Jungchan Cho, Chong-Ho Choi, and Songhwai Oh. “Procrustean normal distribution for non-rigid structure from motion.” In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013, pp. 1280–1287.

[42] Sijin Li and Antoni B. Chan. “3D human pose estimation from monocular images with deep convolutional neural network.” English. In: *Asian Conference on Computer Vision (ACCV)*. Ed. by Ming-Hsuan Yang, Hideo Saito, Daniel Cremers, and Ian Reid. Vol. 9004. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Germany: Springer Verlag, Nov. 2014, pp. 332–347.

[43] Sijin Li, Weichen Zhang, and Antoni B. Chan. “Maximum-Margin Structured Learning with Deep Networks for 3D Human Pose Estimation.” In: *International Conference on Computer Vision (ICCV)*. ICCV ’15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 2848–2856.

[44] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. “SMPL: A Skinned Multi-Person Linear Model.” In: *ACM Transactions on Graphics* 34.6 (Oct. 2015), 248:1–248:16.

[45] Chenxu Luo, Xiao Chu, and Alan L. Yuille. “OriNet: A Fully Convolutional Network for 3D Human Pose Estimation.” In: *British Machine Vision Conference (BMVC)*. 2018, p. 92.

[46] Zhi-Quan Luo and Paul Tseng. “On the Convergence of the Coordinate Descent Method for Convex Differentiable Minimization.” In: *Journal of Optimization Theory and Applications* 72.1 (Jan. 1992), pp. 7–35.

[47] Hiroya Maeda, Yoshihide Sekimoto, Toshikazu Seto, Takehiro Kashiyama, and Hiroshi Omata. “Road Damage Detection and Classification Using Deep Neural Networks with Smartphone Images.” In: *Computer-Aided Civil and Infrastructure Engineering* (2018).

[48] Timo von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. "Human Pose Estimation from Video and IMUs." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 38.8 (2016), pp. 1533–1547.

[49] Timo von Marcard, Bodo Rosenhahn, Michael J. Black, and Gerard Pons-Moll. "Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs." In: *Computer Graphics Forum* 36.2 (2017), pp. 349–360.

[50] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. "Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera." In: *European Conference on Computer Vision (ECCV)*. Vol. Lecture Notes in Computer Science, vol 11214. Springer, Cham, Sept. 2018, pp. 614–631.

[51] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. "A simple yet effective baseline for 3d human pose estimation." In: *International Conference on Computer Vision (ICCV)*. 2017.

[52] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. "Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision." In: *International Conference on 3D Vision (3DV)*. IEEE. 2017.

[53] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. "VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera." In: *ACM Transactions on Graphics*. Vol. 36. 4. July 2017.

[54] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. "Single-Shot Multi-Person 3D Pose Estimation From Monocular RGB." In: *International Conference on 3D Vision (3DV)*. IEEE. 2018.

[55] MetaMotion. `https://metamotion.com/gypsy/gypsy-motion-capture-system.htm`, Last accessed on 2020-02-05.

[56] Francesc Moreno-Noguer. "3D Human Pose Estimation from a Single Image via Distance Matrix Regression." In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[57] Richard M. Murray, Zexiang Li, and S. Shankar Sastry. *A Mathematical Introduction to Robotic Manipulation*. 1994.

[58] Alejandro Newell, Kaiyu Yang, and Jia Deng. "Stacked Hourglass Networks for Human Pose Estimation." In: *European Conference on Computer Vision (ECCV)*. Vol. 9912. Lecture Notes in Computer Science. Springer, 2016, pp. 483–499.

[59] Optitrack. `https://optitrack.com/`, Last accessed on 2020-01-28.

[60] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. "A survey of structure from motion." In: *Acta Numerica* 26 (2017), pp. 305–364.

[61] Hyun Soo Park and Yaser Sheikh. "3D reconstruction of a smooth articulated trajectory from a monocular image sequence." In: *International Conference on Computer Vision (ICCV)*. Ed. by Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc J. Van Gool. IEEE Computer Society, 2011, pp. 201–208.

[62] Hyun Soo Park, Takaaki Shiratori, Iain Matthews, and Yaser Sheikh. "3D Reconstruction of a Moving Point from a Series of 2D Projections." In: *European Conference on Computer Vision (ECCV)* (2010).

[63] Sungheon Park, Jihye Hwang, and Nojun Kwak. "3D Human Pose Estimation Using Convolutional Neural Networks with 2D Pose Information." In: *European Conference on Computer Vision (ECCV)*. 2016, pp. 156–169.

[64] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. "Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose." In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 1263–1272.

[65] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. "DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[66] Polhemus. `https://polhemus.com/applications/electromagnetics/`, Last accessed on 2020-02-05.

[67] Qualisys. `https://www.qualisys.com/`, Last accessed on 2020-01-28.

[68] Varun Ramakrishna, Takeo Kanade, and Yaser Ajmal Sheikh. "Reconstructing 3D Human Pose from 2D Image Landmarks." In: *European Conference on Computer Vision (ECCV)*. 2012.

[69] Mir Rayat Imtiaz Hossain and James J. Little. "Exploiting temporal information for 3D human pose estimation." In: *European Conference on Computer Vision (ECCV)*. 2018.

[70] Ali Rehan, Aamer Zaheer, Ijaz Akhter, Arfah Saeed, Bilal Mahmood, Muhammad Usmani, and Sohaib Khan. "NRSfM using Local Rigidity." In: *Winter Conference on Applications of Computer Vision (WACV)*. Steamboat Springs, CO, USA: IEEE, Mar. 2014, pp. 69–74.

[71] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. "Learning monocular 3D human pose estimation from multi-view images." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 8437–8446.

[72] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. "LCR-Net: Localization-Classification-Regression for Human Pose." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, United States: IEEE, July 2017, pp. 1216–1224.

[73] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. "Structuring Autoencoders." In: *International Conference on Computer Vision (ICCV) Workshops*. Aug. 2019.

[74] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A. Kakadiaris. "3D Human pose estimation: A review of the literature and analysis of covariates." In: *Computer Vision and Image Understanding* 152 (2016), pp. 1–20.

[75] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. "HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion." In: *International Journal of Computer Vision* 87.1-2 (2010), pp. 4–27.

[76] Leonid Sigal and Michael J Black. "Predicting 3d people from 2d pictures." In: *International Conference on Articulated Motion and Deformable Objects*. Springer. 2006, pp. 185–195.

[77] Simi Reality Motion Systems. `http://www.simi.com/`, Last accessed on 2020-01-28.

[78] Edgar Simo-Serra, Arnau Ramisa, Guillem Alenyà, Carme Torras, and Francesc Moreno-Noguer. "Single image 3D human pose estimation from noisy observations." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2673–2680.

[79] Graham W Taylor, Leonid Sigal, David J Fleet, and Geoffrey E Hinton. "Dynamical binary latent variable models for 3d human pose tracking." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2010, pp. 631–638.

[80] Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua. "Direct Prediction of 3D Body Poses from Motion Compensated Sequences." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 991–1000.

[81] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. "Structured Prediction of 3D Human Pose with Deep Neural Networks." In: *British Machine Vision Conference (BMVC)*. 2016.

[82] The Captury. `https://thecaptury.com/`, Last accessed on 2020-01-28.

[83] Yan Tian, Leonid Sigal, Fernando De la Torre, and Yonghua Jia. "Canonical locality preserving latent variable model for discriminative pose inference." In: *Image and Vision Computing* 31.3 (2013), pp. 223–230.

[84] Carlo Tomasi and Takeo Kanade. "Shape and motion from image streams under orthography: a factorization method." In: *International Conference on Computer Vision (ICCV)* 9 (1992), pp. 137–154.

[85] Lorenzo Torresani, Aaron Hertzmann, and Christoph Bregler. "Learning Non-Rigid 3D Shape from 2D Motion." In: *Conference on Neural Information Processing Systems (NIPS)*. Ed. by Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf. MIT Press, 2003.

[86] Lorenzo Torresani, Aaron Hertzmann, and Christoph Bregler. "Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2008).

[87] Lorenzo Torresani, Danny B. Yang, Eugene J. Alexander, and Christoph Bregler. "Tracking and Modeling Non-Rigid Objects with Rank Constraints." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2001, pp. 493–500.

[88] Nikolaus F. Troje. "Decomposing biological motion: A framework for analysis and synthesis of human gait patterns." In: *Journal of Vision* 2.5 (2002), pp. 371–387.

[89] Nikolaus F. Troje. "The little difference: Fourier based synthesis of gender-specific biological motion." In: *AKA Press* (2002), pp. 115–120.

[90] Paul Tseng and Sangwoon Yun. "A coordinate gradient descent method for nonsmooth separable minimization." In: *Mathematical Programming* 117.1-2 (2009), pp. 387–423.

[91] Hsiao-Yu F. Tung, Adam W. Harley, William Seto, and Katerina Fragkiadaki. "Adversarial Inverse Graphics Networks: Learning 2D-to-3D Lifting and Image-to-Image Translation from Unpaired Supervision." In: *International Conference on Computer Vision (ICCV)*. 2017, pp. 4364–4372.

[92] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. "Self-supervised learning of motion capture." In: *Advances in Neural Information Processing Systems*. 2017, pp. 5236–5246.

[93] Jack Valmadre and Simon Lucey. "Deterministic 3D Human Pose Estimation Using Rigid Structure." In: *Proceedings of the 11th European Conference on Computer Vision Conference on Computer Vision: Part III*. ECCV'10. Springer-Verlag, 2010, pp. 467–480.

[94] Jack Valmadre, Yingying Zhu, Sridha Sridharan, and Simon Lucey. "Efficient Articulated Trajectory Reconstruction Using Dynamic Programming and Filters." In: *European Conference on Computer Vision (ECCV)*. Ed. by Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid. Vol. 7572. Lecture Notes in Computer Science. Springer, 2012, pp. 72–85.

[95] Vicon. `https://www.vicon.com/`, Last accessed on 2020-01-28.

[96] Bastian Wandt, Hanno Ackermann, and Bodo Rosenhahn. "3D Human Motion Capture from Monocular Image Sequences." In: *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2015.

[97] Bastian Wandt, Hanno Ackermann, and Bodo Rosenhahn. "3D Reconstruction of Human Motion from Monocular Image Sequences." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.8 (2016), pp. 1505–1516.

[98] Bastian Wandt, Hanno Ackermann, and Bodo Rosenhahn. "A Kinematic Chain Space for Monocular Motion Capture." In: *European Conference on Computer Vision (ECCV) Workshops*. Sept. 2018.

[99] Bastian Wandt and Bodo Rosenhahn. "RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[100] Bastian Wandt, Thorsten Laude, Yiqun Liu, Bodo Rosenhahn, and Jörn Ostermann. "Extending HEVC Using Texture Synthesis." In: *Visual Communications and Image Processing (VCIP)*. Dec. 2017.

[101] Bastian Wandt, Thorsten Laude, Bodo Rosenhahn, and Jörn Ostermann. "Detail-aware image decomposition for an HEVC-based texture synthesis framework." In: *Data Compression Conference (DCC)*. Mar. 2018.

[102] Bastian Wandt, Thorsten Laude, Bodo Rosenhahn, and Jörn Ostermann. "Extending HEVC with a Texture Synthesis Framework using Detail-aware Image Decomposition." In: *Picture Coding Symposium (PCS)*. June 2018.

[103] Chunyu Wang, Yizhou Wang, Zhouchen Lin, Alan Yuille, and Wen Gao. "Robust Estimation of 3D Human Poses from a Single Image." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[104] X. K. Wei and Jinxiang Chai. "Modeling 3D human poses from uncalibrated monocular images." In: *International Conference on Computer Vision (ICCV)*. 2009, pp. 1873–1880.

[105] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. "Single Image 3D Interpreter Network." In: *European Conference on Computer Vision (ECCV)*. 2016.

[106] Jing Xiao, Jinxiang Chai, and Takeo Kanade. "A Closed-Form Solution to Non-Rigid Shape and Motion Recovery." In: *European Conference on Computer Vision (ECCV)*. 2004.

[107] Xsens. `https://www.xsens.com/`, Last accessed on 2020-02-05.

[108] Angela Yao, Juergen Gall, Luc V Gool, and Raquel Urtasun. "Learning probabilistic non-linear latent variable models for tracking complex activities." In: *Advances in Neural Information Processing Systems*. 2011, pp. 1359–1367.

[109] Petrissa Zell, Bastian Wandt, and Bodo Rosenhahn. "Joint 3D Human Motion Capture and Physical Analysis from Monocular Videos." In: *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. July 2017.

[110] Zhengyou Zhang. "Microsoft kinect sensor and its effect." In: *IEEE multimedia* 19.2 (2012), pp. 4–10.

[111] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G. Derpanis, and Kostas Daniilidis. "Sparseness Meets Deepness: 3D Human Pose Estimation From Monocular Video." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[112] Xiaowei Zhou, Menglong Zhu, Spyridion Leonardos, and Kostas Daniilidis. “Sparse Representation for 3D Shape Estimation: A Convex Relaxation Approach.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.8 (2017), pp. 1648–1661.

[113] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. “Deep Kinematic Pose Regression.” In: (2016), pp. 186–201.

[114] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. “Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach.” In: *International Conference on Computer Vision (ICCV)*. 2017.

[115] Yingying Zhu, Mark Cox, and Simon Lucey. “3D motion reconstruction for real-world camera motion.” In: *CVPR*. IEEE Computer Society, 2011, pp. 1–8.

Bastian Wandt

Personal Data

Year of Birth:	1984
Place of Birth:	Peine, Germany
email:	wandt@tnt.uni-hannover.de

Work Experience

Current research associate at the **Leibniz University Hannover**, Germany,

Institut für Informationsverarbeitung,

Main Focus: Human pose estimation, Machine learning

Thesis: "Human Pose Estimation from Monocular Images" | Advisor: Prof. Dr.-Ing. Bodo Rosenhahn

Education

Sep 2014	Master of Science in Mechatronics from the **Leibniz University Hannover**
Sep 2012	Bachelor of Science in Mechatronics from the **Leibniz University Hannover**
Jun 2004	Abitur at Ratsgymnasium Peine

Die Reihen der Fortschritt-Berichte VDI:

1 Konstruktionstechnik/Maschinenelemente
2 Fertigungstechnik
3 Verfahrenstechnik
4 Bauingenieurwesen
5 Grund- und Werkstoffe/Kunststoffe
6 Energietechnik
7 Strömungstechnik
8 Mess-, Steuerungs- und Regelungstechnik
9 Elektronik/Mikro- und Nanotechnik
10 Informatik/Kommunikation
11 Schwingungstechnik
12 Verkehrstechnik/Fahrzeugtechnik
13 Fördertechnik/Logistik
14 Landtechnik/Lebensmitteltechnik
15 Umwelttechnik
16 Technik und Wirtschaft
17 Biotechnik/Medizintechnik
18 Mechanik/Bruchmechanik
19 Wärmetechnik/Kältetechnik
20 Rechnerunterstützte Verfahren (CAD, CAM, CAE CAQ, CIM ...)
21 Elektrotechnik
22 Mensch-Maschine-Systeme
23 Technische Gebäudeausrüstung

ISBN 978-3-18-386910-7