UnderPressure: Deep Learning for Foot Contact Detection, Ground Reaction Force Estimation and Footskate Cleanup

Lucas Mourot, Ludovic Hoyet, François Le Clerc, Pierre Hellier

UnderPressure: Deep Learning for Foot Contact Detection, Ground Reaction Force Estimation and Footskate Cleanup

Lucas Mourot[1,2] , Ludovic Hoyet[1] , François Le Clerc[2] and Pierre Hellier[2]

[1]Inria, Univ Rennes, CNRS, IRISA
[2]InterDigital, Inc

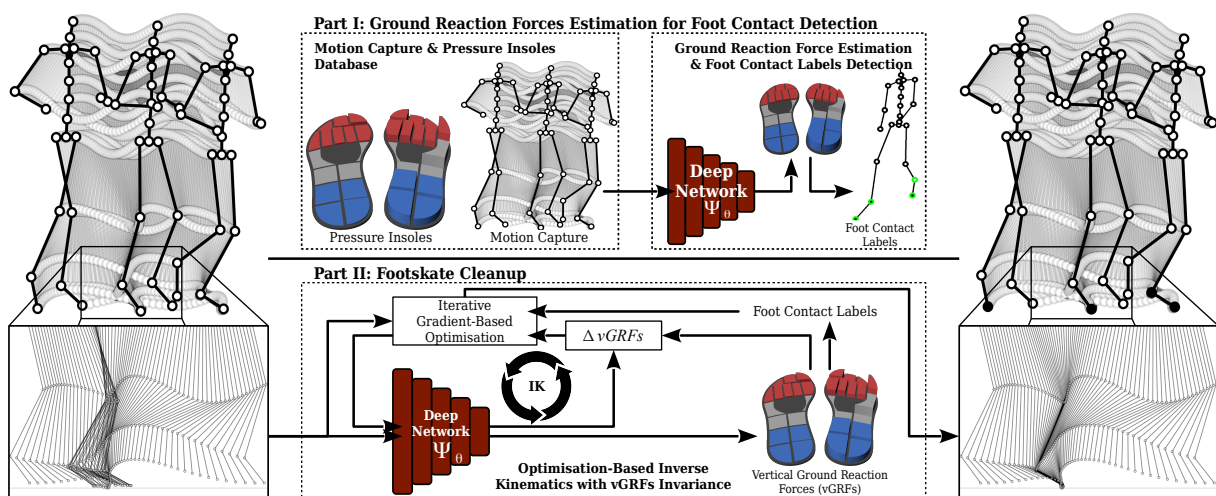**Figure 1:** *We leverage* UNDERPRESSURE, *a novel publicly available dataset of motion capture synchronized with pressure insoles data, to estimate vertical Ground Reaction Forces (vGRFs) from motion data with a deep neural network and derive foot contact labels. We further clean up footskate artefacts through an optimisation-based inverse kinematics algorithm while enforcing vGRFs invariance.*

Abstract

Human motion synthesis and editing are essential to many applications like video games, virtual reality, and film post-production. However, they often introduce artefacts in motion capture data, which can be detrimental to the perceived realism. In particular, footskating is a frequent and disturbing artefact, which requires knowledge of foot contacts to be cleaned up. Current approaches to obtain foot contact labels rely either on unreliable threshold-based heuristics or on tedious manual annotation. In this article, we address automatic foot contact label detection from motion capture data with a deep learning based method. To this end, we first publicly release UNDERPRESSURE, *a novel motion capture database labelled with pressure insoles data serving as reliable knowledge of foot contact with the ground. Then, we design and train a deep neural network to estimate ground reaction forces exerted on the feet from motion data and then derive accurate foot contact labels. The evaluation of our model shows that we significantly outperform heuristic approaches based on height and velocity thresholds and that our approach is much more robust when applied on motion sequences suffering from perturbations like noise or footskate. We further propose a fully automatic workflow for footskate cleanup: foot contact labels are first derived from estimated ground reaction forces. Then, footskate is removed by solving foot constraints through an optimisation-based inverse kinematics (IK) approach that ensures consistency with the estimated ground reaction forces. Beyond footskate cleanup, both the database and the method we propose could help to improve many approaches based on foot contact labels or ground reaction forces, including inverse dynamics problems like motion reconstruction and learning of deep motion models in motion synthesis or character animation. Our implementation, pre-trained model as well as links to database can be found at* github.com/InterDigitalInc/UnderPressure.

CCS Concepts

• Computing methodologies → Motion capture; Neural networks; Motion processing;

1. Introduction

Perceived realism is central to human animation; however, artefacts are often introduced whenever editing or synthesising motion data. These include feet sliding on, passing through or floating above the ground, which are known to disturb human perception even at very low intensities [PHO11]. In this paper we propose a fully automatic approach to foot contact detection and footskate artefacts removal.

As of today, foot contacts are derived from motion sequences using simple heuristics, most commonly hand-crafted thresholds over velocity and proximity relative to the ground. These approaches suffer from three main limitations. First, the terrain height map must be known, meaning that these approaches are most of the time not applicable to uneven or inclined ground. Second, optimal thresholds are not universal and must be manually tuned, ideally for every type of motion, morphology and contact location (e.g. heel or toe). Finally, even optimal thresholds fail at accurately detecting every foot contact, which implies that tedious manual checking is necessary when using these approaches. Finally, even optimal thresholds are far from being 100% accurate and lack of robustness, which implies that tedious manual checking is necessary when using these approaches.

In this paper, we propose a data-driven foot contact detection method from motion that outperforms traditional heuristic approaches. First, we introduce UNDERPRESSURE, a novel publicly available dataset composed of diverse motion capture data from 10 individuals, synchronised with pressure insoles data, from which corresponding vertical Ground Reaction Forces (vGRFs) and foot contact labels can be derived. Then, our key idea is to model vGRFs with a deep neural network, providing a finer representation of interactions of the feet with the ground than binary contact labels. Furthermore, vGRFs are related to the dynamics of motion unlike binary labels. To this end, we train a deep neural network to estimate the distribution of vGRFs over the feet, from which foot contact labels can be calculated, encouraging robustness and generalisability since the network needs a relatively high-level understanding of the motion dynamics to estimate accurate vGRFs.

We then provide a quantitative evaluation of our model against an *optimal thresholds (OT)* baseline that relies on thresholding the velocity and height above ground of foot joints, linear and non-linear learned generalisations of the OT baseline, as well as an ablative study of the proposed architecture. We further experiment how these models behave in different challenging conditions representative of real-world applications.

Finally, we demonstrate the generalisability of our approach on motion sequences from other databases, as well as its integration in a fully automatic footskate cleanup workflow. The novelty of this workflow is to leverage our deep neural network by enforcing invariance between vGRFs estimated from input and optimised motion sequences to better guide the IK optimisation and maintain the consistency of interactions between feet and ground.

To the best of our knowledge, we propose the first method of foot contact detection for animation learned on a significant amount of diverse motion data labelled with foot pressures. Here is a summary of our key contributions:

- a unique and publicly available motion capture dataset, called

UNDERPRESSURE, composed of 5.6 hours of diverse motions from 10 individuals and synchronized with pressure insoles data,
- a deep neural network modelling part of the motion dynamics from joint positions, i.e. vGRF distribution over the feet,
- a robust state-of-the-art method for foot contact detection, providing about 95% correct foot contact labels at 100 Hz on new subjects,
- a fully automatic workflow for footskate cleanup built on top of our foot contact detection method, preserving vGRFs estimated by our deep neural network.

2. Related Work

Early works in human animation already considered kinematic constraints such as foot contacts [BB98, LCR*02, KSG02, IAF06, LCB06]. Foot contact labels are helpful in numerous applications: they are often necessary to clean up foot artefacts such as foot sliding [KSG02], e.g. by enforcing foot constraints via Inverse Kinematics (IK) [MC12, HKS17, SZKS19, ALL*20, SZKZ20, HYNP20]; likewise, they are required to quantify such foot artefacts, e.g. for evaluation purposes [SZKS19, SZKZ20, ZSKS18, LZCvdP20]; they are also helpful to mitigate the well-known problem of mean collapse / drift away in human motion prediction, particularly with deterministic and recurrent models [HAB20, MBR17, WCX19, HYNP20], to disambiguate human motion modelling [XWCC15, HSK16, PGA18, PFAG20, HKS17], and more generally to leverage contact-based loss functions for increased quality and robustness [HKS17, LLL18, WHSZ19, PRL*19, SAA*20]; foot contact information is also deeply involved in character control based on physical simulation [ALX*19, KLvdP20, WL19, XLKvdP20, YTL18] where ground reaction forces are explicitly modelled in the physics engine, and also relates to foot-placement strategies that are a real challenge for locomotion policies [PvdP17]. In the following, we overview existing approaches for foot contacts labels detection (2.1) and ground reaction forces estimation (2.2), as well as existing databases of motion data labelled with information on foot contacts (2.3).

2.1. Foot Contact Labels Detection

The most widespread family of approaches in both animation research and industry to extract foot contact labels from motion data relies on simple heuristics with hand-crafted thresholds, applied to velocity and proximity relative to the ground as proposed by Bindiganavale and Badler [BB98] or Lee et al. [LCR*02]. However, foot contact identification methods relying on heuristics based on joint position and velocity lack temporal precision and are not reliable [LCB06]. Therefore, they are either limited to algorithms that are insensitive to the accuracy of contact labels [PRL*19], or require tedious manual checking and corrections [WCX19, MSZ*18, HKS17] to produce faithful contact detection results. Moreover, the accuracy of these approaches often dramatically drops when decreasing motion sequences quality, e.g. increasing noise or artefacts.

Other heuristic approaches have been sparsely investigated, including the work of Le Callennec and Boulic [LCB06]. Stationary or rotating point constraints are computed by solving linear equation systems assuming rigid transformations and a uniform noise

pattern, itself estimated from known user-defined constraint(s). While addressing the unreliability of common heuristics-based methods due to zero velocity assumption invalidated by noise, this approach is not fully automatic. Moreover, the noise pattern is assumed to be uniform along time, which is not always true.

Researchers also investigated learned contact detection models. E.g. Ikemoto et al. [IAF06] proposed a K-Nearest Neighbors (KNN) classifier to determine when the feet should be planted and claimed to be more accurate than heuristics-based approaches. However, a very low amount of manually annotated motion data was used (about 3 minutes, single subject) preventing the approach to generalise well. The proposed KNN classifier achieves 90.78% accuracy over about 33 seconds of hand-labelled motion data, compared to 57.45% and 57.00% for speed-based and height-based thresholding baselines, respectively. However, these relatively low accuracies suggest that the corresponding thresholds are not optimal.

More recently, researchers leveraged neural networks for exploring this problem. Smith et al. [SCNW19] detected foot contact labels using a dedicated Multilayer Perceptron (MLP) to remove footskate artefacts through IK in a motion style transfer pipeline. Zou et al. [ZYC*20], Shi et al. [SAA*20] and Rempe et al. [RGH*20] concurrently proposed to leverage foot contact detection to refine human motion estimation. In these works, foot contact labels are detected from 2D keypoints, themselves estimated from images using OpenPose [CHS*19], a state-of-the-art 2D pose estimator. Both Zou et al. [ZYC*20] and Rempe et al. [RGH*20] used a dedicated module for contact detection while Shi et al. [SAA*20] detected them together with 3D joint rotations and global root positions. Shimada et al. [SGXT20] extended the contact detection module from Zou et al. [ZYC*20] to additionally detect at each frame whether the subject is stationary or not. However, in all these works the foot contact detection modules were trained with ground truth contacts obtained from simple heuristics, as previously described. Tedious manual screening has occasionally been used to correct or label motion sequences, but this approach limits the amount of labelled data which is crucial with neural networks.

Other works also embedded foot contact detection into their model with the goal of improving foot positioning consistency, but do not support detection from motion data. Yang et al. [YKL21] estimated lower-body pose and foot contacts from upper-body AR/VR tracking devices. Harvey et al. [HYNP20] proposed a motion in-betweening method where pose and contact labels are interpolated between their known values at user-specified keyframes. Likewise, Holden et al. [HKS17] and Starke et al. [SZKS19] include foot contact detection at the next frame in character control frameworks, which enables foot contact continuation. Min and Chain [MC12] modelled foot contacts into a graph-based framework called *Motion Graphs++*, by annotating individual motion primitives with embedded contact information, able to randomly synthesise motion along with contacts at runtime.

2.2. Ground Reaction Forces Estimation

In physics, the force exerted by the ground on a body in contact, such as the human body, is called a ground reaction force (GRF). It is generally difficult to measure but nonetheless important in many fields of study including biomechanics, biomedical engineering and physics-based animation. Researchers in biomechanics and biomedical investigated GRF estimation from plantar pressure sensors [RFCA10, JJLK14, MMPD21], inertial and optical motion capture systems [FCK*15, KBS*16, FHS*16, MKD*20], 3D accelerometers [LBMN15], and Kinect [EKA*17]. However, GRF estimation for biomechanics or biomedical applications is beyond the scope of our approach that uses vertical GRF (vGRF) distribution as a proxy representation and is intended for human animation applications. For more details on GRF estimation in biomechanics and biomedical engineering, we refer the reader to the systematic review by Ancilloa et al. [ATBO18].

Early works in motion reconstruction leveraged pressure sensors to measure GRFs, because of their importance in dynamics. Ha et al. [HBL11] formulated the problem as a per-frame optimisation of end-effector positions obtained from a hand tracking device, and linear and angular momentums measured with pressure platforms. Later, Zhang et al. [ZSZ*14] leveraged a pair of pressure sensing shoes as well as three depth cameras to develop a full-body motion reconstruction framework consisting in kinematic pose reconstruction followed by physics-based motion optimisation.

More recently, several approaches instead estimated GRFs from monocular images, starting with 2D and 3D pose estimation and then solving physical optimisation problems. Zell et al. [ZWR17] proposed to estimate inner and exterior forces by optimising camera parameters and 2D pose reconstruction with a linear combination of base poses in a first step, and then GRFs and inner joint torques to satisfy the equations of motion and resolve camera projection ambiguities. Li et al. [LSC*19] estimated 3D motion and forces between a subject and its environment by minimizing the discrepancy between the observed and reprojected 2D poses, with priors on estimated 3D poses, trajectory smoothness and physical plausibility for regularization. Rempe et al. [RGH*20] and Shimada et al. [SGXT20] proposed a similar pipeline but focusing on more dynamic and diverse human motions, without object interactions. Shimada et al. further correct imbalanced stationary poses. Later on, they extended their approach with additional neural networks [SGX*21]: *TPNet* first regresses target 3D poses and contact states from 2D keypoints. Then, *GRFNet* and *DyNet* iteratively estimate GRFs and PD controller gain parameters in a dynamic cycle where the character pose is updated at each step after forward kinematics.

Different from motion reconstruction from images, Zell et al. [ZRW20] proposed a weakly-supervised approach to inverse dynamics. An MLP is trained to estimate GRFs, moments and joint torques from motion such that the input motion is reconstructed using forward dynamics in an optimisation loop. Motion capture data synchronized with force plates enable supervision during training: reconstructed GRF+M and joint torque divergences are penalized while GRF are minimized whenever feet are not in contact with the ground.

To the best of our knowledge, the closest work to the proposed method is the deep learning approach to improve human pose estimation computing stability proposed by Scott et al. [SRF*20]. In this work, body dynamics analysis from joint positions is investi-

gated while most papers on human pose estimation only focus on skeleton kinematics. A convolutional neural network called Press-Net is proposed to estimate 2D foot pressure maps from joint positions and validated on a novel dataset of Tai Chi sequences (see Section 2.3). Center of Pressure (CoP) and Base of Support (BoS) are computed from pressure values and used for validation. Unlike Scott et al., our work specifically targets foot contact detection and footskate cleanup, tasks that are relevant to human character animation. We explore more diverse motion sequences including different types of locomotion at different paces performed in different ways (forward, backward, sideways) as well as sequences in non-flat environments such as stair climbing and stepping on solid obstacles.

2.3. Foot Contact & Ground Reaction Force Databases

As of today, motion capture data annotated with accurate foot contact information are scarce. Researchers in biomechanics and biomedical engineering have released a few databases of motion capture with GRFs e.g. [KSN14], however most of them are not suitable for animation purposes as they typically focus on specific aspects of movement, or target stage and symptoms recognition of diseases in pathological subjects.

To the best of our knowledge, the closest database to the one we propose is PSU-TMM100. Recently released to the computer vision community by Scott et al. [SRF*20], this dataset provides videos from two views, motion capture markers, body joints and foot pressure recorded with insole sensors. It contains about 7.6 hours of data during which 10 subjects are performing 24-form simplified Tai Chi. Although similar in terms of scale and nature of the captured data, the database we propose has quite different types of motion. While PSU-TMM100 contains specific Tai Chi sequences mostly composed of slow body movements with long and stable foot supports, UNDER PRESSURE provides diversified sequences focused on but not restricted to locomotion at different paces including on non-flat environments (see Table 1), i.e. more challenging conditions for foot contacts detection.

3. Database

In this work, we release UNDER PRESSURE, a motion capture database annotated with pressure insoles data, designed primarily for character animation purposes. In the following, we provide information about the capture, the motion characteristics, and the pre-processing steps.

We recorded 10 healthy adult volunteers (2F, 8M) with diverse morphologies aged between 21 and 55 years ($32 \pm 11\ yr$), weighing between 65 and 91 kilograms ($79 \pm 9\ kg$), and measuring between 167 and 187 centimeters ($177 \pm 5\ cm$). Each subject performed the same set of activities, including forward and backward locomotion at different paces, sitting, standing, passing obstacles, climbing stairs, as well as motions on uneven terrain like going up and down stairs. The detailed composition of our dataset is provided in Table 1. Motion capture data for each subject last approximately 34 minutes, for a total of 5.6 hours of motion capture.

Table 1: *Motion sequence categories in* UNDER PRESSURE.

Category	Motion Type	Duration [mn]
Locomotion, forwards	slow walking	43.9
	normal walking	42.0
	fast walking	42.9
	running	30.1
Locomotion, backward	slow walking	21.1
	normal walking	21.8
	fast walking	20.8
	running	15.5
Locomotion, miscellaneous	running sideways	12.3
	hopping	13.9
	stairs 1 at a time	18.0
	stairs 2 at a time	13.9
Locomotion with obstacles	stepping on obstacles	5.7
	stepping over obstacles	11.4
	jumping over obstacles	10.5
Idle	leg stand-up	4.8
	sit-down	4.8
	crouched down	4.8
Total		338.2

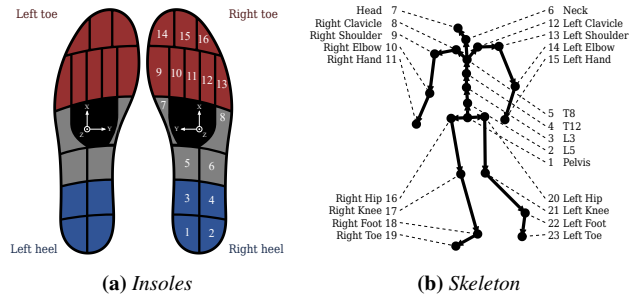
(a) *Insoles*   **(b)** *Skeleton*

Figure 2: *Left) Pressure cell layout of* Moticon's OpenGo Sensor Insoles *[AG21]. Blue (1 to 4) and red (9 to 16) cells are the groups of cells used to compute heel and toe contacts, respectively. Axes at insole centers represent inertial measurement units (IMUs). Right)* Xsens MVN's *[SGB18] motion capture skeleton with 23 joints.*

3.1. Motion Capture

Subjects were equipped with an *Xsens MVN Link* motion capture system [SGB18]. The hardware consists of 17 inertial measurement units (IMUs) running at 240 *Hz* embedded in the *MVN Link* suit. Each IMU contains a 3*D* accelerometer, a gyroscope, and a magnetometer. Capturing was performed using the *Xsens MVN Animate* software, an engine customised for 3*D* character animation that combines tracking data of the 17 individual IMUs with a 23-segment biomechanical model to obtain segment positions and orientations. We calibrated *MVN Animate* for each subject with height, arm span and shoe length measurements as inputs while other body dimensions and proportions were estimated through the calibration, as well as the orientation of motion trackers w.r.t. the corresponding segments. After *MVN Animate* processing, motion data consist of pose sequences with 23 segments sampled at 240 *Hz*.
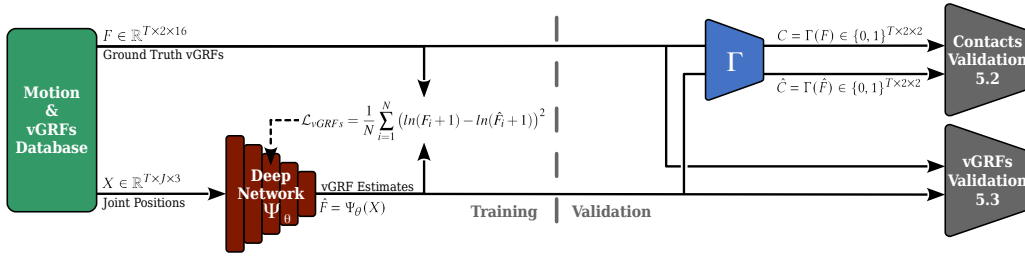
Figure 3: *Overview of our approach. Synchronized vGRFS and motion capture database (green rectangle) provide inputs and targets to train deep network Ψ_θ depicted in red. The blue trapezoid represents the contact function Γ (see Section 3.3). At runtime, our deep network augments motion data with vGRFs from which foot contact labels can be derived using the contact function Γ, both useful in many applications, e.g., reconstructing motion from images, cleaning footskate, finding suitable transition frames for motion blending, adapting animations to uneven terrain, and many more. As illustrated, both estimated vGRFs and detected contacts are evaluated in Sections 5.2 and 5.3, respectively.*

3.2. Foot Pressure

In addition to motion capture, we recorded the spatial distribution of plantar foot pressures. To this end, subjects were also equipped with *Moticon's OpenGo Sensor Insoles* [AG21] placed into their shoes. Each insole has 16 plantar pressure sensors with a resolution of $0.25 \, N/cm^2$ and a 6-axis inertial measurement unit (IMU), both running at $100 \, Hz$ (See Figure 2a). Moreover, we weighed each subject with full equipment to enable vGRFs normalisation and equipped subjects with the same shoes whose soles are thin and flexible insole for accurate and faithful pressure measures as well as for controlling the grip.

3.3. Post-Capture Processing

Vertical GRFs Captured data include motion sequences, plantar pressure distribution and foot acceleration. Since pressure is defined as the perpendicular force per unit area, we additionally compute vertical GRF components (vGRFs) by multiplying pressure values by the corresponding cell areas. The motivation here is that groups of vGRFs are easier to aggregate (by summation). We also normalise these values to express them as subject weight proportions.

Foot Contact Labels We derive ground truth foot contact labels deterministically from vGRFs. In this work, we consider two contact locations per foot, i.e. heels and toes as commonly done in human animation [IAF06, HKS17, SGX*21]. To compute contact labels, vGRF components are first smoothed with a Gaussian filter to avoid rapidly alternating labels due to threshold effects. Then, smoothed vGRF components are summed per contact location (see red and blue cells in Figure 2a), rescaled such that the sum of blue and red cell vGRFs is equal to total vGRF (to properly ignore gray cells which particularly suffer from noised measures and can be activated during either toes or heels contact) and then a threshold at 5% of the body weight is applied to obtain raw labels. Finally, raw labels are discarded whenever per-foot total vGRF (including gray cells) is below 10% of the body weight to avoid false positives triggered by noise. Contact phases shorter than $0.1s$ are also discarded

for the same reason. In the following, we refer to this binary contact labels calculation as the contact function Γ.

Synchronization Since we jointly capture motion and foot pressure data with separate devices, our records must be accurately synchronized in absence of a genlock signal. To this end, subjects were asked to perform a simple control movement at the beginning and end of each capture sequence, consisting in an in-place double-leg jump. This enables matching vertical acceleration peaks measured on pressure insole IMUs with peaks computed from motion captured foot positions. Although numerical differentiation is known to amplify high frequency noise, we found that the framerate of our motion capture data was sufficiently high and measurement noise was small enough for synchronization.

Downsampling and Trimming After synchronizing our data, we downsampled motion capture data from $240 \, Hz$ to $100 \, Hz$ using spherical linear interpolation to match the framerate of pressure insole data. Original motion sequences at $240 \, Hz$ are also provided in our database. We also automatically trimmed the beginning and end of each sequence to remove the synchronization patterns.

4. Deep Neural Network

In this section, we describe the proposed method to learn a deep neural network model estimating vGRF distributions from motion capture data. Learning vGRFs instead of binary contact labels encourages our deep neural network to more accurately model interactions between feet and ground, and enforces motion dynamics understanding. See Figure 3 for an overview of our approach.

4.1. Data Representation

Input At each frame t, the human pose $X_t \in \mathbb{R}^{J \times 3}$ is represented by the position of its $J = 23$ joints in a global Euclidean space. We design our deep network Ψ_θ to output vGRFs and contact labels at each frame from a few surrounding input frames with padding when needed. The full input pose sequence is then $X \in \mathbb{R}^{T \times J \times 3}$, where T is a variable number of frames.

Output As previously described, our deep network Ψ_θ estimates the vGRF distribution from motion data. At each frame, it outputs $\hat{F} = \Psi_\theta(X)$ with $\hat{F} \in \mathbb{R}^{T \times 2 \times 16}$, i.e. 16 positive real-valued vGRF components, corresponding to the 16 insole pressure cells for each foot, expressed proportionally to subject weight.

4.2. Network Architecture

We designed our network Ψ_θ to process variable-length sequences. To this end, the network is composed of four 1D temporal convolutional layers with 7-frame wide kernels, followed by three fully-connected layers applied at each frame independently to preserve the support variable-length sequence. Each convolutional or fully-connected layer is followed by exponential linear units (ELU) as activation, except for the last one which is a softplus activation to output nonnegative vGRF components.

4.3. Training and Inference

During training, our network is iteratively exposed to sequences of human poses and tries to estimate corresponding vGRF components as depicted in Figure 3. To encourage robustness and smooth convergence, we make use of stochastic data augmentation. First, similar to random crops and rotations used on images in computer vision, we apply random vGRF-invariant transformations on input pose sequences including translations, horizontal rotations, scaling, and left-right mirroring.

For each sequence, we also randomly draw its skeleton which is then animated by joint angles to further robustify our network. To do so, we precomputed (offline) an SVD basis of the skeletons captured in our database. At training time, we draw new skeletons by linearly combining precomputed singular vectors with randomly sampled weights. We then further edit these skeletons by randomly moving joint relative positions and rescaling bone lengths to obtain morphology variations. The resulting input motion sequences purposely suffer from artefacts since kinematic chains (i.e. from root joint to feet) have been randomly edited, which encourages the network to be resilient w.r.t. perturbed inputs. Joint positions are finally computed through forward kinematics and fed to the network.

To train our deep network Ψ_θ, we minimize a reconstruction loss of vGRF components. Instead of the standard mean squared error (MSE), we minimize the mean squared logarithmic error (MSLE). It has the property to only focus on the relative difference between target and estimated values (see right-hand side of Equation 1), which is convenient when the values considered can be several orders of magnitude apart. In our case, actual vGRFs can be strictly positive and arbitrarily low (e.g. during transition from the double leg stance to the single leg stance) as well as very high (e.g. during jump landing). The loss function used to train our network is then

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(ln(F_i + 1) - ln(\hat{F}_i + 1) \right)^2 = \frac{1}{N} \sum_{i=1}^{N} ln^2 \left(\frac{F_i + 1}{\hat{F}_i + 1} \right) \quad (1)$$

where F_i are the ground truth vGRF components, and $\hat{F}_i = \Psi_\theta(X)_i$ their estimated counterpart. Adding 1 to both F and $\hat{F}$ ensures that the loss is defined when vGRF component value goes to zero.

At inference, our deep network estimates vGRF components

from joint positions as inputs. Then, foot contact labels can be calculated from vGRF estimates using the contact function Γ (see Section 3.3).

5. Evaluation

In this section we present results of our method to assess estimated vGRFs and detected foot contacts. After providing implementation details necessary for reproducibility in Section 5.1, we assess foot contact labels detection in Section 5.2 then vGRFs estimation in Section 5.3 on ground truth motion sequences. In Section 5.4, we also evaluate foot contact detection performance on different perturbed motion sequences, simulating challenging conditions encountered in concrete applications. Qualitative results illustrating estimated vGRFs and foot contact labels as well as sample motion sequences of our database are available in the supplementary material.

5.1. Implementation Details

Data We split the proposed UNDERPRESSURE database into two subsets, i.e. a training set and a testing set. To ensure robust evaluation, the testing set is composed of the sequences performed by three out of the ten subjects (1F+2M, $\{S8, S9, S10\}$), representing approximately 30% of the overall database. For training, we further divide the remaining 70% to keep a validation set ($\sim 10\%$) and use early stopping during training. Moreover, we split each training motion sequence into overlapping windows of $T = 240$ frames, i.e. $2.4s$.

Architecture The four convolutional layers at the beginning of our network have respectively 128, 128, 256, and 256 7-frame wide filters while the back-end fully-connected layers have 256 neurons each. Dropout with probability $p = 0.2$ is applied before each fully-connected layer. The total number of weights in our network is approximately 1.1 million.

Training We implemented our deep neural network using Py-Torch. Training and validation were executed on an NVidia Tesla V100 GPU while other results were obtained either on an NVidia GeForce RTX 2060 GPU or on CPU. We trained our deep neural network through stochastic gradient descent (see Equation 1) for about 2500 epochs (about 10 days) at each of which a new version of the training set was randomly generated (see Section 4.3). We used Adam optimization with a batch size of 64, learning rate $\alpha = 3 * 10^{-5}$ and hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

5.2. Foot Contacts Detection

Metric To evaluate foot contact detection, we use the F_1 score which is the harmonic mean of precision (fraction of correctly detected labels among detected labels) and recall (fraction of correctly detected labels among expected contact labels).

Baseline To evaluate our model against commonly used heuristics-based approaches using thresholds (see Section 2), we define an *optimal thresholds (OT)* baseline. This model has two parameters

Table 2: *F_1 score on foot contact labels detection of our method, its variants for ablative study purposes, and the OT baseline. Bold and underline respectively indicate per-column best and second best. Our method outperforms the OT baseline and its linear generalisations, and the proposed architecture seems relevant w.r.t. the 3-layer MLP.*

Model	Walking	Running	Obstacles	Hopping	Stairs	Idle	Overall
OT baseline	0.927	0.869	0.926	0.859	0.882	0.826	0.909
Linear-Feet	0.936	0.863	0.921	0.824	0.906	0.812	0.913
Linear	0.937	0.868	0.926	0.855	0.925	0.912	0.923
3-layer MLP	0.940	0.883	0.926	0.882	0.947	0.921	0.930
Ours-C	0.946	0.917	0.941	<u>0.930</u>	<u>0.956</u>	0.941	0.942
Ours-C&F	<u>0.948</u>	<u>0.918</u>	<u>0.942</u>	0.923	0.954	<u>0.946</u>	<u>0.943</u>
Ours	**0.949**	**0.930**	**0.944**	**0.931**	**0.959**	**0.948**	**0.947**

that are thresholds on foot height relative to the ground and velocity norm. To demonstrate the effectiveness of our approach, we set these two parameters to the values which maximize the F_1 score over the training set, computed by recursive grid search. Note that in practical applications, such optimal thresholds are not available and must be guessed (hence our naming heuristics-based). This OT baseline therefore constitutes an upper bound in terms of performance over the F_1 score of such threshold-based approaches.

We also investigate the relevance of the proposed architecture with an ablative study. We first consider two linear models taking as inputs joint positions and velocities, as learned generalisations of the OT baseline: the former, called *Linear-Feet*, takes only foot and ankle joints as inputs while the latter, called *Linear*, takes all joints. Then, we consider a *3-layer MLP* model, i.e. the architecture of the foot contact label detection module proposed by Smith et al. [SCNW19] in their style transfer framework, which has three 128-neurons-wide layers and ReLU activations, introducing non-linearities w.r.t. Linear-Feet and Linear models. These three models have real-valued outputs, for which positive values are considered as foot contacts with the ground, and are trained with a binary cross-entropy (BCE) loss. Finally, we explore two variants of our deep network. First, the *Ours-C* variant has the architecture proposed in Section 4.2 except for the last layer and activation which are adapted (i.e. number of neurons and sigmoid instead of softplus) to detect foot contact labels instead of estimating vGRF components, and is trained with BCE loss. The second variant corresponds to the combination of *Ours-C* with the main proposed model *Ours*. In this extended variant called *Ours-C&F*, our architecture is adapted to output both contacts and vGRF components (convolutional layers are shared while fully-connected layers are duplicated and separately learned for each output) and trained with both MSLE (see Section 4.2) over vGRF components and BCE over contact labels.

Table 2 reports the F_1 score for each model and each motion category as well as overall results (right-most column). First, learned linear regressions show improved performances w.r.t. the OT baseline, which tends to confirm that thresholds based approaches lack complexity. The linear model (all joints) improvement w.r.t. Linear-
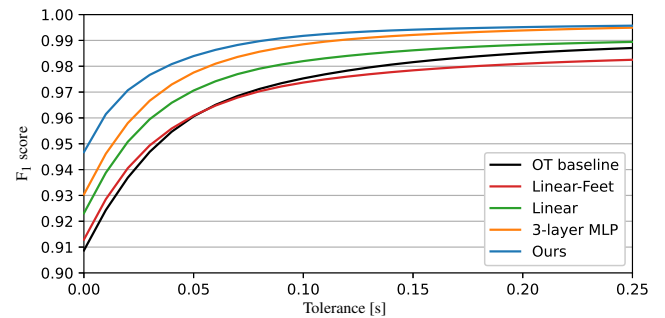

Figure 4: *F_1 score curves against temporal tolerance. At any given tolerance t, the F_1 score is computed with contact labels considered as incorrect if and only if they are wrong and located at least t seconds away from the closest contact phase. Without any tolerance (left-most), F_1 scores correspond to the* Overall *column in table 2. The OT baseline and its linear generalisations have large errors far from contact phase changes, resulting in relatively low F_1 scores compared to our model even with high temporal tolerances.*
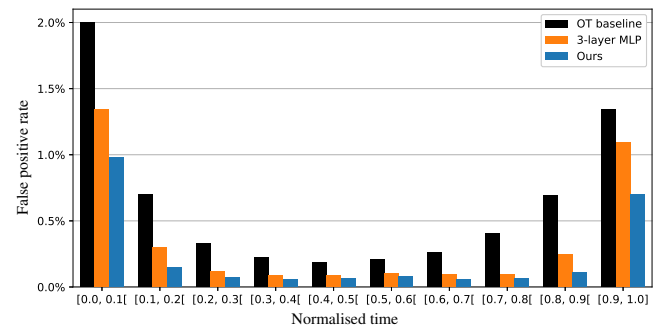

Figure 5: *False positive rate distributions w.r.t. normalised off-contact phases. Ground truth off-contact phase intervals are mapped to $[0,1]$ and false positives are aggregated in this normalised temporal frame. Intuitively, the farther false positives are from the closest contact phase (i.e. 0 or 1 in the normalised temporal frame), the more severe they are. False positive rates of learned models like ours quickly decrease while the OT baseline keeps it significantly higher in-between contacts.*

Feet (only foot and ankle) suggests that contact labels detection can also benefit from other body joints, pointing out one of the limitation of thresholds based approaches that typically process only foot joints. The higher F_1 score of the 3-layer MLP model w.r.t. linear models confirms their limitations for foot contact label detection. Finally the proposed architecture further increases the detection accuracy with relatively small differences among variants.

Temporal Analysis of Foot Contact Detection Up to now, we provided temporally global foot contact detection results. However, misdetected labels located closer to contact phase changes are less severe in the sense that they would result in less severe biases or artefacts in most applications, e.g. footskate cleanup. To this end, we provide a finer analysis of foot contact labels. In Figure 4, we

Table 3: *Root mean squared error of the estimated vGRF normalised by body weight. Estimating foot contacts in addition to vGRFs (Ours-C&F) seems slightly detrimental compared to estimating only vGRFs (Ours).*

Model	Walking	Running	Obstacles	Hopping	Stairs	Idle	Overall
Ours-C&F	9.5%	14.8%	12.4%	14.3%	11.8%	13.1%	11.4%
Ours	**9.1%**	**14.3%**	**11.6%**	**14.1%**	**10.9%**	**11.9%**	**10.9%**

plot the F_1 score against an increasing temporal tolerance to detection errors. In Figure 5, the distribution of false positive rates is given according to a normalised temporal frame where ground truth off-contact phase intervals are mapped to $[0, 1]$. Both figures confirm the limitations of heuristics approaches represented by the OT baseline whose false positive rate is significant in the middle of off-contact phases. On the contrary, learned models show convergence of accuracy close to 100% with increasing tolerance as well as low false positive rates in the middle of off-contact phases.

5.3. vGRFs Estimation

Metrics In this section we assess performances of our method on vGRFs estimation from motion capture data. We use the root mean squared error (RMSE) of per-foot total vGRF which is mainly sensitive to global biases (at foot scale). In complement, we also evaluated the Center of Pressure (CoP) computed from estimated vGRF components, which is mostly sensitive to local errors (at pressure cell scale) but we provide the corresponding results in the supplementary material.

Table 3 gives the RMSE of the estimated vGRFs proportionally to the subjects' body weight. As for contact detection, these results suggest that learning to model foot contact labels in addition to vGRF (Ours-C&F) reduces accuracy, although the performances of both variants are close. Note that other variants evaluated on contact detection in the previous section only detect contact labels and hence cannot be evaluated on vGRF estimation.

In biomechanics, force plates are considered gold standard to measure vGRF and CoP; however, the environment in which we captured the proposed database (significant area including obstacles and stairs) prevented us to use force plates. For this reason, our evaluation considers pressure insole measures as ground truth for evaluation purposes. Existing works in biomechanics [NSM11, JMNB19] evaluated pressure insoles accuracy w.r.t. force plates, and tell us that vGRFs measured with pressure insoles suffer from a RMSE up to approximately 10% of the subject weight, being subject to variations depending on experimental conditions. Thus, Table 3 indicate that vGRF estimation errors of our deep neural network are approximately of the same order of magnitude as the measurement error expected with pressure insoles.

5.4. Foot Contacts Detection in Challenging Conditions

As explained in Section 2, many applications requiring foot contact labels detection consist in quantifying or correcting foot artefacts.
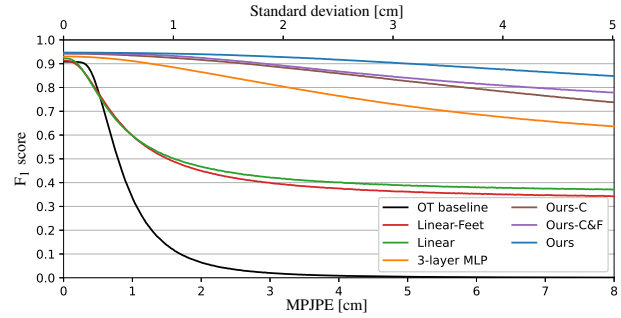

Figure 6: *F_1 score on foot contacts detection from motion sequences purposely noised with additive isotropic Gaussian noise. Each curve represents the F_1 score against the amount of noise introduced, measured with the MPJPE in centimeters indicated by the bottom horizontal axis, while the top horizontal axis gives the corresponding standard deviations of the Gaussian noise. The results indicate that our method is more robust to noise.*
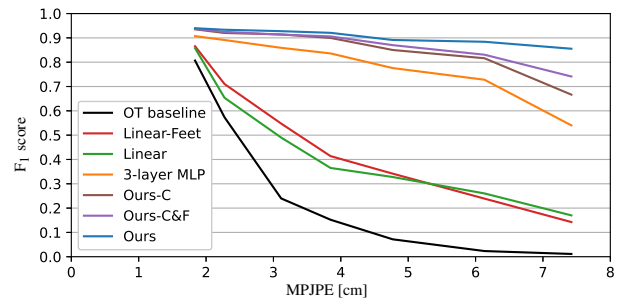

Figure 7: *F_1 score on foot contacts detection from motion sequences distorted by a motion autoencoder early-stopped at different epochs to emulate different amounts of distorsion. Each curve displays the F_1 score against the amount of distorsion introduced, measured with the MPJPE in centimeters. The results indicate that our method is more robust to distorted sequences.*

By definition, motion sequences to be processed in such applications are not expected to be flawless. As a consequence, contact detection performance is not solely relevant for clean motion capture data (as evaluated in Section 5.2), but also for perturbed sequences. In this section, we evaluate the performance of our model on motion sequences in which we introduce three types of artefacts: additive Gaussian noise, distorsions caused by going through a partially trained autoencoder, and artefacts (mostly footskate) obtained by blending different motion sequences.

Motions Perturbed with Gaussian Noise First, we simply evaluate how the contact detection performance of the various models evaluated in Section 5.2 is affected by an increasing amount of Gaussian noise added to joint positions. Figure 6 displays the F_1 score obtained by the different models against the amount of noise introduced.

Motions Distorted by an Autoencoder To simulate perturbations more faithful to what is encountered in real-world applications, we trained an autoencoder to reconstruct motions from our testset. We used a similar architecture to the convolutional autoencoder proposed by Holden et al. [HSK16] and trained it for 100 epochs (about 10 minutes). We kept a selection of epochs at which the amount of perturbations introduced by the autoencoder follows a nice monotonically decreasing curve to avoid outlier epochs. At each of these epochs, we evaluate foot contact detection on motion sequences that have been encoded and then decoded to introduce increasing perturbations. Again we use the F_1 score as a foot contact labels detection performance metric and display in Figure 7 the results against the observed amount of noise introduced by the autoencoder at different epochs.

Footskate Generated with Motion Blending To push further our evaluation toward concrete applications, we additionally evaluate contact detection on motion sequences that have been *blended*. Motion blending (or motion interpolation) is a well-known technique widely used in animation that consists in mixing existing motions with dynamic blending weights to create new motions. In our case, motion blending is interesting because it is known to easily produce footskate in resulting motions, which can then help us to evaluate contact detection on such perturbed motions. To do so, we randomly picked in our testset pairs of motion subsequences having identical foot contact patterns on either left or right foot. This approach enables us to make the hypothesis that foot contact patterns are preserved in blended motions, and thus serve as ground truth. We constituted a set of approximately $40'000$ blended motions of 80 frames long (i.e. $0.8s$) and quantified the amount of footskate introduced using the mean horizontal velocity of the feet during contact phases [SZKS19, SZKZ20, ZSKS18, LZCvdP20]. We then evaluate once again foot contact detection using the F_1 score with these blended motions as inputs. We obtain the curves plotted in Figure 8 using simple moving average.

As depicted in Figures 6 to 8, the proposed approach for foot contact labels detection is much more robust than threshold-based heuristics approaches represented by the OT baseline, regardless of the type of perturbation applied to motion sequences. Moreover, the improvement of modelling vGRFs instead of foot contact labels (Ours vs Ours-C) is much larger when facing perturbed motion sequences. Since vGRFs are much more related to motion dynamics than binary contact label, vGRFs estimation requires a deeper understanding of motion than contact labels detection. Indeed, approximately 90% of contact labels can be correctly detected with foot position and velocity thresholding (see Table 2), i.e. with almost no understanding. Then, in challenging conditions like perturbed input motion sequences, the performances of the variant modelling vGRFs are logically more stable since a deeper understanding of motion is intuitively more robust.

6. Footskate Cleanup

In this section, we leverage our robust foot contact detection for a downstream task and propose a novel fully automatic workflow for footskate cleanup: the foot contact labels lead to kinematic constraints used in a dedicated optimisation-based inverse kinematics algorithm, where foot joints are constrained at ground contact
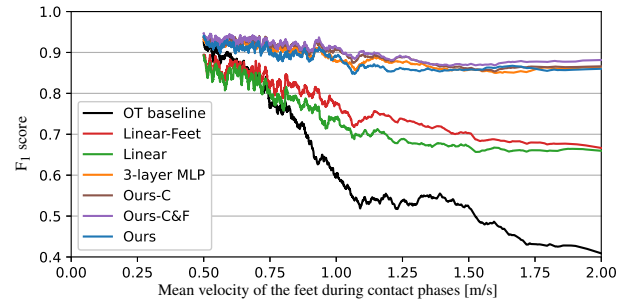

Figure 8: F_1 *score against amount of footskate. Test motions with matching foot contact patterns have been blended to purposely introduce footskate while preserving contact labels. Foot contact detection is then evaluated on such motions suffering from footskate, and compared to the ground truth contact labels. The amount of footskate introduced through blending is measure using the mean velocity of the feet during contact phases. Non-linear learned detection models keep reasonably high accuracy while accuracy of linear models significantly decrease and accuracy of OT baseline quickly collapse.*

points. The novelty of our approach is that the optimisation also includes the preservation of forces estimated by our network.

Let $\tilde{X} = (\tilde{Q}, S, \tilde{P})$ be the input motion sequence suffering from footskate, where $\tilde{Q}$ denote joint angles, S the skeleton and $\tilde{P}$ the global trajectory positions of the root joint. Our goal is to find Q and P such that $X = (Q, S, P)$ is the footskate-cleaned version. First, we compute $\tilde{F} = \Psi_\Theta(FK(\tilde{Q}, S, \tilde{P}))$ and $C = \Gamma(\tilde{F})$, being respectively the vGRF components estimated by our deep neural network Ψ_Θ and the foot contact labels calculated using the contact function Γ defined in Section 3.3. The function $FK(\cdot)$ refers to forward kinematics, i.e. the computation of joint positions from joint angles, skeleton and global trajectory. After initialising Q and P with $\tilde{Q}$ and $\tilde{P}$, we iteratively optimise both for a small fixed number N of iterations to best satisfy foot contact constraints into a gradient-based optimisation loop. In the following paragraphs we describe each term $\mathcal{L}_i$ of the our loss function

$$\mathcal{L} = \omega_q \mathcal{L}_{quat} + \omega_f \mathcal{L}_{foot} + \omega_t \mathcal{L}_{traj} + \omega_v \mathcal{L}_{vGRFs} \qquad (2)$$

where weights ω_i are hyperparameters to balance our objectives.

Quaternions loss $\mathcal{L}_{quat}$: we parameterise joint angles with unit quaternions, which are well-suited for such an optimisation since they are free of singularities, computationally efficient, and numerically stable [MHLC*21]. To keep valid rotations with quaternions throughout the optimisation, $\mathcal{L}_{quat}$ penalise quaternion norm deviations from 1:

$$\mathcal{L}_{quat} = \sum_t \sum_j \left(\left\| Q_j(t) \right\| - 1 \right)^2 \qquad (3)$$

where $Q_j(t)$ is the quaternion representing the orientation of joint j at time t.

Foot contacts loss $\mathcal{L}_{foot}$: satisfying foot contact constraints is not an easy task since we do not have directly access to actual contact locations w.r.t. the skeleton. Moreover, since *foot joints* are lo-
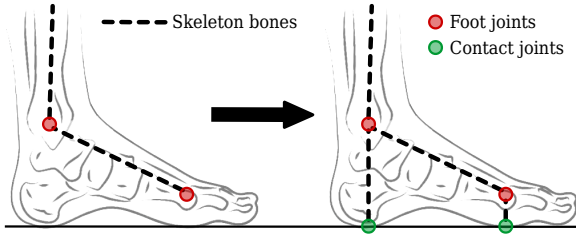
Figure 9: *Illustration of* contact joints *(green dots) inserted under corresponding* foot joints *at the ground level in the skeleton to properly enforce foot contact constraints (See details below).*

cated above actual contact points, they are allowed to rotate around the latter during part of contact phases (e.g. when the heel is in contact with the ground, the ankle is constrained to rotate around). To alleviate these issues, we artificially insert *contact joints* in S corresponding to contact points, i.e. under foot joints at the ground level (see figure 9). We assume that each contact joint is located vertically under its foot joint in the T-pose at the ground height, has constant orientation w.r.t. its foot joint, and is assumed to be static during contact phases. Then, imposing zero velocity on contact joints instead of foot joints during the contact phases better reflects the kinematics of the interaction with the ground, allowing slight rotation of foot joints located above the ground level. To this end, we minimise the mean squared distance from contact joints to contact positions during contact phases to faithfully constrain foot joint positions through forward kinematics:

$$\mathcal{L}_{foot} = \sum_{c=1}^{C} \sum_{t=t_0(c)}^{t_1(c)} \Pi(t) \left\| FK(Q,S,P)_{j_c}(t) - p_c \right\|^2 \qquad (4)$$

where j_c and p_c are respectively the joint in contact with the ground during contact phase c and its contact position during the contact phase spanning from $t_0(c)$ to $t_1(c)$. Moreover, function $\Pi(\cdot)$ is a rectangular wave function with smoothed edges (synchronised with contact phases) to avoid sharp foot position changes.

Trajectory loss $\mathcal{L}_{traj}$: since foot contact constraints are not guaranteed to be reachable by leg extensions, the global trajectory might be affected. However, it is closely related to body dynamics and its optimisation must be carefully controlled to avoid artefacts. In particular, constraining trajectory positions might introduce implausible velocity changes, hurting motion dynamics realism, e.g. slowing down to reach some contact position might lead to speeding up afterwards to catch up the target trajectory. To this end, we minimise the trajectory velocity deviations instead of position deviations:

$$\mathcal{L}_{traj} = \sum_{t} \left(\|V(t)\| - \|\tilde{V}(t)\| \right)^2, \quad V(t) = \frac{P(t+\Delta T) - P(t)}{\Delta T} \quad (5)$$

where $V(t)$ is the global velocity of root joint at time t, computed from global trajectory position P by numerical differentiation.

vGRFs invariance loss $\mathcal{L}_{vGRFs}$: the main novelty in our gradient-based optimisation approach is the use of our deep neural network through which gradients can flow. Since it is able to estimate consistent vGRF components from either clean or perturbed

inputs, we further guide the optimisation by minimising the deviation between initially estimated vGRFs components $\tilde{F}$ and dynamically estimated vGRFs components $F = \Psi_{\Theta}(FK(Q,S,P))$:

$$\mathcal{L}_{vGRFs} = MSLE(F, \tilde{F}) \qquad (6)$$

where MSLE is the mean squared logarithmic error, an alternative to mean squared error suited for vGRFs (see Section 4.3).

We used the same motion sequences blended for evaluation purposes in Section 5.4 to test the proposed footskate cleanup workflow. With parameters $N = 100$, $\omega_q = 10^{-3}$, $\omega_f = 10^{-5}$, $\omega_t = 10^2$ and $\omega_v = 5 \cdot 10^{-5}$ (see Equation 2), we achieve to significantly reduce the amount of footskate (the velocity of the feet during contact phases is approximately halved) while noticeably improving the realism of the sequences, assessed by visual inspection. We provide corresponding visual results of our footskate cleanup workflow in the supplementary video.

7. Conclusion

In this article, we proposed a novel approach that improves the state-of-art of foot contact detection and footskate cleanup in human character animation. Building on U̲NDER P̲RESSURE, a novel motion capture database synchronised with pressure insoles which we release with this article, we proposed to learn the relationship between human motion and interactions between feet and ground with a deep neural network estimating vertical ground reaction forces, from which foot contact constraints can be derived and exploited for an effective removal of foot sliding artefacts.

As of today, the proposed database is one of a kind in animation as it provides synchronised motion capture and pressure insoles data for a wide variety of human motion. Our approach significantly outperforms thresholds-based heuristic approaches in detecting foot contact labels, which are limited in their capability to generalise. We show its robustness to perturbed input motion sequences, which is crucial in concrete use cases. We have shown that estimating forces greatly helps the accurate and robust detection of contacts. Finally, we demonstrate the usefulness of our foot contact detection approach for footskate cleanup, leveraging vGRF estimations to improve the quality of the results.

As described in Section 2, foot contact labels are relevant to a lot of tasks in human animation and are typically obtained with simple heuristics, or manually annotated on an exceptional basis. Thus, many of these downstream tasks could probably benefit from the improved accuracy of our foot contact detection approach. Moreover, motion reconstruction relies more and more on modelling physical interaction of the feet with the ground (see Section 2.2). Resolving inverse dynamics problems like ground reaction forces estimation from motion capture data could provide a valuable regularization for such underconstrained problems as well as other tasks involving physics-based models. We believe our approach is a step in that direction.

References

[AG21] AG M. R.: Opengo sensor insole specification, 2021. Accessed: 2021-12-06. URL: https://moticon.com/wp-content/uploads/2021/09/

`OpenGo-Sensor-Insole-Specification-A4SQ-RGB-EN-03.02.pdf`. 4, 5

[ALL*20] ABERMAN K., LI P., LISCHINSKI D., SORKINE-HORNUNG O., COHEN-OR D., CHEN B.: Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics 39*, 4 (July 2020). `doi:10.1145/3386569.3392462`. 2

[ALX*19] ABDOLHOSSEINI F., LING H. Y., XIE Z., PENG X. B., VAN DE PANNE M.: On learning symmetric locomotion. In *International Conference on Motion, Interaction and Games* (Oct. 2019), Association for Computing Machinery (ACM). `doi:10.1145/3359566.3360070`. 2

[ATBO18] ANCILLAO A., TEDESCO S., BARTON J., O'FLYNN B.: Indirect measurement of ground reaction forces and moments by means of wearable inertial sensors: A systematic review. *Sensors 18*, 8 (Aug. 2018). `doi:10.3390/s18082564`. 3

[BB98] BINDIGANAVALE R., BADLER N. I.: Motion abstraction and mapping with spatial constraints. In *Modelling and Motion Capture Techniques for Virtual Environments* (Mar. 1998), Springer International Publishing, pp. 70–82. `doi:10.1007/3-540-49384-0_6`. 2

[CHS*19] CAO Z., HIDALGO G., SIMON T., WEI S.-E., SHEIKH Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence 43*, 1 (Jan. 2019), 172–186. `doi:10.1109/TPAMI.2019.2929257`. 3

[EKA*17] ELTOUKHY M., KUENZE C., ANDERSEN M. S., OH J., SIGNORILE J.: Prediction of ground reaction forces for parkinson's disease patients using a kinect-driven musculoskeletal gait analysis model. *Medical Engineering & Physics 50* (Oct. 2017), 75–82. `doi:10.1016/j.medengphy.2017.10.004`. 3

[FCK*15] FABER G. S., CHANG C.-C., KINGMA I., DENNERLEIN J. T., VAN DIEËN J. H.: Estimating 3d l5/s1 moments and ground reaction forces during trunk bending using a full-body ambulatory inertial motion capture system. *Journal of Biomechanics 49*, 6 (Nov. 2015), 904–912. `doi:10.1016/j.jbiomech.2015.11.042`. 3

[FHS*16] FRY A. C., HERDA T. J., STERCZALA A. J., COOPER M. A., ANDRE M. J.: Validation of a motion capture system for deriving accurate ground reaction forces without a force plate. *Big Data Analytics 1* (Sept. 2016). `doi:10.1186/s41044-016-0008-y`. 3

[HAB20] HENTER G. E., ALEXANDERSON S., BESKOW J.: Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics 39*, 6 (Nov. 2020). `doi:10.1145/3414685.3417836`. 2

[HBL11] HA S., BAI Y., LIU C. K.: Human motion reconstruction from force sensors. In *2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Aug. 2011), Association for Computing Machinery (ACM), p. 129–138. `doi:10.1145/2019406.2019424`. 3

[HKS17] HOLDEN D., KOMURA T., SAITO J.: Phase-functioned neural networks for character control. *ACM Transactions on Graphics 36*, 4 (July 2017). `doi:10.1145/3072959.3073663`. 2, 3, 5

[HSK16] HOLDEN D., SAITO J., KOMURA T.: A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics 35*, 4 (July 2016). `doi:10.1145/2897824.2925975`. 2, 9

[HYNP20] HARVEY F. G., YURICK M., NOWROUZEZAHRAI D., PAL C.: Robust motion in-betweening. *ACM Transactions on Graphics 39*, 4 (July 2020). `doi:10.1145/3386569.3392480`. 2, 3

[IAF06] IKEMOTO L., ARIKAN O., FORSYTH D.: Knowing when to put your foot down. In *2006 Symposium on Interactive 3D Graphics and Games* (Mar. 2006), Association for Computing Machinery (ACM), p. 49–53. `doi:10.1145/1111411.1111420`. 2, 3, 5

[JJLK14] JUNG Y., JUNG M., LEE K., KOO S.: Ground reaction force estimation using an insole-type pressure mat and joint kinematics during walking. *Journal of Biomechanics 47*, 11 (May 2014), 2693–2699. `doi:10.1016/j.jbiomech.2014.05.007`. 3

[JMNB19] JÖNSSON M., MUNKHAMMAR T., NORRBRAND L., BERG H. E.: Foot centre of pressure and ground reaction force during quadriceps resistance exercises; a comparison between force plates and a pressure insole system. *Journal of Biomechanics 87* (Mar. 2019), 206–210. `doi:10.1016/j.jbiomech.2019.03.004`. 8

[KBS*16] KARATSIDIS A., BELLUSCI G., SCHEPERS H. M., DE ZEE M., ANDERSEN M. S., VELTINK P. H.: Estimation of ground reaction forces and moments during gait using only inertial motion capture. *Sensors 17*, 1 (Dec. 2016). `doi:10.3390/s17010075`. 3

[KLvdP20] KWON T., LEE Y., VAN DE PANNE M.: Fast and flexible multilegged locomotion using learned centroidal dynamics. *ACM Transactions on Graphics 39*, 4 (July 2020). `doi:10.1145/3386569.3392432`. 2

[KSG02] KOVAR L., SCHREINER J., GLEICHER M.: Footskate cleanup for motion capture editing. In *2002 ACM SIGGRAPH/Eurographics symposium on Computer animation* (Mar. 2002), Association for Computing Machinery (ACM), p. 97–104. `doi:10.1145/545261.545277`. 2

[KSN14] KULBACKI M., SEGEN J., NOWACKI J. P.: 4gait: Synchronized mocap, video, grf and emg datasets: Acquisition, management and applications. In *6th Asian Conference on Intelligent Information and Database Systems* (Apr. 2014), Springer International Publishing, pp. 555–564. `doi:10.1007/978-3-319-05458-2_57`. 4

[LBMN15] LEPORACE G., BATISTA L. A., METSAVAHT L., NADAL J.: Residual analysis of ground reaction forces simulation during gait using neural networks with different configurations. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Nov. 2015), IEEE Computer Society, pp. 2812–2815. `doi:10.1109/EMBC.2015.7318976`. 3

[LCB06] LE CALLENNEC B., BOULIC R.: Robust kinematic constraint detection for motion data. In *2006 ACM SIGGRAPH/Eurographics symposium on Computer animation* (Mar. 2006), Association for Computing Machinery (ACM). URL: `https://dl.acm.org/doi/10.5555/1218064.1218103`. 2

[LCR*02] LEE J., CHAI J., REITSMA P. S. A., HODGINS J. K., POLLARD N. S.: Interactive control of avatars animated with human motion data. In *29th International Conference on Computer Graphics and Interactive Techniques* (Mar. 2002), Association for Computing Machinery (ACM). `doi:10.1145/566570.566607`. 2

[LLL18] LEE K., LEE S., LEE J.: Interactive character animation by learning multi-objective control. *ACM Transactions on Graphics 37*, 6 (Dec. 2018). `doi:10.1145/3272127.3275071`. 2

[LSC*19] LI Z., SEDLAR J., CARPENTIER J., LAPTEV I., MANSARD N., SIVIC J.: Estimating 3d motion and forces of person-object interactions from monocular video. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Mar. 2019), IEEE Computer Society/Computer Vision Foundation (CVF), pp. 8640–8649. `doi:10.1109/CVPR.2019.00884`. 3

[LZCvdP20] LING H. Y., ZINNO F., CHENG G., VAN DE PANNE M.: Character controllers using motion vaes. *ACM Transactions on Graphics 39*, 4 (July 2020). `doi:10.1145/3386569.3392422`. 2, 9

[MBR17] MARTINEZ J., BLACK M. J., ROMERO J.: On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017), IEEE Computer Society, pp. 4674–4683. `doi:10.1109/CVPR.2017.497`. 2

[MC12] MIN J., CHAI J.: Motion graphs++: a compact generative model for semantic motion analysis and synthesis. *ACM Transactions on Graphics 31*, 6 (Nov. 2012). `doi:10.1145/2366145.2366172`. 2, 3

[MHLC*21] MOUROT L., HOYET L., LE CLERC F., SCHNITZLER F., HELLIER P.: A survey on deep learning for skeleton-based human animation. *Computer Graphics Forum 41*, 1 (Nov. 2021). `doi:10.1111/cgf.14426`. 9

[MKD*20] MUNDT M., KOEPPE A., DAVID S., BAMER F., POTTHAST

W., MARKERT B.: Prediction of ground reaction force and joint moments based on optical motion capture data during gait. *Medical Engineering & Physics 86* (Oct. 2020), 29–34. doi:10.1016/j.medengphy.2020.10.001. 3

[MMPD21] MORIN P., MULLER A., PONTONNIER C., DUMONT G.: Foot contact detection through pressure insoles for the estimation of external forces and moments: application to running and walking. *Computer Methods in Biomechanics and Biomedical Engineering* (June 2021), 1–2. URL: https://hal.inria.fr/hal-03273616/document. 3

[MSZ*18] MASON I., STARKE S., ZHANG H., BILEN H., KOMURA T.: Few-shot learning of homogeneous human locomotion styles. *Computer Graphics Forum 37*, 7 (Oct. 2018), 143–153. doi:10.1111/cgf.13555. 2

[NSM11] NAKAZATO K., SCHEIBER P., MÜLLER E.: A comparison of ground reaction forces determined by portable force-plate and pressure-insole systems in alpine skiing. *Journal of Sports Science & Medicine 10*, 4 (Dec. 2011), 754–762. URL: https://pubmed.ncbi.nlm.nih.gov/24149570/. 8

[PFAG20] PAVLLO D., FEICHTENHOFER C., AULI M., GRANGIER D.: Modeling human motion with quaternion-based neural networks. *International Journal of Computer Vision 128*, 4 (Apr. 2020), 855–872. doi:10.1007/s11263-019-01245-6. 2

[PGA18] PAVLLO D., GRANGIER D., AULI M.: Quaternet: A quaternion-based recurrent model for human motion. In *British Machine Vision Conference 2018* (Sept. 2018), BMVA Press. URL: https://dblp.org/rec/conf/bmvc/PavlloGA18. 2

[PHO11] PRAŽÁK M., HOYET L., O'SULLIVAN C.: Perceptual evaluation of footskate cleanup. In *2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Aug. 2011), Association for Computing Machinery (ACM), p. 287–294. doi:10.1145/2019406.2019444. 2

[PRL*19] PARK S., RYU H., LEE S., LEE S., LEE J.: Learning predict-and-simulate policies from unorganized human motion data. *ACM Transactions on Graphics 38*, 6 (Nov. 2019). doi:10.1145/3355089.3356501. 2

[PvdP17] PENG X. B., VAN DE PANNE M.: Learning locomotion skills using deeprl: Does the choice of action space matter? In *16th ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (July 2017), Association for Computing Machinery (ACM). doi:10.1145/3099564.3099567. 2

[RFCA10] ROUHANI H., FAVRE J., CREVOISIER X., AMINIAN K.: Ambulatory assessment of 3d ground reaction force using plantar pressure distribution. *Gait & Posture 32*, 3 (June 2010), 311–316. doi:10.1016/j.gaitpost.2010.05.014. 3

[RGH*20] REMPE D., GUIBAS L. J., HERTZMANN A., RUSSELL B., VILLEGAS R., YANG J.: Contact and human dynamics from monocular video. In *16th European Conference on Computer Vision (ECCV)* (Sept. 2020), Springer International Publishing, pp. 71–87. doi:10.1007/978-3-030-58558-7_5. 3

[SAA*20] SHI M., ABERMAN K., ARISTIDOU A., KOMURA T., LISCHINSKI D., COHEN-OR D., CHEN B.: Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM Transactions on Graphics 40*, 1 (Sept. 2020). doi:10.1145/3407659. 2, 3

[SCNW19] SMITH H. J., CAO C., NEFF M., WANG Y.: Efficient neural networks for real-time motion style transfer. *Proceedings of the ACM on Computer Graphics and Interactive Techniques 2*, 2 (July 2019). doi:10.1145/3340254. 3, 7

[SGB18] SCHEPERS H. M., GIUBERTI M., BELLUSCI G.: *Xsens MVN: Consistent Tracking of Human Motion Using Inertial Sensing*. Tech. rep., Xsens Technologies B.V., P.O. Box 559, 7500 Enschede, The Netherlands, Mar. 2018. URL: https://www.researchgate.net/publication/324007368_Xsens_MVN_Consistent_Tracking_of_Human_Motion_Using_Inertial_Sensing. 4

[SGX*21] SHIMADA S., GOLYANIK V., XU W., PÉREZ P., THEOBALT C.: Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics 40*, 4 (July 2021). doi:10.1145/3450626.3459825. 3, 5

[SGXT20] SHIMADA S., GOLYANIK V., XU W., THEOBALT C.: Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics 39*, 6 (Dec. 2020). doi:10.1145/3414685.3417877. 3

[SRF*20] SCOTT J., RAVICHANDRAN B., FUNK C., COLLINS R. T., LIU Y.: From image to stability: Learning dynamics from human pose. In *16th European Conference on Computer Vision (ECCV)* (Sept. 2020), Springer International Publishing, pp. 536–554. doi:10.1007/978-3-030-58592-1_32. 3, 4

[SZKS19] STARKE S., ZHANG H., KOMURA T., SAITO J.: Neural state machine for character-scene interactions. *ACM Transactions on Graphics 38*, 6 (Nov. 2019). doi:10.1145/3355089.3356505. 2, 3, 9

[SZKZ20] STARKE S., ZHAO Y., KOMURA T., ZAMA K.: Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics 39*, 4 (July 2020). doi:10.1145/3386569.3392450. 2, 9

[WCX19] WANG Z., CHAI J., XIA S.: Combining recurrent neural networks and adversarial training for human motion synthesis and control. *IEEE Transactions on Visualization and Computer Graphics 27*, 1 (Sept. 2019), 14–28. doi:10.1109/TVCG.2019.2938520. 2

[WHSZ19] WANG H., HO E. S.-L., SHUM H. P. H., ZHU Z.: Spatio-temporal manifold learning for human motions via long-horizon modeling. *IEEE Transactions on Visualization and Computer Graphics 27*, 1 (Aug. 2019). doi:10.1109/TVCG.2019.2936810. 2

[WL19] WON J., LEE J.: Learning body shape variation in physics-based characters. *ACM Transactions on Graphics 38*, 6 (Nov. 2019). doi:10.1145/3355089.3356499. 2

[XLKvdP20] XIE Z., LING H. Y., KIM N. H., VAN DE PANNE M.: Allsteps: Curriculum-driven learning of stepping stone skills. *Computer Graphics Forum 39*, 8 (Oct. 2020). doi:10.1111/cgf.14115. 2

[XWCC15] XIA S., WANG C., CHAI J., CHAI J.: Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics 34*, 4 (July 2015). doi:10.1145/2766999. 2

[YKL21] YANG D., KIM D., LEE S.-H.: Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals. *Computer Graphics Forum 40*, 2 (June 2021), 265–275. doi:10.1111/cgf.142631. 3

[YTL18] YU W., TURK G., LIU C. K.: Learning symmetric and low-energy locomotion. *ACM Transactions on Graphics 37*, 4 (July 2018). doi:10.1145/3197517.3201397. 2

[ZRW20] ZELL P., ROSENHAHN B., WANDT B.: Weakly-supervised learning of human dynamics. In *16th European Conference on Computer Vision (ECCV)* (Sept. 2020), Springer International Publishing, pp. 68–84. doi:10.1007/978-3-030-58574-7_5. 3

[ZSKS18] ZHANG H., STARKE S., KOMURA T., SAITO J.: Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics 37*, 4 (July 2018). doi:10.1145/3197517.3201366. 2, 9

[ZSZ*14] ZHANG P., SIU K., ZHANG J., LIU C. K., CHAI J.: Leveraging depth cameras and wearable pressure sensors for full-body kinematics and dynamics capture. *ACM Transactions on Graphics 33*, 6 (Nov. 2014). doi:10.1145/2661229.2661286. 3

[ZWR17] ZELL P., WANDT B., ROSENHAHN B.: Joint 3d human motion capture and physical analysis from monocular videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (July 2017), IEEE Computer Society, pp. 17–26. doi:10.1109/CVPRW.2017.9. 3

[ZYC*20] ZOU Y., YANG J., CEYLAN D., ZHANG J., PERAZZI F., HUANG J.-B.: Reducing footskate in human motion reconstruction with

ground contact constraints. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)* (Mar. 2020), IEEE Computer Society, pp. 448–457. doi:10.1109/WACV45572.2020.9093329. 3