4D-Foot: A Fully Automated Pipeline of Four-Dimensional Analysis of the Foot Bones Using Bi-plane X-Ray Video and CT

Shuntaro Mizoe[1], Yoshito Otake[1(✉)], Takuma Miyamoto[2], Mazen Soufi[1],
Satoko Nakao[2], Yasuhito Tanaka[2], and Yoshinobu Sato[1]

[1] Division of Information Science, Nara Institute of Science
and Technology, Ikoma, Japan
otake@is.naist.jp

[2] Department of Orthopedics, Nara Medical University, Kashihara, Japan

Abstract. We aim to elucidate the mechanism of the foot by automated measurement of its multiple bone movement using 2D-3D registration of bi-plane x-ray video and a stationary 3D CT. Conventional analyses allowed tracking of only 3 large proximal tarsal bones due to the requirement of manual segmentation and manual initialization of 2D-3D registration. The learning-based 2D-3D registration, on the other hand, has been actively studied and demonstrating a large capture range, but the accuracy is inferior to conventional optimization-based methods. We propose a fully automated pipeline using a cost function that seamlessly incorporates the reprojection error at the landmarks in CT and x-ray detected by off-the-shelf CNNs into the conventional image similarity cost, combined with the automated bone segmentation. We experimentally demonstrated that the pipeline allowed a robust and accurate 2D-3D registration to track all 12 tarsal bones, including the metatarsals at the foot arch, which is especially important in the foot biomechanics but has been unmeasurable with previous methods. We evaluated the proposed fully automated pipeline in studies using a bone phantom and real x-ray images of human subjects. The real image study showed the registration error of 0.38 ± 1.95 mm in translation and $0.38 \pm 1.20^\circ$ in rotation for the proximal tarsal bones.

Keywords: 2D/3D registration · Automated segmentation · Automated initialization

1 Introduction

The foot consists of flexible structures of bones, joints, muscles, and soft tissues, allowing complex movements and shock absorption in human motion. We aim to accurately track the foot bones for biomechanical analysis (e.g., interaction between the small bones at multiple joints). While its importance is acknowledged, especially in injury prevention and rehabilitation of the ankle disease,

S. Mizoe and Y. Otake—Equal contribution.

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12904, pp. 182–192, 2021.
https://doi.org/10.1007/978-3-030-87202-1_18

most conventional methods are limited to either static anatomical analyses using CT [1,2] or skin-marker-based motion capture [3–5] which is prone to error due to the skin movement.

Some recent studies employ a 2D-3D registration between x-ray videos acquired by a biplane imaging system and a CT image for the analysis of the 3D bone movement [6,7]. The approach demonstrated a high accuracy, however, the target bones have been limited to only the proximal tarsal bones, namely talus, calcaneus, and navicular bones, and the methods required laborious manual segmentation of each bone from the CT and manual initialization of the 2D-3D registration. While Esteban et al. [8] and Grupp et al. [9] studied 2D-3D registration in the analysis of pelvis anatomy using a CNN-based landmark detection for initialization of the intensity-based registration, both works assumed manual segmentation of the target anatomy in CT, which is prohibitive especially in the clinical analysis of the foot bones. On the other hand, several attempts have been made by training CNNs for directly solving the 2D-3D alignment in an end-to-end manner (e.g. [10,11]), showing better stability due to a large capture range but inferior accuracy compared to the conventional intensity-based method. Our approach is to achieve stable and accurate registration using a cost function incorporating similarities of both intensity and landmark positions, unlike landmark only for the initialization.

We propose a fully automated pipeline of 2D-3D registration between x-ray video and CT for the motion analysis of all 12 tarsal bones (i.e., 3 proximal tarsal, 4 distal tarsals, and 5 metatarsal bones) and tibia-fibula (as one rigid object). The contribution of this paper is threefold: (1) Proposal of a 4D foot analysis system including the movement of the foot arch (metatarsal bones) which was previously unmeasurable, (2) introduction of a cost term in 2D-3D registration that incorporates reprojection error of the landmarks detected by CNNs allowing robust and accurate registration without any manual interactions, (3) quantitative evaluation of impacts of the errors in automated segmentation and landmark detection on the final registration accuracy.

2 Method

2.1 Overview of the Proposed Pipeline

Figure 1 shows the overview of the proposed pipeline. The input CT and biplane x-ray videos are first processed by CNNs, Bayesian U-net [14] for bone segmentation and landmark extraction in CT, and DeepLabCut [13] for landmark extraction in x-ray video. Then the intensity-based 2D-3D registration is performed frame-by-frame using the proposed cost terms incorporating information of the landmark and intensity similarities, resulting in a robust and accurate registration for multiple small bones in the foot.

2.2 Automated Segmentation and Landmark Detection

Segmentation of each lower leg and foot bones (2 lower leg bones, 7 tarsal bones, 5 metatarsal bones, and 14 phalanges bones) in CT is performed by

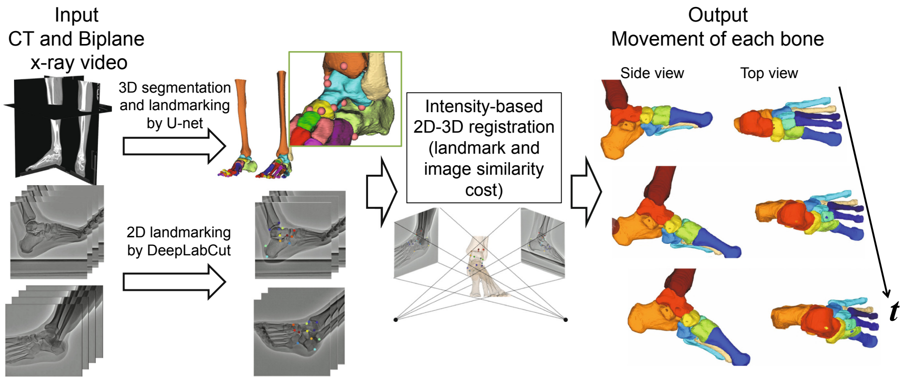
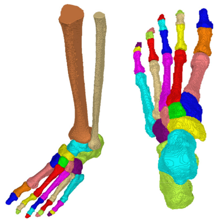


Fig. 1. Overview of the proposed automated pipeline for 4D analysis of the foot bones. The CT images and biplane x-ray video are automatically annotated (segmentation and landmarking) using CNNs and the movement of each tarsal and metatarsal bones are estimated using the proposed intensity-based 2D-3D registration.

the Bayesian U-net [14], that previously demonstrated a significantly superior accuracy than the previous multi-atlas method in segmentation of the hip and thigh muscles and bones. Our implementation including the network architecture, hyper-parameters, and pre- and post-processing follows [14][1] except for the size of convolution kernel of 7 × 7 for leveraging a larger receptive field. Figure 2 shows detail of the target bones.



Bone name	Color	Bone name	Color	Bone name	Color
Tibia	🟧	1st metatarsal	🟥	2nd middle phalanges	🟨
Fibula	🟨	2nd metatarsal	🟥	3rd middle phalanges	🟦
Talus	🟦	3rd metatarsal	🟪	4th middle phalanges	🟩
Calcaneus	🟩	4th metatarsal	🟨	5th middle phalanges	🟨
Navicular	🟨	5th metatarsal	🟦	1st distal phalanges	🟦
Medial cuneiform	🟦	1st proximal phalanges	🟧	2nd distal phalanges	🟩
Intermediate cuneiform	🟩	2nd proximal phalanges	🟨	3rd distal phalanges	🟥
Lateral cuneiform	🟥	3rd proximal phalanges	🟦	4th distal phalanges	🟥
Cuboid	🟪	4th proximal phalanges	🟥	5th distal phalanges	🟪
		5th proximal phalanges	🟪		

Fig. 2. List of the foot bones annotated in this study. (note that the phalanges bones are not included in our 2D-3D registration analysis due to the limited field-of-view of the x-ray video).

[1] Source code was obtained from https://github.com/yuta-hi/bayesian_unet.

2.3 2D-3D Registration Incorporating Landmark Reprojection Error

The intensity-based 2D-3D registration optimizes similarity between the x-ray image (fixed image) and digitally reconstructed radiograph (DRR) generated from CT (moving image). DRRs were generated using the tri-linear interpolation ray-tracing algorithm [12] implemented on the graphics processing unit (GPU). In this study, we parameterized the rigid transformation of each bone with a 6 degree-of-freedom variable (3 rotation parameters represented as Euler angle around the geometrical centroid of each bone and 3 translation parameters), resulting in a $6N$ parameter optimization problem for N bones. Following [12], we employed covarience matrix adaptation evolutionary strategy (CMA-ES) [15] for optimization and the gradient correlation similarity measure [16] for the cost function. Initialization of the translation parameters was derived by the paired point registration of the landmarks for each frame independently, assuming all bones moved rigidly. The registration of 14 bones (bones in Fig. 2 except for the 14 phalanges bones) was split into 3 stages, 1) proximal tarsal, tibia, and fibula (5 bones), distal tarsal (4 bones), and metatarsals (5 bones), to reduce the optimization parameters. The proposed cost function incorporating the landmark reprojection error derived from CNNs and the conventional image similarity is defined as follows.

$$\begin{aligned}\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}}\{&(1-\alpha)C_{landmark}(p_i^{2D}, p_i^{3D}, \boldsymbol{\Theta}) \\ &-\alpha GC(I^{Xp}, \sum_{k=1}^{N} I_k^{DRR}(\boldsymbol{\Theta})) + \lambda g_{rigidity}(\boldsymbol{\Theta})\}\end{aligned} \tag{1}$$

The parameter α changes balance between the two data fitness terms, the landmark fitness and the image fitness defined by the gradient correlation (denoted by GC) between the X-ray image I^{Xp} and sum of DRRs of each bone I_k^{DRR}. The third term encourages rigidity of the target bones and λ is the weight parameter. The rigidity term was effective only for the bones with no landmark identified, such as metatarsal bones in this study. $C_{landmark}(p_i^{2D}, p_i^{3D}, \boldsymbol{\Theta}) = \sum_{i=1}^{M} ||p_i^{2D} - P(T(\boldsymbol{\Theta}))p_i^{3D}||$ represents the reprojection error of i_{th} landmark, where p_i^{2D} and p_i^{3D} are the landmark location identified by CNNs in 2D and 3D. $g_{rigidity}(\boldsymbol{\Theta}) = \sum_{k=2}^{N} d(T_1(\boldsymbol{\Theta}), T_k(\boldsymbol{\Theta}))$, $T_k(\boldsymbol{\Theta})$ is the transformation of k_{th} bone, $P(T_k)$ is the projection matrix with the extrinsic parameter defined by T_k, and $d(T_1, T_k)$ indicates difference between the two transformations (in our implementation, assuming small difference, we first concatenate T_1 and T_k^{-1}, convert it to 3 translation and 3 rotation parameters, and calculate Euclidean distance between the two 6-element vectors). Our implementation of the 2D-3D registration is available at https://github.com/YoshitoOtake/4DFoot.

3 Experiment and Results

After evaluation of the accuracy of individual automated segmentation and landmark detection components by the cross-validation, accuracy of the 2D-3D registration was evaluated using; 1) the bone phantom with metallic beads attached to 14 anatomical landmarks, providing the ground truth using the radio stereometric analysis, and 2) the images from 5 volunteer subjects with fully manual annotations. Firstly, using the ground truth in the phantom image, we validate that registration using manually annotated segmentation and landmarks can be used as the quasi-ground-truth. Then, using the manually annotated quasi-ground-truth, we evaluate the accuracy of the proposed fully automated pipeline for real subjects' images.

3.1 Experimental Materials

Thirty-five CTs of the lower leg and the foot obtained from 35 patients, and 18 biplane x-ray videos of the foot during the gait obtained from 5 healthy volunteers, were used in the experiment. The phase from heel contact to toe-off was manually identified by an expert surgeon and used in the experiment. The field of view of the CTs was 323–486 mm^2, the matrix size was 512 × 512, and the slice interval was 0.625 mm. All individual bone regions shown in Fig. 2 and 17 anatomical landmarks (on the tibia and 3 proximal tarsal bones) in the CTs and 12 landmarks (on the same bone for each view) in all frames of the x-ray video were manually annotated by an expert orthopedic surgeon. Since we could not find a sufficient number of 3D landmarks visible in two views simultaneously, 5 landmarks were used only in one x-ray view, the other 5 were used only in the other view, and the remaining 7 were used in both views. Thus, $(5 + 7) = 12$ landmarks were used in each 2D view, which amounts to 17 in 3D. The biplane x-ray imager was equipped so that the two views are aligned to the patient's right-left direction (referred to as *lateral view*) and the oblique direction (referred to as *oblique view*). The distance between the x-ray source and detector was approximately 1200 mm for both views. The matrix size of the x-ray image was 512 × 512, and the pixel spacing was 0.558 × 0.558 mm. Geometric calibration of the two imagers was performed by obtaining 12 x-ray images of a cube-shaped calibration phantom (edge length of 110 mm) having 8 metallic spheres of 10 mm diameter at each corner. In our system, the two x-ray views were not synchronized. They record images alternately at 15 fps with half a frame (1/30 sec) phase offset. In order to obtain a *pseudo synchronized* pair of videos, a CNN-based video interpolation method, SuperSloMo [18], with a pre-trained model was used to double the frame rate of each video.

3.2 Evaluation of Automated Segmentation and Landmark Detection

Three-fold cross-validation using the 35 CTs was performed to evaluate the segmentation accuracy. In training, the right and left sides of the foot were split at the middle of the axial slice, and the right foot was flipped for the data augmentation purpose (note that the training/test split in the cross-validation was performed patient-wise since the right and left side of the same patient are similar). The dice coefficient for each bone used in the 2D-3D registration in the following experiments is summarized in Fig. 3. The dice for the lower leg, tarsal bones, and metatarsal bones were 0.990 ± 0.012, 0.971 ± 0.053, and 0.975 ± 0.022. The phalanges bones were not used in the 2D-3D registration but included in the segmentation target. The dice coefficients were 0.956 ± 0.050, 0.847 ± 0.154, and 0.794 ± 0.210 for phalanx proximalis, medialis, and distalis.

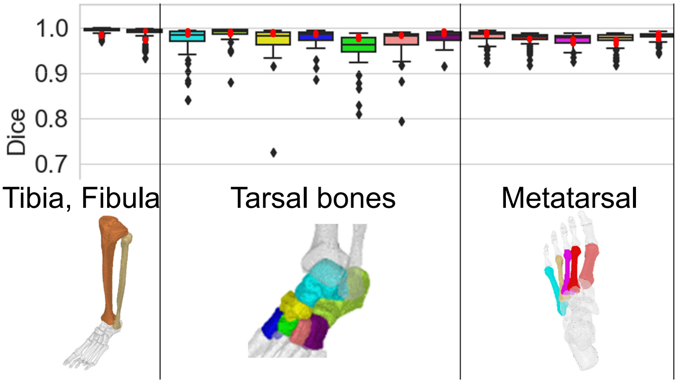


Fig. 3. Results of automated segmentation of the foot bones. Red dots indicates the five cases that were used in the 2D-3D registration experiment.

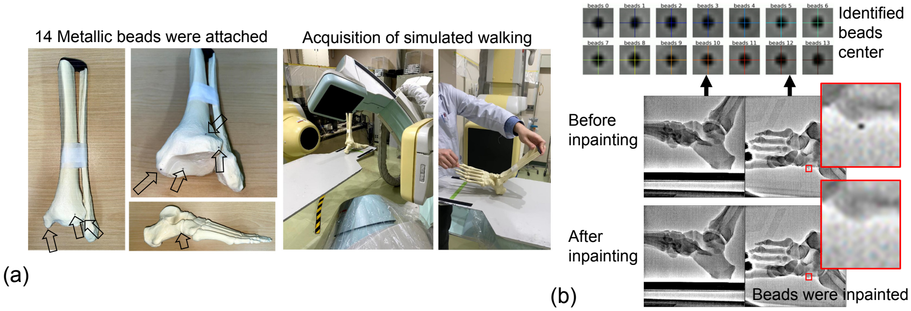


Fig. 4. Accuracy evaluation experiment using the bone phantom. (a) Experimental setting, and (b) preprocessing of the x-ray videos. The ground truth was obtained by radio stereometric analysis (RSA) using the metallic beads attached to the bones. Metallic beads in the x-ray image and CT were removed by inpainting in order to avoid the bias in the 2D-3D registration due to the strong gradient created by the beads.

The landmark detection in CT was performed by the U-net using the heatmap approach [17] with the σ (radius of the Gaussian representing the landmark) of 5 mm. As the result of three-fold cross-validation, the Euclidean distance errors of landmarks on the tibia, talus, calcaneus, and navicular were 4.27 ± 2.26, 3.65 ± 2.04, 4.07 ± 2.01, and 4.13 ± 2.41 mm, respectively.

The landmark detection in the x-ray video was performed using DeepLabCut [13]. We used the pre-trained Resnet-50 with fine-tuning using our own training data set. The leave-one-patient-out evaluation demonstrated the average Euclidean distance error of all the landmarks for the lateral and oblique view was 3.01 ± 2.29 and 2.73 ± 2.00 mm, respectively.

The landmark detection errors in CT and x-ray video were comparable to those reported in [17], where the authors applied their state-of-the-art method in the spine CT data set.

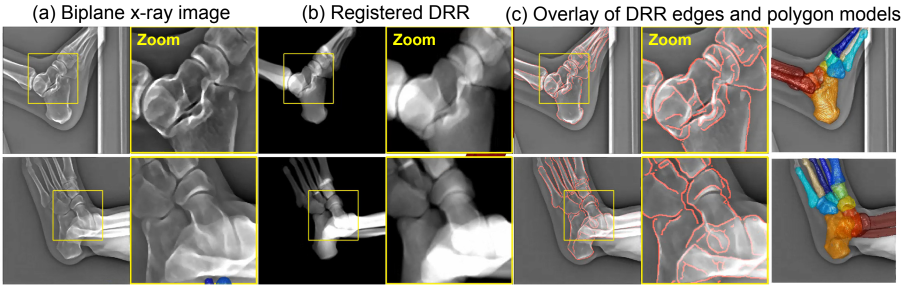


Fig. 5. A representative registration result. The intensity-based 2D-3D registration optimized the similarity between (a) the original biplane x-ray image and (b) DRR. The overlay of the DRR edges and polygon models in (c) demonstrate the accurate alignment between the two images indicating the 3D position of each bone was correctly estimated.

3.3 Evaluation of 2D-3D Registration Using Bone Phantom

The bone phantom and its x-ray videos used in the experiment were shown in Fig. 4. The phantom was moved by hands to simulate the gate. The 14 metallic beads attached to the phantom were localized in the x-ray images first manually and then refined by the Gaussian fitting search at their vicinity. To avoid the strong image gradient at the edge of the beads affecting registration accuracy, the bead regions were inpainted [19] (see Fig. 4b). The localized beads position with the geometric calibration provided the ground truth movement of each bone, while the inpainted x-ray videos were used for the 2D-3D registration. The experimental results were shown in Fig. 6a. The average absolute translation error was 0.40 ± 0.28 mm, and the rotation error was 0.66 ± 0.59°. Thus, we confirmed that 2D-3D registration based on manual annotation is of a level that can be used as a quasi-ground-truth in terms of the clinically required accuracy. The larger error in the navicular bone was likely attributed to its small size and rotationally symmetric shape. Relatively larger error in Y translation and smaller error in X rotation could be attributed to the sensitivity to the imaging direction (i.e., movement in the out-of-plane direction is less sensitive to the in-plane direction).

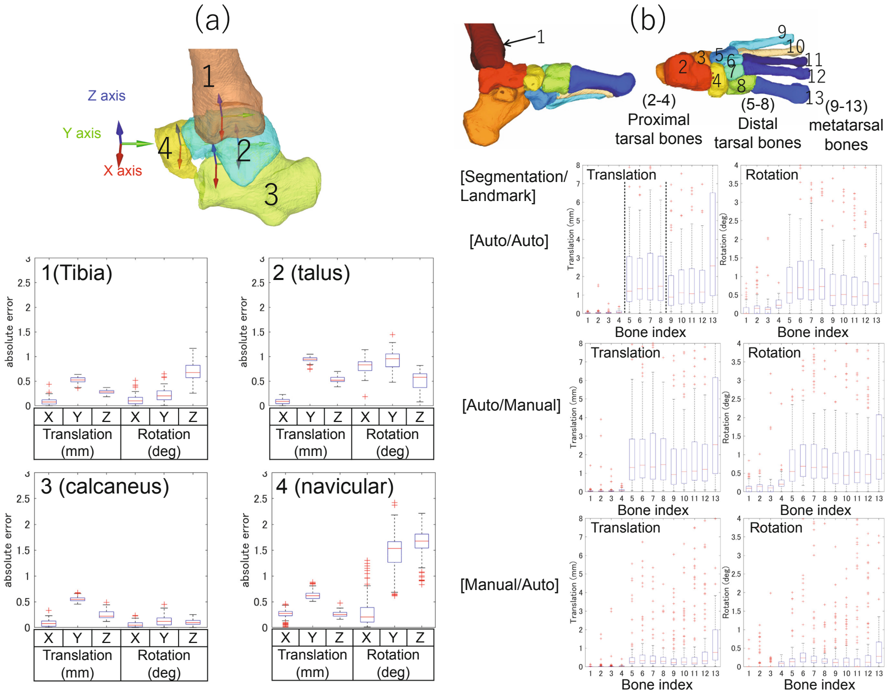


Fig. 6. Quantitative evaluation of the 3D tracking of each bone for the experiment with (a) the bone phantom and (b) real subject images.

3.4 Evaluation of 2D-3D Registration Using Images of Real Subjects

Figure 5 demonstrates a representative registration result using real subject images. DRR at the registered position correctly aligned with the x-ray image providing a visual assessment of the registration accuracy. Figure 6b and Table 1 show the quantitative results. As described above, the registration result using manual segmentation and manual landmark detection was used as the quasi-ground-truth in this experiment. In order to evaluate the effect of using automated annotation in the registration, the results in three scenarios were compared, 1) automated segmentation (Auto)/automated landmark detection (Auto), 2) Auto/Manual, 3) Manual/Auto. Overall, registration of proximal tarsal bones showed excellent accuracy (<0.5-mm translation and <0.5-degree rotation), comparable to the bone phantom experiment regardless of the annotation method. The insensitivity of the registration results to the landmark detection error suggests that the error was in an acceptable range in our 2D-3D registration application.

The distal tarsal bones and metatarsal bones showed relatively lower accuracy (∼3 mm translation and ∼1.5° rotation), especially when we used automated segmentation, likely due to their small size increasing sensitivity to the segmentation error. Parameters for the CMA-ES optimizer were: population size 1000, stopping criterion 0.01 (mm or deg), the two-level multi-resolution pyramid with down-sampling by a factor of 2 and 1. One registration trial required approximately 60,000 function evaluations (i.e., DRR generation and cost calculation), and the computation time was approximately 20 s on a workstation with AMD EPYC 7742 64-core processor and two nVidia GeForce RTX3090.

Table 1. Comparison of the registration accuracy using automated- and manual- segmentation and landmark (trans: translation error, rot: rotation error, bone phantom experiment used manual segmentation and manual landmark)

Segmentation/ Landmark	Proximal tarsal bones				Distal tarsal bones				Metatarsal bones			
	trans(mm)		rot(deg)		trans(mm)		rot(deg)		trans(mm)		rot(deg)	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Auto/Auto	0.38	1.95	0.38	1.20	3.09	4.46	1.41	2.06	3.25	7.79	1.30	2.88
Auto/Manual	0.15	0.90	0.27	0.90	2.79	4.16	1.28	1.95	3.04	8.81	1.21	2.95
Manual/Auto	0.39	2.07	0.34	1.48	1.66	4.93	0.72	1.98	1.88	9.04	0.63	2.79
(bone phantom)	0.40	0.28	0.66	0.59	—	—	—	—	—	—	—	—

4 Discussion and Conclusion

We have presented a fully automated pipeline of segmentation and 2D-3D registration for 4D analysis of the foot bones and evaluated the accuracy with fully manually annotated data sets. Our primary contribution has been the proposal and quantitative evaluation of the registration cost incorporating reprojection error at landmarks derived by CNNs with a conventional image similarity cost. We showed that the combination of simple off-the-shelf CNN-based image recognition and the conventional intensity-based registration allowed highly accurate 4D tracking of the complex movement of small foot bones, including the foot arch, whose shock absorption function is critical in the analysis of foot biomechanics but has been unmeasurable with previous methods. Furthermore, the experiment suggested that the error in the automated segmentation had a larger impact on the registration accuracy than the landmark detection error, especially for the distal part, namely the distal tarsal and metatarsal bones, which is small in size and symmetric in shape. The lower accuracy in those bones is attributed partly to the lack of landmarks since our current landmarks are placed only on the proximal tarsal bones as shown in Fig. 1 and the distal bones are associated with landmarks placed on the bones close to them. We plan to add several landmarks on those distal parts to improve accuracy. Application in a clinical routine and the analysis of patients with ankle disease are also underway.

References

1. Stanković, K., Booth, B.G., Danckaers, F., Burg, F., Vermaelen, P., Duerinck, S., et al.: Three-dimensional quantitative analysis of healthy foot shape: a proof of concept study. J. Foot Ankle Res. **11**(1), 8 (2018)
2. Nozaki, S., Watanabe, K., Kamiya, T., Katayose, M., Ogihara, N.: Three-dimensional morphological variations of the human calcaneus investigated using geometric morphometrics. Clin. Anat. **33**(5), 751–758 (2020)
3. Eichelberger, P., Blasimann, A., Lutz, N., Krause, F., Baur, H.: A minimal markerset for three-dimensional foot function assessment: measuring navicular drop and drift under dynamic conditions. J. Foot Ankle Res. **11**(1), 15 (2018)
4. Kim, T., Park, J.C.: Short-term effects of sports taping on navicular height, navicular drop and peak plantar pressure in healthy elite athletes: a within-subject comparison. Med. (United States) **96**(46), 3–8 (2017)
5. Behling, A.V., Manz, S., von Tscharner, V., Nigg, B.M.: Pronation or foot movement - what is important. J. Sci. Med. Sports **23**(4), 366–371 (2020)
6. Cao, S., et al.: In vivo kinematics of functional ankle instability patients during the stance phase of walking. Gait Posture **73**, 262–268 (2019)
7. Lenz, A.L., et al.: Compensatory motion of the subtalar joint following tibiotalar arthrodesis. J. Bone Joint Surg. **102**(7), 600–608 (2020)
8. Esteban, J., Grimm, M., Unberath, M., Zahnd, G., Navab, N.: Towards fully automatic X-Ray to CT registration. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 631–639. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_70
9. Grupp, R.B., et al.: Automatic annotation of hip anatomy in fluoroscopy for robust and efficient 2D/3D registration. Int. J. Comput. Assist. Radiol. Surg. **15**(5), 759–769 (2020)
10. Gao, C., et al.: Generalizing spatial transformers to projective geometry with applications to 2D/3D registration. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 329–339. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_32
11. Miao, S., et al.: Dilated FCN for multi-agent 2D/3D medical image registration. In: 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, pp. 4694–4701 (2018)
12. Otake, Y., et al.: Intraoperative image-based multiview 2D/3D registration for image-guided orthopaedic surgery: incorporation of fiducial-based C-arm tracking and GPU-acceleration. IEEE Trans. Med. Imaging **31**(4), 948–962 (2012)
13. Mathis, A., et al.: DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat. Neurosci. **21**(9), 1281–1289 (2018)
14. Hiasa, Y., Otake, Y., Takao, M., Ogawa, T., Sugano, N., Sato, Y.: Automated muscle segmentation from clinical CT using Bayesian U-Net for personalized musculoskeletal modeling. IEEE Trans. Med. Imaging **39**(4), 1030–1040 (2019)
15. Nikolaus, H.: The CMA evolution strategy: a comparing review. In: Lozano, J., et al. (ed.) Towards a New Evolutionary Computation, vol. 192, pp. 75–102. Springer, Heidelberg (2006). https://doi.org/10.1007/3-540-32494-1_4
16. Penney, G.P., et al.: A comparison of similarity measures for use in 2-D-3-D medical image registration. IEEE Trans. Med. Imaging **17**(4), 586–595 (1998)
17. Payer, C., Štern, D., Bischof, H., Urschler, M.: Integrating spatial configuration into heatmap regression based CNNs for landmark localization. Med. Image Anal. **54**, 207–219 (2019)

18. Jiang, H., Sun, D., Jampani, V., Yang, M.-H., Learned-Miller, E., Kautz, J.: Super SloMo: high quality estimation of multiple intermediate frames for video interpolation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9000–9008 (2018)
19. Bertalmio, M., Bertozzi, A.L., Sapiro, G.: Navier-stokes, fluid dynamics, and image and video inpainting. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, p. I (2001)