SYNTHESISING 2D VIDEOS FROM 3D DATA: ENLARGING SPARSE 2D VIDEO DATASETS FOR MACHINE LEARNING APPLICATIONS

Marion Mundt[1], Molly Goldacre[1], Jacqueline Alderson[1,2]

Minderoo Tech & Policy Lab, The University of Western Australia, Perth, Australia[1]
Auckland University of Technology, Sports Performance Research Institute New Zealand (SPRINZ), Auckland, New Zealand[2]

This study outlines a technique to repurpose widely available high resolution three-dimensional (3D) motion capture data for training a machine learning model to estimate the ground reaction forces from two-dimensional (2D) pose estimation keypoints. Keypoints describe anatomically related landmarks in 2D image coordinates. The landmarks can be calculated from 3D motion capture data and projected to different image planes, serving to synthesise a near-infinite number of 2D camera views. This highly efficient method of synthesising 2D camera views can be used to enlarge sparse 2D video databases of sporting movements. We show the feasibility of this approach using a sidestepping dataset and evaluate the optimal camera number and location required to estimate 3D ground reaction forces. The method presented and the additional insights gained from this approach can be used to optimise corporeal data capture by sports practitioners.

KEYWORDS: artificial neural networks, 2D pose estimation, enlarging datasets.

INTRODUCTION: Three-dimensional (3D) motion capture in laboratory settings has been the common historical approach to collect gold-standard, accurate and reliable, kinematic and kinetic motion data (Winter, 1991). However, in these settings internal and external factors that influence an athlete's performance in training or competition cannot be captured. Additionally, the commonly used retro-reflective markers affixed to an athlete's body, together with the requirement to contact a ground embedded force plate without targeting, leaves open questions surrounding the ecological validity of laboratory collected data. Alongside this limitation is the ever-increasing desire for technologies that enable the collection of accurate and reliable kinematic and kinetic athlete data in training and competition environments (Morris, Mundt, Goldacre, Weber, Mian & Alderson, 2021).

Two-dimensional (2D) video has been extensively used by sports practitioners to provide athletes with qualitative feedback regarding their performance (Nolan, Patritti, & Simpson, 2012). With advances in camera technologies resulting in the proliferation of low cost, high-resolution cameras with adjustable sampling frequency, this technology makes 2D biomechanical analysis even more accessible. The rise of computer vision machine learning (ML) tools in sport further stimulates this trend. Open-access pose estimation models are used to automatically annotate anatomically related landmarks (keypoints) in images and videos to provide quantitative motion kinematic information (Cao, Hidalgo, Simon, Wei & Sheikh, 2019). Although 2D biomechanical analyses using existing computer vision pose estimation models need to be undertaken with caution, these methods do provide a first step on the pathway towards automated higher resolution on-field motion analysis (Cronin, 2021). As an example, Morris and colleagues recently used keypoints estimated from videos to train a ML model to estimate 3D ground reaction forces (GRFs) with good accuracy (*r* = 0.825) (Morris et al., 2021). Applying ML methods to sports biomechanics problems presents two major initial challenges. First, these methods are reliant on large input datasets; indeed dataset sizes that are rarely encountered in the sports biomechanics discipline, e.g. most pose estimation models are trained on an image dataset of 250.000 people. The second challenge is a lack of generalisability, in that ML models only provide reliable estimates for biomechanical related motion data that has been used to train the model (for example, a model that has been trained to estimate joint kinetics for walking cannot be used to estimate these parameters for sidestepping). Unfortunately, there exist no large sports related 2D video datasets with concurrently collected 3D kinematic and kinetic information available for public use. This problem is only further exacerbated when we consider the data available for very specific movement tasks such as those found in elite sport.

To address these challenges, this paper presents a method to synthesise 2D video keypoints across a range of camera views using only 3D motion capture data. The output keypoints are used to train ML models to estimate 3D GRFs during sidestepping tasks. The optimal camera number and camera location (position and height) for this specific dataset is also investigated, to provide practical information to inform best practice future corporeal video data capture sessions.

METHODS: The dataset comprised 135 (63 to the right, 72 to the left direction) 45-degree sidestepping trials captured from 15 female (semi-) professional Australian Rules Football players (23±3.7 years, 167±4.0 cm, 62.9 ±5.6 kg). All participants were injury free and provided informed consent prior to testing in accordance with UWA Human Ethics approval (RA/4/1/2593). A 23-camera Vicon T40 system (Vicon Peak, Oxford Metrics, Oxford, UK, 200Hz) and a 1.2m x 1.2m synchronised AMTI force plate (Advanced Mechanical Technology Inc., Watertown, MA, 2000 Hz) were used to capture 3D motion and force data. Participants were affixed with 67 retro-reflective markers as per a custom marker set and model (Besier, Sturnieks, Alderson & Lloyd, 2003). Marker and force data were filtered using a 2nd order low-pass Butterworth filter.

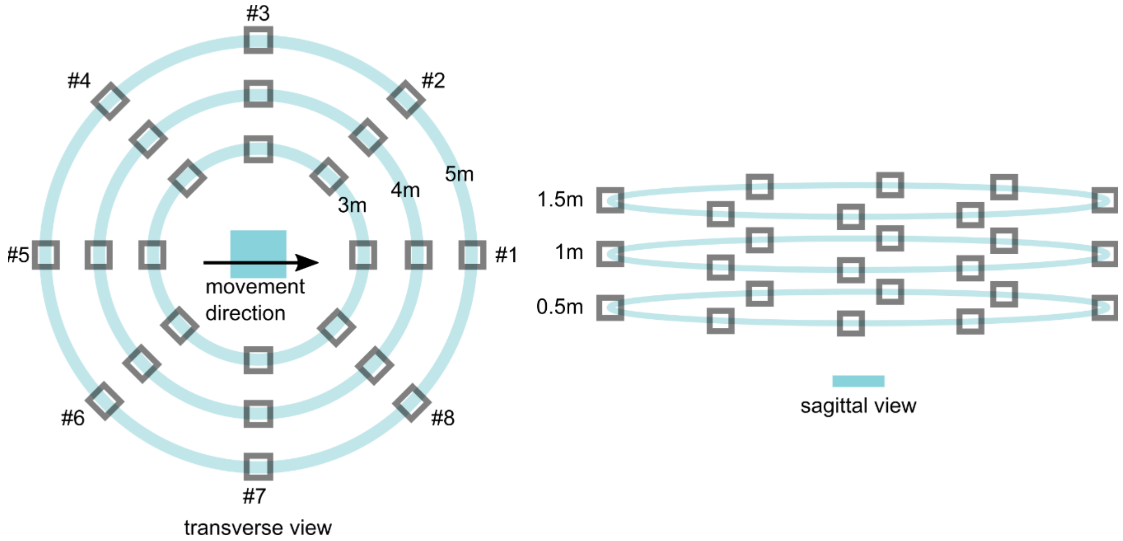

Figure 1: Overview of the simulated camera locations (views) used to estimate ground reaction force using machine learning. The simulated camera was not tilted.

Keypoints were simulated to match those output by the open source ML pose estimation model called OpenPose (Cao et al., 2019), whereby markers placed on the left and right 1st and 5th metatarsal and calcaneus were considered to be similar to the big toe, small toe and heel OpenPose keypoints. The ankle and knee joint centre were calculated as the midpoint between the ankle lateral and medial malleoli and femoral epicondyle markers for left and right respectively. The hip joint centres were calculated according to Shea and colleagues (Shea, Lenhoff, Otis, Backus, 1997). The mid-hip keypoint was defined as the midpoint between left and right hip joint centre. After calculating the coordinates of these 13 keypoints in 3D (x, y, z), they were projected to a 2D image frame (u, v). The projected keypoints should replicate keypoints estimated in images by OpenPose.

Eight different camera views surrounding the force plate were simulated, at three different distances, across three heights; resulting in a total of 8 x 3 x 3 = 72 camera views (Figure 1). Each camera view was treated as a separate input for the neural network, resulting in 72 views x 135 trials = 9,720 samples. Keypoints were translated to be reported relative to the moving coordinate system of the participant mid-hip keypoint. The stance phase was normalised to 101 time steps, resulting in 12 (keypoints) x 2 (image dimensions) x 101 (time steps) = 2424 input features. Samples for each camera view were used to train multi-layer perceptron neural networks (MLP) to estimate 3D GRFs. Sidesteps performed using the left plant foot versus the right plant foot contacting the force plate were analysed separately. In a first analysis, we aimed to identify the optimal camera position and height for estimating GRFs using 2D video input. For this purpose nine MLPs were trained; one for each position/height combination (Table 1) containing all eight camera locations. For this purpose, the dataset was split into a fixed training/validation/test set with no individual participant data split across the training, validation or test set for any MLP model training. In a second step, the camera position/height combination resulting in the highest accuracy (*r*) in GRF prediction was used and eleven different combinations of camera angles were investigated (Table 2). A leave-one-subject-out cross-validation was performed. The correlation coefficient *(r)* and normalised root mean squared error (nRMSE) were used to analyse the accuracy of the predicted GRFs. All data processing was performed using Python and Tensorflow.

RESULTS: For all tests the same neural network architecture was used: input layer: 24x101 features; hidden layer 1: 1500 neurones; hidden layer 2: 500 neurones; output layer: 3x101 features. A dropout rate of 0.7 was used to prevent the model from overfitting, the initial learning rate was set to 3x10^{-5} and training was stopped after 50 epochs. The mean absolute error was used as loss function. The validation loss did not indicate overfitting of the model.

Table 1: Results of the ground reaction force prediction for different camera positions. The best camera position is highlighted bold.

			r			nRMSE [%]	
distance [m]	height [m]	left	right	mean	left	right	mean
3	0.5	0.969	0.968	0.969	11.39	12.02	11.71
3	1	0.969	0.967	0.968	11.43	11.54	11.49
3	**1.5**	**0.973**	**0.97**	**0.972**	**10.59**	**11.39**	**10.99**
4	0.5	0.968	0.966	0.967	11.46	11.26	11.36
4	1	0.969	0.966	0.968	11.1	11.91	11.51
4	1.5	0.972	0.97	0.971	11.18	11.19	11.19
5	0.5	0.969	0.965	0.967	11.31	12.01	11.66
5	1	0.969	0.966	0.968	11.57	11.93	11.75
5	1.5	0.972	0.967	0.970	11.08	11.23	11.16

The prediction accuracy of the GRF is similar from all camera positions investigated with the highest accuracy found to be with a camera distance of 3 m and height of 1.5 m respectively as evidenced by a *r* of 0.978 for the vertical force component, 0.965 in the anterior-posterior direction and 0.972 in the medio-lateral component, with an nRMSE of 8.14%, 12.44% and 12.38% respectively.

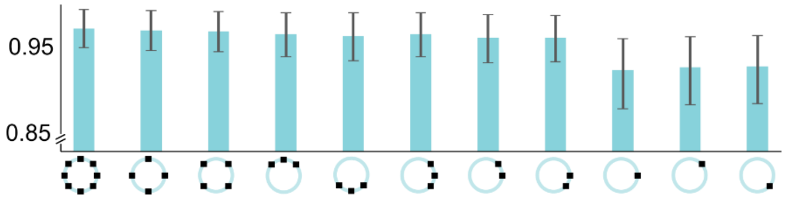

Figure 2: Correlation coefficient of the ground reaction force prediction based on camera combinations.

Camera combinations ranging from a single to eight cameras resulted in minor differences in GRF prediction accuracy. The highest accuracy was achieved with an eight-camera set-up (*r*=0.973), the lowest accuracy for a single camera placed in front (*r*=0.928) (Figure 2).

DISCUSSION: This study used synthesised 2D pose estimation keypoints to predict sidestepping 3D GRFs and investigated the influence of different camera positions and combinations on the resulting accuracy of the prediction. Neither the camera position or number of cameras significantly influenced the high prediction accuracy achieved across all permutations (%change 1 to 8 cameras 31.99%, 2 to 8 cameras 12.44, 3 to 8 cameras 10.26%). Eight cameras resulted in the highest prediction accuracy. This might be due to more information available or based on the larger number of training samples. The keypoints used to train the machine learning model were referenced to the mid-hip coordinate system, resulting in relatively homogeneous inputs across the multiple camera positions, which may explain the similar results across all comparisons.

In one of our previous studies, we used 2D videos from cameras positioned similar to cameras #6, #7 and #8 during the same data capture. We achieved nRMSE values of 44.21% for the medio-lateral, 13.11% for the anterior-posterior and 13.07% for the vertical component using OpenPose to estimate 25 keypoints. The keypoints were not referenced to the mid-hip keypoint and no independent left and right ML models were trained (Morris et al., 2021). In this study we achieved nRMSE of 17.00%, 11.07% and 10.40% for the same camera combination. We further investigated a more computationally expensive method to animate videos and employed OpenPose to estimate keypoints on the same dataset (Mundt et al., 2021, Mundt, Oberlack, Morris, Goldacre, Funken, Potthast, Alderson & Powles, 2022). In this study we used eight simulated camera views and eight keypoints referenced to the mid-hip keypoint. The keypoints used in this study included the trunk but no feet and data was not separated into right and left sidesteps. The GRF components were estimated with an nRMSE of 17.62%, 11.11% and 11.94 %. For the same given camera combination we would achieve nRMSE values of 14.67%, 10.78% and 9.70% in this study while using a computationally more efficient approach and treating left and right steps independently. Using keypoints relative to a moving coordinate system creates more homogeneous inputs and less dependency on camera positions and facilitates improved results.

CONCLUSION: This study presents a computationally efficient method to repurpose 3D motion capture data to synthesise 2D pose estimation keypoints using computer vision methods. A single camera can be used to accurately estimate 3D GRF from 2D video data, thereby creating a mechanism to artificially increase sparse video datasets. Separating left and right movement direction results in higher prediction accuracy than combining both into one – which may be counterintuitive given the split results in a smaller dataset. These results suggest that sport practitioners may be able to develop tools to use a single smart phone camera to capture an athlete during a game or training session in order to estimate GRF and, by extension, gain insights into mechanical load exposure.

Future research should further validate the findings regarding camera location and optimal view combination. Finally, the use of more than eight camera views should be investigated as this study suggests that more cameras may result in improved prediction accuracy.

REFERENCES

Besier, T.F., Sturnieks, D.L., Alderson, J. & Lloyd, D.G. (2003). Repeatability of gait data using a functional hip joint centre and a mean helical knee axis. Journal of Biomechanics, 36, 1159–1168.

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2019). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cronin, N. J. (2021). Using deep neural networks for kinematic analysis: Challenges and opportunities. *Journal of Biomechanics*, 123, 110460.

Morris, C., Mundt, M., Goldacre, M., Weber, J., Mian, A. & Alderson, J. (2021). Predicting 3D ground reaction force from 2D video via neural networks in sidestepping tasks. In *39th International Society of Biomechanics in Sport Conference*. Virtual Conference. September 3-7, 2021.

Mundt, M., Oberlack, H., Morris, C., Funken, J., Potthast, W. & Alderson, J. (2021). No dataset too small! Animating 3D motion data to enlarge 2D video databases. In *39th International Society of Biomechanics in Sport Conference*. Virtual Conference. September 3-7, 2021.

Mundt, M., Oberlack, H., Morris, C., Goldacre, M., Funken, J., Potthast, W., Alderson, J. & Powles, J. (2022). Repurposing biomechanics data: using Machine learning to synthesize 2D camera views, train 2D pose estimation models and predict 3D ground reaction forces from historical 3D motion capture data. Submitted to *Sensors*.

Nolan, L., Patritti, B. L., & Simpson, K. J. (2012). Effect of take-off from prosthetic versus intact limb on transtibial amputee long jump technique. *Prosthetics and Orthotics International*, 36(3), 297–305.

Shea, K.M., Lenhoff, M.W., Otis, J.C., Backus, S.I., 1997. Validation of a method for location of the hip joint center. *Gait & Posture* 5, 157–158.

Winter, D. (1991). *The biomechanics and motor control of human gait : normal, elderly, and pathological*, second ed.; University of Waterloo Press.