A Non-invasive Vision Based Approach to Velocity Measurement of Skeleton Training

Murray Evans
Centre for the Analysis of Motion Entertainment Research and Applications (CAMERA)
Department of Computer Science
University of Bath
m.evans@bath.ac.uk

Laurie Needham
CAMERA
Department of Health
University of Bath
ln424@bath.ac.uk

Steffi L. Colyer
CAMERA
Department of Health
University of Bath
sc356@bath.ac.uk

Darren P. Cosker
CAMERA
Department of Computer Science
University of Bath
d.p.cosker@bath.ac.uk

Abstract

Skeleton is a winter sport where performance is greatly affected by the velocity an athlete can achieve during their start up to the point where they load themselves onto their sled. As such, it is of interest to athletes and coaching staff to be able to monitor the performance of their athletes and how they respond to different training schedules and techniques. This paper proposes a non-invasive vision based method for measuring the velocity of a skeleton athlete and their sled during the push start. Mean differences in estimated velocity between ground truth data and our proposed system were -0.005 ($\pm$ 0.186) $m.s^{-1}$ for the athlete mass centre and -0.017 ($\pm$ 0.133) $m.s^{-1}$ for the sled. The results compare favourably to techniques previously presented in the biomechanics and sport science literature.

1. Introduction

This paper proposes an approach to measure the velocity of skeleton athletes and their sled as they sprint along a push-track training facility. The velocity of the sled has been shown to be a key determinant of overall skeleton performance [4] and so a thorough understanding of sled and athlete speed during the push start is valuable to the athletes and coaching staff for the purposes of performance enhancement.

During running, various parts of the athlete's body move at different rates, however, the overall velocity of the athlete is described by the motion of their centre of mass. This can be accurately estimated using standard optical motion capture systems, by placing many reflective markers on the athlete and sled. These markers are time consuming to place however, and can interfere with the athlete's natural performance. Markers on the sled are less intrusive, but are not permitted to be used in the competition environment and thus a system that can non-invasively measure both the athlete and sled velocity would provide impact to skeleton coaching.

Non-invasive field based measures of athlete velocities include the use of floor mounted laser grids such as the OptoJump™system (Microgate, Bolzano, Italy). Such systems can provide estimates of the average velocity of the athlete's mass centre across each step [9] but would fail in a skeleton environment due the false detections caused by the sled. Laser distance measurement can also provide non-invasive estimates of athlete's mass centre velocity[1] however such methods provide noisy data which requires extensive signal processing. Furthermore, such an approach would also fail in the skeleton environment as the laser would be occluded as the athlete travels over the brow of the track during a push start. Alternately, a vision based approach may provide a non-invasive approach to capturing important skeleton performance characteristics.

The detection of humans has advanced rapidly with the progress of convolutional neural network (CNN) systems. These have been used for basic bounding-box level detections [15], detecting a sparse set of joints on a person (e.g. shoulders, elbows, wrists, hips, knees, ankles, etc) [2, 6], for predicting the pose of a 3D model from single views [13, 18] or for providing per-pixel segmentation results of the whole body [21] and individual body

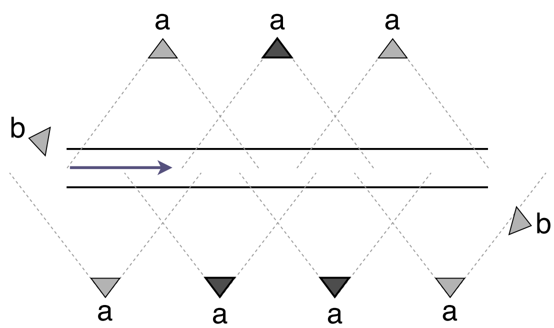

Figure 1. Camera system. Athletes push the sled along the track between a corridor of cameras. Cameras a look across the track, perpendicular to sled motion, while cameras b look along the track, as parallel to sled motion as possible without obstructing training activities.

parts [7, 16]. The aim of this paper is to determine if any of these technologies can enable useful markerless measurement of skeleton athletes and their sled.

First, the camera system used for the experiments will be described in Section 2. Section 3 will propose a method for measuring the speed of the athlete, then section 4 will propose a method for tracking the sled. The performance of these two systems will then be evaluated in Section 5 with final thoughts provided in the conclusion.

2. Camera system

The hardware used for this paper consisted of a 9 camera setup with cameras along both sides of a push-track training facility. Such facilities allow for off season training of skeleton and bob-sleigh and consist of a concrete declined hill with straight metal rails on which a wheeled practice sled can be used. The cameras used were JAI machine vision cameras set to record HD images (1920×1080) at 200 Hz, with a 200 Hz trigger signal used to ensure frame synchronisation. The cameras were arranged along the track as shown in Figure 1, and cover an approximately 8 metre span at the start of the track where it is sloped but has constant gradient of 2%.

The camera system was calibrated using standard techniques. A circle-grid calibration board is presented to each camera in turn for intrinsic calibration [22], and then moved through the scene, ensuring it is seen by multiple cameras at any one time. Camera extrinsic parameters are calculated from these shared observations and then optimised using Bundle Adjustment [19] to reach a globally optimal calibration. Marks on the ground are used to align the calibration such that $z = 0$ is the floor plane, with $+z$ up. The y-axis of the system is aligned to be parallel with the sled track, such that $+y$ is down the hill parallel to the sled rails. Fixing the alignment of the calibration in this way means that the sled orientation is both known and easy to work with.

In parallel with the machine vision camera system, ground truth data were captured using a 15 camera marker-based motion capture system (Oqus, Qualysis AB, Gothenburg, Sweden). 12 international skeleton athletes performed up to 4 maximal pushes on a dry land skeleton training track. A full body marker set comprising of 44 individual markers and four clusters were attached to each participant to create a full body six degrees of freedom (6DoF) model (bilateral feet, shanks and thighs, pelvis and thorax,upper and lower arms, and hands). Four additional markers were placed on the sled to track position and orientation. Following labelling and gap filling of trajectories (Qualysis Track Manager v2019.3, Qualysis, Gothenburg, Sweden) data were exported to Visual 3D (v6, C-Motion Inc, Germantown, USA) where raw trajectories were low-pass filtered (Butterworth 4th order, cut-off 12 Hz) and a 6DoF inverse kinematics (IK) constrained model was computed. Athlete mass centres were computed using the model described by de Leva [5]. Additionally, the sled was modelled as a rigid object with uniformly distributed mass. Filtered marker data and mass centre locations were used to compute mass centre derivatives using a finite central differences method and touch-down (TD) and toe-off (TO) events were computed based on foot marker kinematics [8]. Computing TD and TO events permitted the calculation of the average velocity of the sled and athlete mass centre across each step.

3. Estimating athlete velocity

To measure the velocity of the athlete as they proceed along the track, this paper proposes an algorithm based on multiple processing stages.

1. Label athlete body parts in each camera view using a CNN based segmentation algorithm.

2. Get bounding boxes of head and torso regions in each view.

3. Use back-projection to fuse the head and torso observations.

4. Optimise a 3D bounding box to fit the head and torso observations.

5. Track the bounding box over time using a Kalman filter.

A number of different approaches could be taken for detecting the athlete in the images. One of the most widely seen approaches to human detection and pose estimation uses CNNs to detect a sparse collection of points corresponding to the approximate image location of body joints (e.g. shoulders, elbows, wrists, hips, knees, ankles). Initial testing with OpenPose [2] suggested that the highly occluded, bent over position of the athlete caused quite noisy and unreliable detections.

Systems which estimate a full 3D model from uncalibrated single views such as [13] could provide an alternative, but it is unclear exactly how to fuse the results of individual cameras as the athlete traverses the scene. Ideally each individual camera view would produce the same result, however in practice the 2D ambiguity is not fully resolved by these systems and the models will be different. There is also a question over accuracy when these systems typically forgo a calibrated camera in favour of a generic projection matrix during training, and additionally, questions of how scale can be fully resolved.

The velocity of the athlete is taken to be the rate of change of their centre of mass displacement. Previous works have shown that centre of mass can be estimated using volumetric reconstructions such as a voxel hull [12]. These can generally be constructed from segmentation masks of the athlete from multiple views, but require that the cameras are arranged to maximise the performance of space carving - ensuring that individual limbs can be separated from each other and the rest of the body, and that no negative space is filled in. The pose of the skeleton athlete while running with their sled makes it very difficult to accommodate this need, especially if cameras are required to also span a significant length of track and the recording system is to remain economical.

Rather than attempt to recover a high fidelity volumetric reconstruction of the athlete in the scene, this work demonstrates that the motion of the athlete's mass centre along the track can be recovered by tracking a 3D bounding box that is optimised to best fit the observations from each individual view.

For the first stage, human segmentation is performed using the approach of [16]. This provides a segmentation that labels pixels of the image as belonging to various parts of the body - lower and upper leg, torso, lower and upper arm, and head. Despite the pose of the athlete, and their wearing of a helmet, the algorithm produced useable segmentations. Some common failures were observed, such as partial or fragmented segmentations, particularly of the head, and some mislabelling of body parts; the most common miss-labelling being to label the arm that pushes the sled as a leg.

Although it can be guaranteed that only one athlete will be performing during any one video, it cannot be guaranteed that there will not be other people in the background of images or that the athlete's body parts will be segmented into single perfect blobs. As such, a robust method of fusing the segmentations between views is needed to get a 3D estimate of the athlete's body parts.

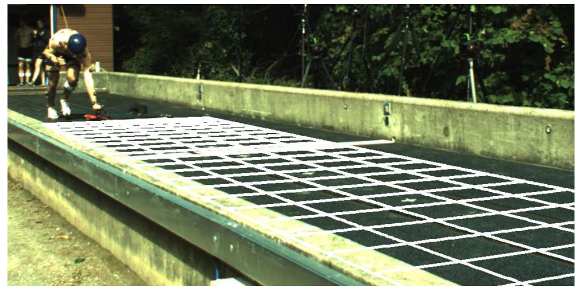

Figure 2. Example occupancy grid overlaid on the push-track (actual grid is higher resolution).

To this end, a variant of occupancy maps (also termed synergy maps [14]) is proposed. The camera system is calibrated and the extents of the observed region of the push-track are known, as well as the plane of the ground - which is assumed to be the $z = 0$ plane. An occupancy map is typically created by dividing the region of the ground plane up into a grid of cells as seen in Figure 2, and projecting each cell into the images. Usually, the occupancy of a cell is computed as the sum of images for which the projection is inside the segmentation, although variants exist which allow for the occupancy to be based on multiple scene planes above the ground [20]. Due to occlusions, miss-labelling, broken up segmentation regions and indeterminate height of the athlete above the ground, this work changes the computation of occupancy slightly.

First, connected components is used to place a bounding box around each body part segmentation region. Specifically, only the head and torso regions are currently used to minimise the impact of moving legs and arms on estimates of the centre of mass. An occupancy map for each body part is computed by projecting each cell of the occupancy map into each camera view to provide a resulting point p_i in the image. If p_i lies between the horizontal extents of one or more bounding boxes in the view, the occupancy of the cell is incremented by 1. An example can be seen in Figure 3.

As there will only ever be one athlete on the push-track at one time, localising the head or torso body parts can be done by searching for the peak value in the occupancy map. Once the peak has been found, a tracking bounding box can be initialised at the location corresponding to that occupancy cell. If the location of the body part is known from previous frames then the occupancy map can serve for verifying that the part is still present.

3.1. Modelling and tracking parts using bounding boxes

Body parts are tracked by modelling them as 3D axis aligned bounding boxes and fitting these bounding boxes to the segmentation. On first detection, the occupancy map provides an approximate x, y location for the body part, but does not provide a vertical z coordinate. The z component

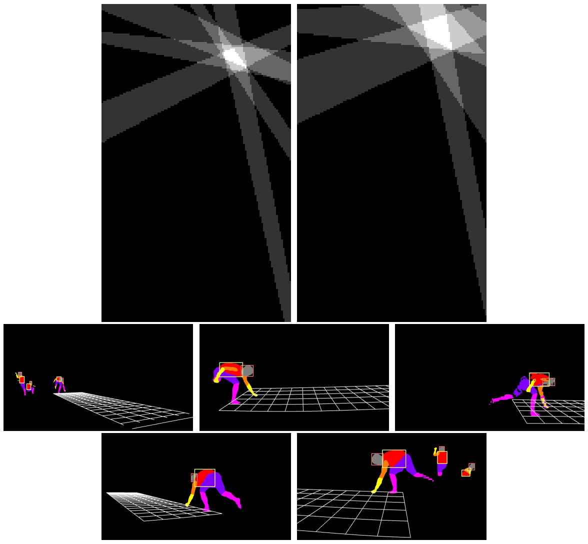

Figure 3. The top two images show the occupancy for the head (left) and torso (right). Beneath this are the segmentation images that generated the occupancy. Each segmentation image is overlaid with a 2D bounding box around the torso and head segments against which grid cells are tested, and a low-resolution visualisation of the grid.

can be determined as part of the process for optimising the position of the bounding box.

The 3D part box is described by its centre (x, y, z) and its size (w, l, h). If the part box is known from previous frames, then that previous box is used to initialise the search for the new state in the current frame. If it is not known, then x and y position are initialised to the peak of the occupancy map. Width w and length l can be set to reasonable guesses and held fixed, which for the head are both 300 mm and for the torso are both 400 mm. z can be initialised to 800 mm, and the height of the box initialised to a large size of 700 mm and solved for as part of the search for the optimal bounding box position.

Tracking the part from frame to frame consists of optimising the parameters of the 3D part box so that it best fits the segmentations in the image.

First, the initial 3D part box B_3 is projected into each camera view. This means that the 8 corners of B_3 are projected into image i, and then the smallest 2D bounding box B_{2i} that can enclose the 8 resulting projections is determined.

For each image there are a set of 2D bounding boxes resulting from connected components of the segmentation map. These segment boxes are checked for intersection with B_{2i}. The subset that are shown to intersect with B_{2i} are merged together into a bounding box S_{2i}. This process is designed to handle problems induced by low quality segmentations where a part might be detected as disconnected partial segmentations.

The 3D part box B_3 is optimised to minimise the error in Equation 1:

$$e = \sum_i 10 \cdot O(B_{2i}, S_{2i}) + O(S_{2i}, B_{2i}) + C(B_{2i}, S_{2i}) \quad (1)$$

In Equation 1 the function $O(a, b)$ computes a bounding box overlap error and $C(a, b)$ computes an error based on the distance between bounding box centres:

$$O(a, b) = 1.0 - (A(I(a, b))/A(b)) \quad (2)$$

$$C(a, b) = 1.0 - e^{\left(\frac{-(c(a)-c(b))^2}{1000000}\right)} \quad (3)$$

where $A(x)$ computes the area of the shape x, and $I(a, b)$ computes the intersection of a and b, while $c(x)$ computes the centre of the shape x. The functions $O(a, b)$ and $C(a, b)$ are written so that good results return errors near 0, hence the $1.0 - x$ form of each equation. For $C(a, b)$, the x of the $1.0 - x$ form is a Gaussian shaped function, and the large constant in the denominator is chosen to provide a broad basin to the error term, selected based on the size of objects in the image and the image resolution. The constant of 10.0 on the first term of Equation 1 was used because it was considered to be especially important that B_{2i} not be smaller than S_{2i}.

Examples of the resulting bounding boxes can be seen in Figure 4. The optimisation to minimise the error can be carried out using any suitable algoirthm - this paper uses the SBPLX algorithm from the open source C++ NLOpt library [11]. Once the bounding box has been optimised, the final parameters are used to update a Kalman filter to help keep the track smooth.

4. Estimating sled velocity

To detect the sled, DeepLabCut (Version 2.1.6) [17] was used to train a CNN to detect the four corners of the sled across all camera views. DeepLabCut is a toolbox to facilitate transfer learning, taking a pre-trained feature detector and specialising it for a different, but related, task. Approximately 400 images of the sled were labelled, with 95% being used for training. A ResNet-50 encoder and DeeperCut [10] decoder architecture were combined and trained for 200,000 iterations. Test error was: 4.16 pixels, train error: 2.13 pixels, and a confidence-cutoff of 0.4 was implemented to condition the (x, y) image coordinates for further analysis. Figure 5 shows several examples of the detected sled corners.

Once detected, corners are back-projected from each image down to the $z = 100$ plane, which is approximately the height of the sled on the track. This would ideally leave four

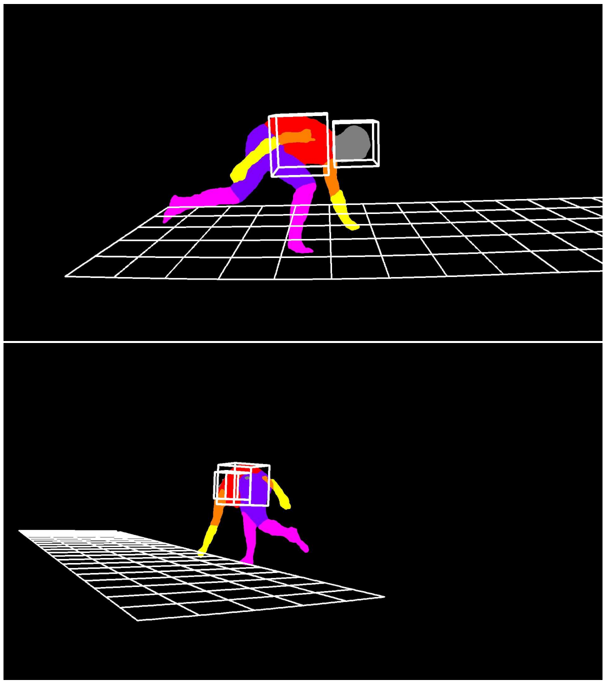

Figure 4. Final 3D bounding boxes resulting from fitting to segmentation information.

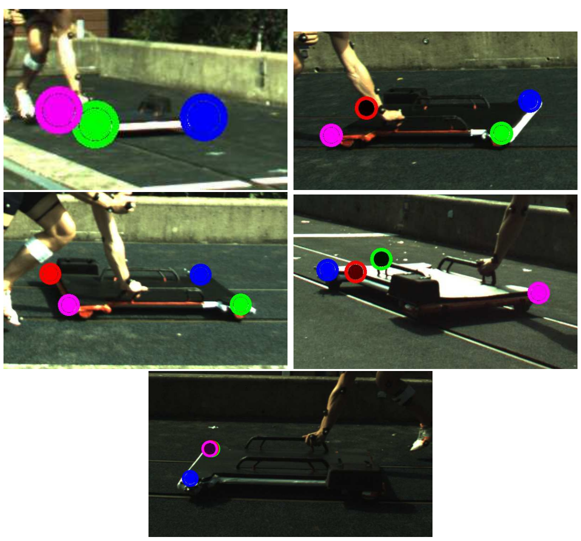

Figure 5. Examples of sled corner detections. Circles are coloured based on the corner label, with confidence indicated by the brightness of the circle interior. Many of the detections are of the quality of the top right and middle left examples, but a number of different failure cases have been observed as evidenced here.

tight clusters of points, one for each corner, though in practice there can be some outliers. Another problem can be that the CNN can confuse which corner is which when labelling. However, knowing the layout of the track and the shape of the sled means that labelling can be ignored. An example of the back-projected points can be seen in Figure 6.

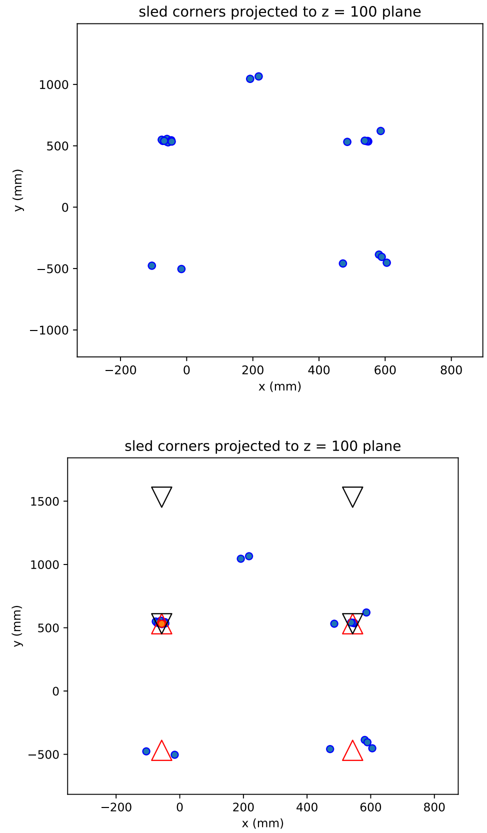

Figure 6. The top figure shows the detected sled corners back-projected to the $z = 100$ plane. The four clusters for the correct point corners can be seen, as well as a fifth cluster for some erroneous points. The orange highlighted (orange) point near (-50,500) can be assumed to be either a back-left corner, or a front left corner of the sled. If back left, then the hypothesised sled corners must be located where the black, down-poiting triangle markers are shown. If back right, then the red, up-pointing triangle markers show where the hypothesised sled corners would be. As the up-pointing triangle markers are more consistent with the clusters of back-projected corner detections, it can be inferred that the highlighted point must be a front-left corner of he sled.

Each back-projected corner point is taken in turn and assumed to be either a front corner or a back corner. Left vs. right can be inferred from the x coordinate of the point and the known geometry of the track. Next, a check is conducted to determine how many of the other corner points would be consistent with this assumption. The configuration that produces the most consistent result is used to initialise the position of the sled and label the corner points. The consistency check is as simple as counting how many points are within 100 mm of each hypothesised corner. An example of this can be seen in Figure 6.

Once the hypothesis most consistent with the point data is chosen, points that are consistent with the hypothesis are

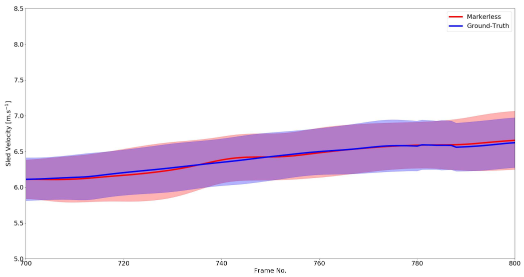

Figure 7. Mean markerless sled velocity (± SD) (blue line) and mean marker-based (ground truth) sled velocity (± SD) (red line).

associated to their nearest sled corner. The (x, y, z) position of the centre of the sled is then optimised by minimising an objective function that sums the distance of each detected corner from the projection of the associated sled corner in image space. As with the athlete body part box fitting, this minimisation is carried out using the NLOpt SBPLX algorithm [11].

5. Evaluation

Accuracy of the proposed system was assessed against marker-based motion capture which provided ground truth data as described in section 2. In total, 33 push trials from 12 athletes were used for system evaluation. This data was independent of that used for training/testing but was collected in the same environment (skeleton training push track). In order to evaluate system performance, results were compared using linear regression and Bland-Altman analysis.

Figure 7 demonstrates the mean (± SD) markerless and mean ground truth sled velocity as a function of time for all push trials. Mean differences between systems across the step were -0.005 (± 0.186) m.s^{-1} for the athlete mass centre and -0.017 (± 0.133) m.s^{-1} for the sled (Table 1).

Bland-Altman analysis of the athlete mass centre velocity and sled velocity are given in Figure 8 and Figure 9 respectively. Very good agreement is reported between both systems (proposed vs. ground truth) with a very low bias for both the athlete and sled step velocities. Furthermore the standard deviations fall well within the limits of agreement, further supporting the validity of the proposed method.

When compared to a commonly used field based approach to measuring athlete running velocities - laser distance measurement, which exhibits mean errors of up to 0.41 (± 0.18) m.s^{-1} [1], the proposed method provides substantial improvements. Previously, combined sled and athlete velocities have been measured using photocells placed at five metre intervals along the push track or using a magnetic encoder placed on the sled's wheel [3]. However, discrete photocells provide only discrete data across

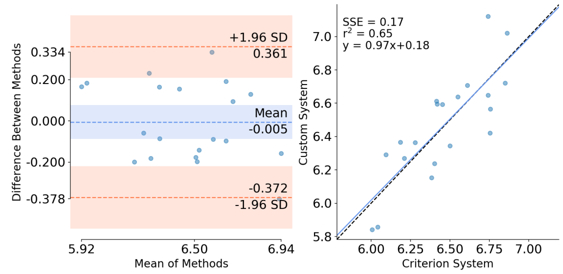

Figure 8. Bland-Altman and linear regression plots comparing step averaged athlete CoM velocity between systems. Confidence intervals are given around the mean difference and 95% limits of agreement.

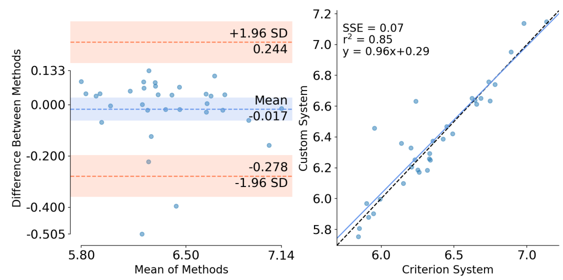

Figure 9. Bland-Altman and linear regression plots comparing step averaged sled velocity between systems. Confidence intervals are given around the mean difference and 95% limits of agreement.

each five metre section of the track and inconsistencies in timings can be caused when different body parts trigger the photocell. The sled's magnetic encoder can provide data points at a higher resolution, but this is still limited to one sample every rotation of the wheel or 0.1984 m (the sled wheel's circumference). Additionally, the wheel is susceptible to slipping which creates spurious readings. The proposed method in this paper provides higher resolution information than the current approaches being utilised. Furthermore, it can also determine the velocities of the athlete and sled separately without the need for two measurement systems and as such provides further novel insights into push start performance.

6. Conclusion

A vision based approach to non-invasively collect velocity data was proposed and validated for skeleton push start performance. The proposed method was applied in a challenging real world environment and application (skeleton push starts) and was able to capture representative velocity data of the athlete mass centre and sled. Such an approach provides a viable and accurate alternative to a) marker-based motion capture which is invasive due to the need for athlete's to wear markers and b) non-invasive field

Variable	Mean Difference (Bias)	$\pm$ SD	Bias $\pm$ 1.96 SD	R^2
Athlete CoM Velocity ($m.s^{-1}$)	-0.005	0.186	0.244	0.65
Sled Velocity ($m.s^{-1}$)	-0.017	0.133	0.361	0.85

Table 1. Comparison of computed step velocities.

based measures that would ultimately be difficult to deploy into such a challenging environment. Furthermore the proposed method could be utilised by coaches and sports science support staff to monitor technique where traditional motion capture techniques may not.

Acknowledgement

This investigation was part-funded by CAMERA, the RCUK Centre for the Analysis of Motion, Entertainment Research and Applications, EP/M023281/1. Thank you to British Skeleton for their time and support with this project.

References

[1] N. E. Bezodis, A. I. Salo, and G. Trewartha. Measurement error in estimates of sprint velocity from a laser displacement measurement device. *International Journal of Sports Medicine*, 33(6):439–444, 2012. 1, 6

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018. 1, 3

[3] S. L. Colyer, K. A. Stokes, J. L. Bilzon, , and A. I. Salo. Skeleton sled velocity profiles: a novel approach to understand critical aspects of the elite athletes' start phases. *Sports biomechanics*, 17(2):168–179, 2018. 6

[4] S. L. Colyer, K. A. Stokes, J. L. Bilzon, D. Holdcroft, and A. I. Salo. The effect of altering loading distance on skeleton start performance: Is higher pre-load velocity always beneficial? *Journal of Sports Science*, 36(17):1930–1936, 2018. 1

[5] Paolo De Leva. Adjustments to zatsiorsky-seluyanov's segment inertia parameters. *Journal of Biomechanics*, 29(9):1223–1230, 1996. 2

[6] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 1

[7] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. *arXiv*, 2018. 2

[8] J. C. Handsaker, S. E. Forrester, J. P. Folland, M. I. Black, and S. J. Allen. A kinematic algorithm to identify gait events during running at different speeds and with different footstrike types. *Journal of Biomechanics*, 49(16):4128–4133, 2016. 2

[9] R. Healy, I. C. Kenny, and A. J. Harrison. Estimating step parameters using photoelectric cells. In *Proceedings of the 33rd International Conference of Biomechanics in Sports*, 2015. 1

[10] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, , and B Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *In European Conference on Computer Vision*, pages 34–50, 2016. 4

[11] Steven G. Johnson. The nlopt nonlinear-optimization package. http://ab-initio.mit.edu/nlopt, 2018. 4, 6

[12] T. Kaichi, S. Mori, H. Saito, K. Takahashi, D. Mikami, M. Isogawa, and H. Kimata. Estimation of center of mass for sports scene using weighted visual hull. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1890–18906, June 2018. 3

[13] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 1, 3

[14] S.M Khan and M. Shah. Tracking multiple occluding people by localizing on multiple scene planes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3):505–519, March 2009. 3

[15] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan. Scale-aware fast r-cnn for pedestrian detection. *IEEE Transactions on Multimedia*, 20(4):985–996, April 2018. 1

[16] Kevin Lin, Lijuan Wang, Kun Luo, Yinpeng Chen, Zicheng Liu, and Ming-Ting Sun. Cross-domain complementary learning with synthetic data for multi-person part segmentation. *arXiv preprint arXiv:1907.05193*, 2019. 2, 3

[17] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21:1281–1289, 2018. 4

[18] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1

[19] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment — a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*, pages 298–372, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg. 2

[20] Alparslan Yildiz and Yusuf Sinan Akgul. A fast method for tracking people with multiple cameras. In Kiriakos N. Kutulakos, editor, *Trends and Topics in Computer Vision: ECCV 2010 Workshops, Heraklion, Crete, Greece, September 10-11, 2010, Revised Selected Papers, Part I*, pages 128–138, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. 3

[21] S. Zhang, R. Li, X. Dong, P. Rosin, Z. Cai, X. Han, D. Yang, H. Huang, and S. Hu. Pose2seg: Detection free human instance segmentation. In *2019 IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 889–898, June 2019. 1

[22] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, Nov 2000. 2