A Single-Camera Computer Vision-Based Method for 3D L5/S1 Moment Estimation During Lifting Tasks

Hanwen Wang, Ziyang Xie, Lu Lu, Li Li, Xu Xu
Edward P. Fitts Department of Industrial & Systems Engineering,
North Carolina State University, Raleigh, NC, USA

Abstract: Excessive low back joint loading during material handling tasks is considered a critical risk factor of musculoskeletal disorders (MSD). Therefore, it is necessary to understand the low-back joint loading during manual material handling to prevent low-back injuries. Recently, computer vision-based pose reconstruction methods have shown the potential in human kinematics and kinetics analysis. This study performed L5/S1 joint moment estimation by combining VideoPose3D, an open-source pose reconstruction library, and a biomechanical model. Twelve participants lifting a 10 kg plastic crate from the floor to a knuckle-height shelf were captured by a camera and a laboratory-based motion tracking system. The L5/S1 joint moments obtained from the camera video were compared with those obtained from the motion tracking system. The comparison results indicate that estimated total peak L5/S1 moments during lifting tasks were positively correlated to the reference L5/S1 joint moment, and the percentage error is 7.7%.

INTRODUCTION

Manual materials handling (MMH) is considered as one of the risk factors of low-back pain (Da Costa & Vieira, 2010; Hoogendoorn et al., 2000; Kuiper et al., 1999; Schaffer, H., & United States, 1982). From the ergonomics perspective, it is critical to ensure the low-back joint loadings during working are within the failure tolerance to avoid low-back injuries (McGill, 1997). To date, numerous studies have investigated the L5/S1 joint moment to identify the risks associated with a variety of lifting tasks. These studies have focused on the peak value of the L5/S1 joint moment (Jiang et al., 2005), as well as the cumulative L5/S1 joint moment (Callaghan et al., 2001)

In order to calculate low-back joint moment, workers' body motion needs to be captured first so that human kinetics methods can be further applied. One method to capture workers' body motion is to use an optical marker-based motion tracking system. Such a system is able to obtain three-dimensional coordinates of markers that are attached to workers' body in a laboratory environment. The dynamic moments at L5/S1 joint are then calculated using worker's body motion together with body segment inertial properties (Pfister et al., 2014). However, this method is less practical for field studies due to the bulky size, high cost and expertise that are associated with a laboratory-based motion tracking system.

To overcome this limitation, a few studies sought to develop video-based coding systems that use human raters to observe workers' posture from the videos recorded in field studies. The raters estimate body pose in selected key frames extracted from the recorded videos by fitting the poses to a predefined digital manikin (Xu et al., 2012). The workers' body motion is then reconstructed by interpolating the rater-identified poses in the keyframes. While this method does not rely on a laboratory-based motion tracking system for capturing workers' body motion, it is highly labor-intensive as the raters would need to observe a large number of video frames. In addition, the accuracy of the reconstructed body motion heavily relies on the experience of the raters as well as the view angle of the videos.

With the recent development of the advanced deep neural network, various computer vision algorithms have been presented to estimate 3D human poses through videos (Mehrizi et al., 2019). For example, previous studies attempted to reconstruct 3D pose from multiple two-dimensional (2D) calibrated images derived from Openpose (Simon et al., 2017; Cao et al.,2021). D'Antonio etc. (D'Antonio et al., 2020)used two synchronized videos during walking to compute lower limb joint kinematics by applying a triangulation algorithm on the 2D joint center coordinates assessed in Openpose. While these methods can yield 3D poses from synchronized videos captured from multiple view angles, the reconstruction accuracy inevitably depends on the number of the cameras. In addition, camera calibration among multiple cameras is time-consuming and requires expertise in computer vision, which could be a technical burden for ergonomics practitioners.

Very recently, Pavllo et al. (Pavllo et al., 2019)developed a single-camera-based 3D pose reconstruction algorithm named VideoPose3D. This algorithm can estimate 3D poses using a fully convolutional model generated by dilated temporal convolutions over 2D joint points. A semi-supervised approach was introduced in their work and could process unlabeled video without any 2D ground truth annotations. Because this algorithm only relies on the video captured from a single camera, it has a good potential for ergonomists to investigate workers' body posture and the associated joint loadings in the field.

In this study, we developed a computer vision-based method for analyzing low-back joint moments during lifting tasks. Particularly, Detectron2 (Pavllo et al., 2019)is adopted for 2D key-point detection, and VideoPose3D (Pavllo et al., 2019) is applied to process the unlabeled video data and reconstruct workers' 3D pose. A top-down inverse dynamic biomechanical model was then adopted to calculate the

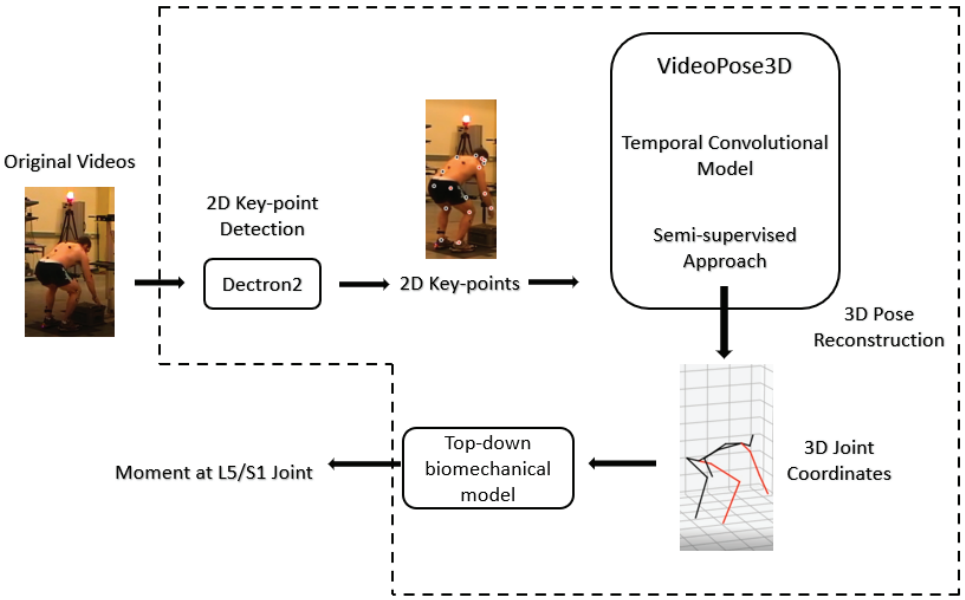

Figure 1. Workflow of the single camera-based computer vision method for estimating L5/S1 joint moment.

moments at L5/S1 joint. To test the validity of this proposed method, we conducted an experiment where participants perform a variety of lifting tasks and their motions were concurrently captured by a camera and a laboratory-based motion tracking system. The L5/S1 joint moments derived from the proposed method were then compared with those derived from a laboratory-based motion tracking system.

METHODS

Experiment design

Twelve male participants (age 47.50 ± 11.30 years; height 1.74 ± 0.07 m; weight 84.50 ± 12.70 kg) lifted a plastic crate of 10 kilograms from floor to a knuckle-height shelf. Each lifting task was performed twice. The lifting trials were captured at 30 frame per second by a camcorder (GR-850U, JVC) with a resolution of 720 × 480 pixels. The camera was placed on the rear-right side (135 degrees from the sagittal plane). Participants' body motions were also recorded by a motion tracking system (Motion Analysis, Santa Rosa, CA) through 45 reflective markers attached to the bony landmarks of the participants at 100 Hz.

Computer vision-based method

The workflow of the proposed video-based L5/S1 joint moment estimation method includes three steps: 2D key-point detection and 3D reconstruction and moment calculation, which is summarized in Figure 1. The input is the videos of each participant, and the output is the L5/S1 joint moment.

2D key-point detection and 3D reconstruction. The recorded videos are first processed in Detectron2 to estimate 2D key points in each frame. Since the script assumes exactly one person is depicted, it will select the person corresponding to the bounding box with the highest confidence. In the case of multiple people visible at once in the video, the background is blurred in advance. The 2D key points from each video are converted to a dataset in the form of "NumPy archives" for inputting into VideoPose3D. VideoPose3D is a fully convolutional architecture that uses 2D key-point sequences as input and processes them through temporal convolutions (see Figure 2). In the input layer, the estimated 2D (x, y) coordinates of the J joints in each frame are applied in a temporal convolution with C output channels and W kernel size. B ResNet-style blocks surrounded by a skip-connection (He et al., 2016) first perform a 1D convolution with kernel size W and dilation factor $D = W^B$, followed by convolution with kernel size = 1. In this study, J = 17 joints, C = 1024 output channels, W = 3, B = 4 blocks. Each convolution process is followed by batch normalization (Ioffe & Szegedy, 2015), rectified linear units (Nair & Hinton, 2010), and a dropout layer (Srivastava et al., 2014) except the last layer. The receptive field of each block increases exponentially by a factor of W, while the quantity of parameters increases linearly. Thus, the receptive field for any output frame will include information extracted from all input frames (see Figure 2). Finally, the last layer predicts the 3D poses for video frames using 2D key-point data generated in Detectron2. Since we do not include any ground truth pose data or the camera extrinsic parameters for the recorded videos, this method does not train a traditional supervised loss where the ground truth 3D poses data is set as a target. The semi-supervised training method introduced in Pavllo et al. (Pavllo et al., 2019)is applied in this study. A projection layer is added after 3D pose estimation, and the 3D predicted poses are regressed and projected back to 2D coordinates. The projected results are then compared with the input to check for consistency. A penalty is applied if the 2D coordinates from the projection process are far from the 2D data input. As the global position of key joints can be arbitrary for

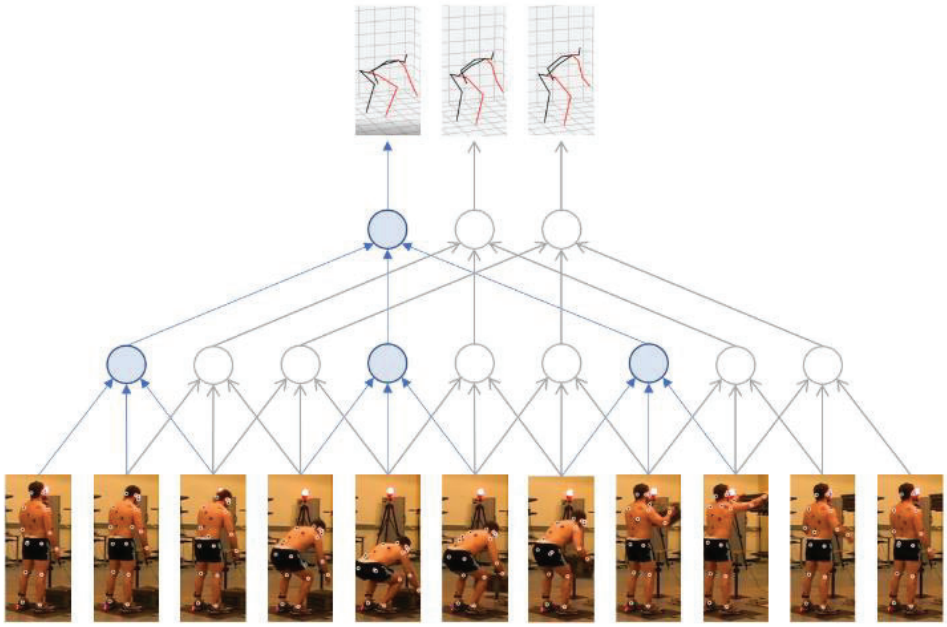

Figure 2. The temporal convolutional model used in the proposed method. The inputs are the 2D key-point sequences (bottom), the middle is the intermediate convolution process, and the outputs are the 3D poses (top). The implementation for a single-frame prediction is highlighted.

human kinetics analysis, the coordinates of the reconstructed key joints are translated in a way that the coordinates of mid-hip joint are considered as the origin.

Moment calculation. To estimate L5/S1 moment during a lifting task, we developed a biomechanical model in MATLAB programs (R2020b, The MathWorks, Boston, MA) following a top-down inverse dynamic algorithm. This model selects 9 of 17 key-point joints output from VideoPose3D and 13 of 45 markers in motion tracking system, respectively, to estimate positions of 10 key joint centers, including left/right hip, left/right shoulder, left/right elbow, left/right wrist, C7, and L5/S1 joint. Based on the approaches presented by De Leva(Leva, 1996), body segments including upper arms, forearms, hands, trunk above L5/S1 joint are defined in this model.

Body segment inertial properties, including mass (m) and moment of inertia (I), are estimated from previous anthropometry studies (Zatsiorsky, 2002) as well as participants' weight and stature. The center of mass location (CoM_i) of each body segment i is determined as a proportional location of the segment length, which can be determined from the distal and proximal joint center location.

L5/S1 joint moments (M_{L5S1}) is then calculated by an inverse dynamics model presented in (Leva, 1996). The equation applied in this model is described as:

$$M_{L5S1} = -(r_r - r_{L5S1}) \times F_r - \sum_{i=1}^{k}[(r_i - r_{L5S1}) \times m_i g] + \sum_{i=1}^{k}[(r_i - r_{L5S1}) \times m_i a_i] + \sum_{i=1}^{k}(I_i \alpha_i) \quad (1)$$

where F_r is the external force applied on the hands; $m_i g$, a_i and $I_i \alpha_i$ are gravity, acceleration and angular momentum of body segment i that are above L5/S1 joint; r_r, r_i and r_{L5S1} are the position vectors of the external force, center of segment mass and L5/S1 joint, k is the number of segments included in this model (upper arms, forearms, hands, and trunk). Note the external force applied on the hands is estimated based on the hand acceleration and the mass of the crater.

Low-back joint moment validation

The performance of our proposed single-camera computer vision-based method is validated against the motion tracking system-based method. The 3D motion coordinates from both methods were first filtered by a fourth-order Butterworth low-pass filter at 8 Hz. A comprehensive top-down biomechanical model (Leva, 1996) was then applied to estimate the L5/S1 joint moment. The peak moment estimated from both the computer vision-based method and the motion tracking system are extracted. Linear regression is then performed between them. The root mean squared error (*RMSE*), the absolute percent error, and correlation coefficient (r) are also calculated to describe the performance of the computer vision-based system. A histogram of the estimation error across all trails is constructed for peak total L5/S1 moment.

RESULTS

An example of a lifting trial is presented in Figure 3, showing the total moment variation at L5/S1 joint over a lifting task calculated based on the proposed computer vision-based method against the motion tracking system-based method. The

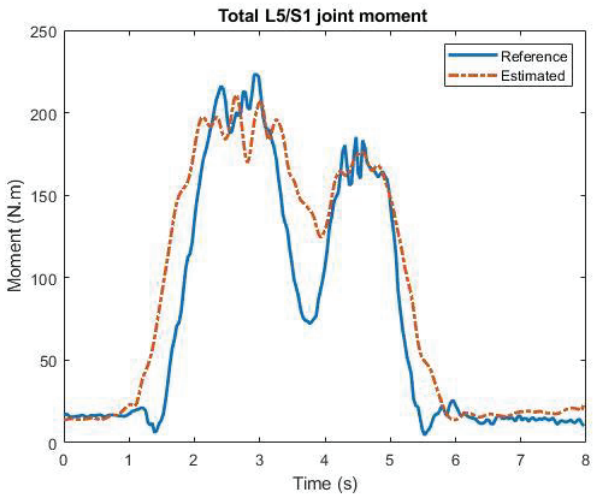

Figure 3. An example of the estimated total L5/S1 moment (computer vision-based method) vs. reference total L5/S1 moment (motion tracking system-based method).

estimated L5/S1 joint moment and the reference are in good correspondence. The computer vision-based method yields a good estimate on the peak total L5/S1 moment (Figure 4). The correlation coefficient r is 0.832. Although the root mean square error (RMSE) is 13.64 N·m, the absolute percentage error is only 7.7% since the magnitude of the total L5/S1 moment is relatively large.

To indicate overestimation and underestimation, the estimation error is also computed from the reference moment. The histogram of the estimation error across all trails for peak total L5/S1 moments reveals that the error distribution is symmetric and approximately zero centered (Figure 5).

DISCUSSION

In this study, we presented a single-camera computer vision-based method to estimate 3D L5/S1 joint moment. The input of this method is the videos capture by a single camcorder. This method was then validated against the reference L5/S1 moment derived from a laboratory-based motion tracking system. The correlation coefficient and the linear regression outcomes indicate that the estimated total peak moments are positively correlated to the reference L5/S1 joint moment measured by a motion tracking system.

There are a few limitations that need to be addressed for this proposed method. First, due to the resolution of the videos, small body motion may not be precisely captured. Thus, the sensitivity of this method is lower than a motion tracking system. Second, the performance of the joint detection can be affected when the view of a body segment is blocked by other objects. Once a view block occurs, disturbances in the body trajectories will lead to errors in kinematic calculation, especially for body segment acceleration estimation This error then reduce the accuracy of the estimated joint moment. Third, because VideoPose3D can only output the 3D location of certain joints, we have to extrapolate the location of few key joints for the inverse dynamic calculation. For example, the

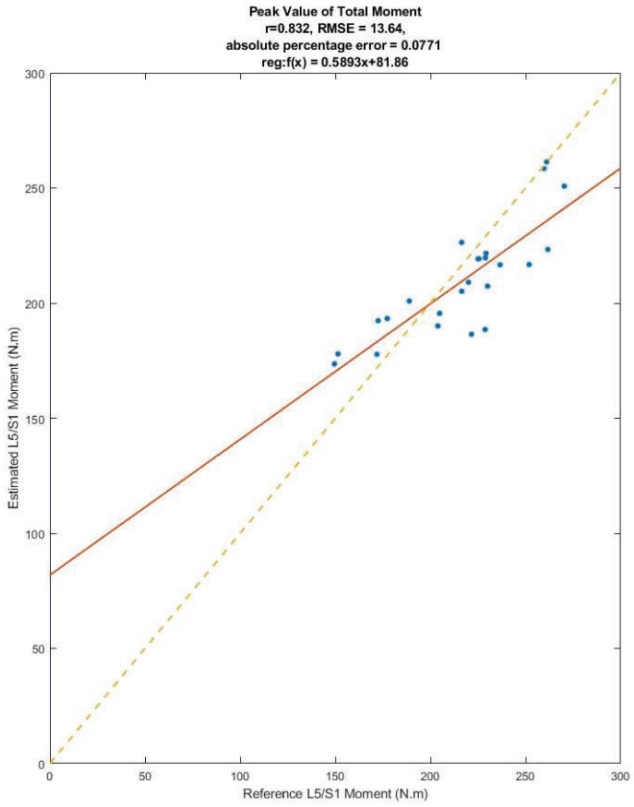

Figure 4. The comparison between the reference total L5/S1 peak moment and the estimated total L5/S1 peak moment. r is the correlation coefficient. *RMSE* is the root-mean-square error. *Reg* refers to the linear regression between the reference and estimated moments. The solid line is the linear regression line that generated from the data points and the dashed diagonal line is the identity line.

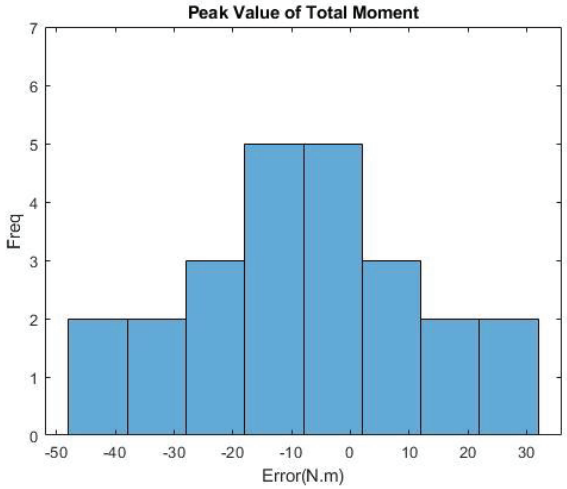

Figure 5. Histograms of the estimation error of the peak total L5/S1 moments.

position of C7 is not identified through the video. Thus, we estimated the positions of C7 based on the positions of shoulder joints and hip joints in our model (Chaffin & Anderson, 1991). The error introduced in position extrapolation could also

contribute to the L5/S1 moment estimation error. Finally, we assumed an equal weight distribution on both hands in our top-down inverse dynamics model. This assumption may be violated if the weight in a crate is not well balanced.

CONCLUSION

The results of this study show a good potential of using a single RGB camera to perform low-back joint loading estimation for manual material handling tasks. While the accuracy of the L5/S1 moment estimation can be further improved, this single-camera real-time method could facilitate ergonomics practitioners to quickly catch the jobs in the field with high risks of low-back injuries.

ACKNOWLEDGEMENT

The authors are grateful to Dr. Jacob Banks, Niall O'Brien and Amanda Rivard for assistance in data collection and data post-processing. This manuscript is based upon work supported by the National Science Foundation under Grant # 2013451.

REFERENCES

Da Costa, B. R., & Vieira, E. R. (2010). Risk factor for work-related musculoskeletal disorders: A systematic review of recent longitudinal studies. American Journal of Industrial Medicine, 53, 285–323.

Hoogendoorn, W. E., Bongers, P. M., de Vet, H. C. W., Douwes, M., Koes, B. W., Miedema, M. C., Ariens, G. A. M., & Bouter, L. M. (2000). Flexion and rotation of the trunk and lifting at work are risk factors for low back pain: Results of a prospec- tive cohort study. Spine, 25, 3087–3092.

Kuiper, J. I., Burdorf, A., Verbeek, J. H. A. M., Frings-Dresen, M. H. W., van der Beek, A. J., & Viikari-Juntrra, E. R. A. (1999). Epidemiologic evidence on manual materials handling as a risk factor for back disorders: a systematic review. International Journal of Industrial Ergonomics, 24, 389–404.

Schaffer, H., & United States. (1982). Back injuries associated with lifting. Washington, D.C: The Bureau.

McGill, S. M. (1997). The biomechanics of low back injury: Impli- cations on current practice in industry and the clinic. Journal of Biomechanics, 30, 465–475.

Jiang, Z. L., Shin, G., Freeman, J., Reid, S., & Mirka, G. A. (2005). A study of lifting tasks performed on laterally slanted ground surfaces. Ergonomics, 48, 782–795.

Callaghan, J. P., Salewytsch, A. J., & Andrews, D. M. (2001). An evaluation of predictive methods for estimating cumulative spinal loading. Ergonomics, 44, 825–837.

Pfister, A. M. West, S. Bronner, and J. A. Noah, "Comparative abilities of microsoft kinect and vicon 3Dmotion capture for gait analysis," J. Med. Eng. Technol., vol. 38, pp. 274–280, 2014.

Xu, X., Chang, C. C., Faber, G. S., Kingma, I., & Dennerlein, J. T. (2012). Estimating 3-D L5/S1 moments during manual lifting using a video coding system: Validity and interrater reliability. In Human Factors (Vol. 54, Issue 6, pp. 1053–1065).

Mehrizi, R., Peng, X., Metaxas, D. N., Xu, X., Zhang, S., & Li, K. (2019). Predicting 3-D lower back joint load in lifting: A deep pose estimation approach. IEEE Transactions on Human-Machine Systems, 49(1), 85–94.

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2021). OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In IEEE Transactions on Pattern Analysis and Machine Intelligence (Vol. 43, Issue 1, pp. 172–186).

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 (Vols. 2017-January, pp. 4645–4653). https://doi.org/10.1109/CVPR.2017.494

D'Antonio, E., Taborri, J., Palermo, E., Rossi, S., & Patane, F. (2020). A markerless system for gait analysis based on OpenPose library. I2MTC 2020 - International Instrumentation and Measurement Technology Conference, Proceedings, 19–24.

Pavllo, D., Feichtenhofer, C., Grangier, D., & Auli, M. (2019). 3D human pose estimation in video with temporal convolutions and semi-supervised training. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019-June, 7745–7754. https://doi.org/10.1109/CVPR.2019.00794

He. K, Zhang. X, Ren. S, and Sun. J. Deep residual learning for image recognition. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016. 3

Ioffe. S and Szegedy. C . Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International Conference on Machine Learning (ICML), pages 448–456, 2015. 3, 5

Nair. V and Hinton. G. E. Rectified linear units improve re-stricted boltzmann machines. In International Conference on Machine Learning (ICML), pages 807–814, 2010. 3

Srivastava. N, Hinton. G, Krizhevsky. A, Sutskever. I, and Salakhutdinov. R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):1929–1958, 2014. 3

Zatsiorsky, V. M. (2002). Kinetics of human motion. Champaign, IL: Human Kinetics

De Leva .P, "Adjustments to Zatsiorsky-Seluyanov's segment inertia parameters," J. Biomech., vol. 29, no. 9, pp. 1223–1230, 1996.

Chaffin, D. B., & Anderson, C. K. (1991). Occupational biomechanics. New York, NY: Wiley.