A Video Is All You Need: Learning Fine-Grained Running Form Representations from Monocular Sources

Beck LaBash
Northeastern University
labash.b@northeastern.edu

Oliver Marker
Northeastern University
marker.o@northeastern.edu

Abstract

We train a sequence to sequence model to learn a dense representation of a subject's running form, given a single input video of them running. To achieve this, we compile a dataset of running videos consisting of diverse styles and body types. The videos are first preprocessed to extract 2D limb positions. We then fine-tune a Dual-stream Spatio-temporal Transformer on this dataset with the 2D to 3D pose lifting objective. We show that the latent representation learned by the transformer, combined with simple vector classification methods, can be used to identify flaws in form, potentially predicting future injuries.

1 Introduction

The specific nuances in an individual's running form (especially at an elite level), can affect their overall speed, efficiency, and ability to recover quickly. For example, Barnes et al. (2014) found that an individual's running economy (energy efficiency) was highly correlated to form attributes like leg stiffness and moment arm length. In more extreme cases, imbalances and deficiencies in one's running form can lead to a myriad of overuse injuries. Daoud et al. (2012) found that runners who habitually rearfoot strike have significantly higher rates of repetitive stress injury than those who mostly forefoot strike. Additionally, Schubert et al. (2014) found that increased stride rate appears to reduce the magnitude of several key biomechanical factors associated with running injuries. As such, being able to systematically analyze one's running form is essential to improving efficiency and mitigating injury.

Current SOTA running form analysis methods involve a combination of multi camera systems, motion capture systems, and medical expertise. Such methods are extremely expensive in terms of equipment and clinician hours, making them inaccessible to a large proportion of athletes. We propose a cheap, flexible running form analysis framework that significantly reduces this barrier to entry. Specifically, we first learn fine-grained representations of human running form by fine-tuning MotionBERT from Zhu et al. (2023), a transformer-based general human motion encoder, on a curated dataset of 2D running pose sequences. Since 2D pose sequences can be extracted from monocular video, equipment cost is reduced to a single camera, which most carry around in their pockets. Then, we demonstrate that a simple SVM can be used to classify whether a runner is overstriding or not, given their form representation. This eliminates the clinician hours required to review and analyze footage case by case. While we label the SVM's training dataset ourselves, in practice, clinician hours would be reduced to the initial development of a high quality training dataset. To further demonstrate the flexibility of the learned representations, we use cosine similarity to compare an input running sequence to a to a set of professional runners.

Preprint. Under review.

2 Methods

2.1 Dataset Curation

We knew we would need hundreds of running sequences to fine tune the model and accurately make conclusions about one's running form. However, a major hurdle arose due to the absence of a comprehensive and easily accessible dataset containing videos of people running.

To overcome this obstacle, we had to manually collect data by scraping YouTube and Instagram for videos/reels under the #running or similar tags. Subsequently, we screened each of these videos, extracting specific uncut and unedited segments to build our dataset.

Our efforts resulted in a dataset comprising 756 unlabeled videos, sourced primarily from workout videos and professional races, with each clip ranging from 2 to 15 seconds. The curation process aimed at creating a diverse dataset, including clips of runners of different genders, skill levels, overall speeds, varying qualities, and distances to the runner. Notably, each clip could feature multiple runners, significantly increasing the number of sequences available for fine-tuning the model beyond the initial 756.

Following the collection of the unlabeled dataset, we recognized the need for smaller, labeled datasets for our downstream tasks. Two datasets were chosen: one with 18 videos of professional runners and another with 18 videos labeled as 'overstriding' or 'optimal' running form. The first dataset aimed to assess the similarity between amateur runners who might use this model and professional runners. The second dataset was intended to set up a classification problem to detect a common running deficiency, overstriding, which is known to lead to injuries.

2.2 2D Pose Extraction

We utilize RTMPose-M from Jiang et al. (2023), a top-down, CNN-based model, and its accompanying python inference library MMPose to extract 2D poses from videos in our dataset. We then convert the predicted keypoints from COCO (Lin et al. (2015)) format to Human3.6m (Ionescu et al. (2014)) format such that they are compatible with the MotionBERT pretraining data.

2.3 Fine-Tuning MotionBERT

MotionBERT is an encoder-only, dual-stream transformer based on the BERT architecture introduced by Devlin et al. (2019). It's dual-stream attention mechanism allows it attend to both a single keypoint across all timesteps, and all keypoints within each timestep, simultaneously.

MotionBERT is pretrained on the 2D to 3D pose lifting objective. Specifically, MotionBERT takes a sequence of size (T, J, C_{in}), projects it to sequence of size (T, J, C_e), and outputs a sequence of size (T, J, C_{out}) where T is the number of timesteps, J is the number of joint keypoints, C_{in} consists of the 2D input channels and a keypoint confidence channel, C_e consists of 512 latent representation channels, and C_{out} consists of the 3D prediction channels. MotionBERT's pretraining dataset consisted of a large, 3D human motion capture dataset with 17 body keypoints called Human3.6m from Ionescu et al. (2014). Since we only have access to 2D pose data, we follow Zhu et al. (2023) and calculate loss based on a re-projection of the predicted 3D poses into the 2D plane:

$$\mathcal{L}_{2D} = \sum_{t=1}^{T} \sum_{j=1}^{J} \delta_{t,j} \| \hat{X}_{t,j} - X_{t,j} \|_2 \tag{1}$$

Where $\delta_{t,j}$, $\hat{X}_{t,j}$, and $X_{t,j}$ are 2D detection confidence, predicted 3D joint position re-projected into 2D, and ground-truth 2D joint position respectively, at a single timestep, t and joint keypoint j.

Due to computational overhead, we decided to fine-tune MotionBERT-Lite: a smaller, 16M parameter version of MotionBERT with comparable performance. The architectural details of the DSTFormer backbone used by MotionBERT-Lite are given in the table below:

Parameter	Value
Maximum Sequence Length	243
Feature Dimension	256
MLP Ratio	4
Transformer Depth	5
Representation Dimension	512
Number of Attention Heads	8

Table 1: MotionBERT-Lite Architecture Configuration

We trained on samples of timestep length 30 from our dataset whose keypoint confidences were greater than 0.3 on average. We trained for 60 epochs and used an Adam optimizer with a learning rate of 0.0002, weight decay of 0.01 and learning rate decay of 0.99.

2.4 Downstream tasks

Our fine-tuned MotionBERT encoder provides an embedding of size 512 for each joint j, at each timestep t. In order to obtain a single representation for the entire input sequence that we can utilize in downstream tasks, we apply a mean pooling across the time dimension and concatenate across the joint dimensions. Specifically, we transform the (T, J, C_e) embedding into a (1, $J * C_e$) embedding.

In our initial experiment, we created embeddings for nine runners identified as overstriding and nine runners with what we considered to be near-optimal form, setting up a classification problem. For this task, we selected a support vector machine (SVM) as our model of choice. SVMs operate by computing the optimal hyperplane that separates the two classes, maximizing the distance to the nearest points from each class—commonly referred to as support vectors. Points located 'above' the margin are assigned to the positive class, while those 'below' the margin are assigned to the negative class. This methodology enables the SVM to discern distinct patterns between the overstriding and optimal form categories.

As an additional experiment, we build up an embedding database of various pro runner form representations. We then took an embedding representation of an amateur runner's form and computed the pairwise cosine similarity score for each pro runner. Cosine similarity is a scaled dot product in the range (-1, 1) where 1 indicates maximum similarity, 0 indicates no similarity, and -1 indicates maximum dissimilarity. The cosine similarity of two vectors, A and B is given by:

$$\text{cosine similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

3 Experiments & Results

3.1 Identifying Deficiencies: Overstriding

To identify overstriding, we recognized the need for a classifier capable of distinguishing differences in embeddings. Opting for a support vector machine (SVM), we specifically utilized the svm package from sklearn. Given our limited dataset of 18 videos (9 overstriding and 9 good running form), we employed the k-fold cross-validation technique to evaluate the model's accuracy.

Randomly splitting the dataset into 4 folds, the SVM demonstrated a remarkable ability to classify the training data, achieving accuracies of 1.0, 0.923, 0.857, and 0.785 for each fold, resulting in an average training accuracy of 0.891. Likewise, when applied to the test data, the model exhibited accuracies of 0.8, 0.6, 0.75, and 0.75 for the respective folds, yielding an average test accuracy of 0.725.

Additionally, we used t-NSE to visualize the clusters in two dimensions.

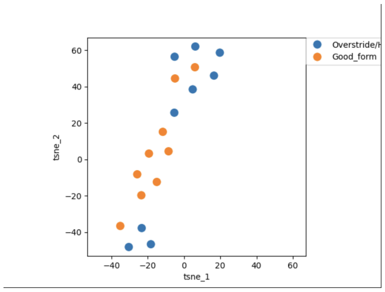

Figure 1: t-NSE visualization of overstriding clusters

3.2 Pro Runner Similarity

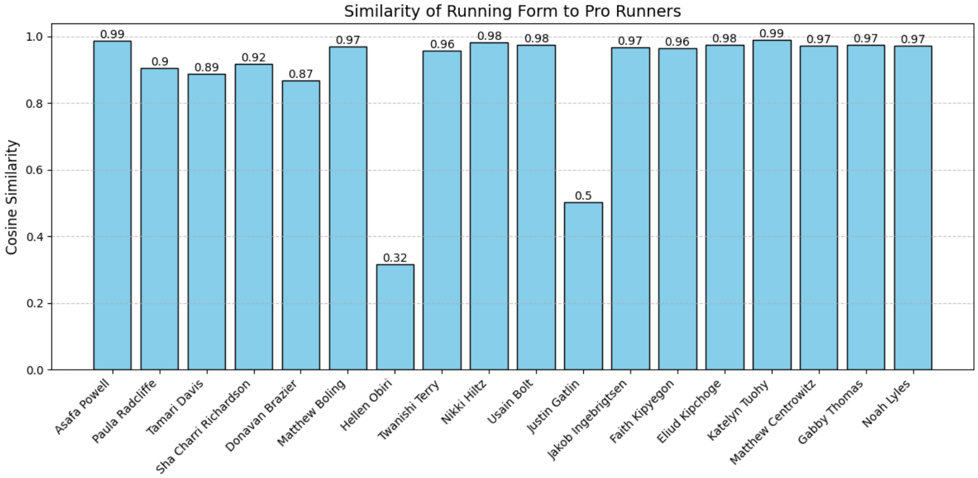

Figure 2: Similarity of a sample running sequence to various pro runners

4 Discussion & Conclusion

Despite our efforts, our training data distribution was skewed toward professional and experienced runners due to the nature of running content on the internet. The effect of this can be observed in the pro runner similarity experiment, where minimial differentiation is made between running forms in the embedding space. In the future, we would like to devote more time to curating an even more diverse running dataset to fine-tune on.

Additionally, we would like to investigate more robust multi-subject tracking methods across videos, as subject switching was observed in some dataset videos that contained occlusions.

In conclusion, we show that human running form representations can be learned by fine-tuning a general human motion encoder on a curated running dataset, and that the learned representations can be used to accurately classify common deficiencies in running form, at a significantly reduced cost to current SOTA methods.

5 Code Release

We publicly release the code for fine-tuning and downstream experiments here.

References

Barnes, K. R., McGuigan, M. R., and Kilding, A. E. (2014). Lower-body determinants of running economy in male and female distance runners. *Journal of Strength and Conditioning Research*, 28:1289–1297.

Daoud, A. I., Geissler, G. J., Wang, F., Saretsky, J., Daoud, Y., and Lieberman, D. E. (2012). Foot strike and injury rates in endurance runners: a retrospective study. *Medicine and science in sports and exercise*, 44 7:1325–34.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339.

Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., and Chen, K. (2023). Rtmpose: Real-time multi-person pose estimation based on mmpose.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft coco: Common objects in context.

Schubert, A. G., Kempf, J., and Heiderscheit, B. C. (2014). Influence of stride frequency and length on running mechanics. *Sports Health*, 6:210 – 217.

Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., and Wang, Y. (2023). Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.