APE-V: Athlete Performance Evaluation using Video

Chaitanya Roygaga[1], Dhruva Patil[2], Michael Boyle[2], William Pickard [2], Raoul Reiser[2], Aparna Bharati[1], Nathaniel Blanchard[2]
[1] Lehigh University, PA, USA [2] Colorado State University, CO, USA

Abstract

Athletes typically undergo regular evaluations by trainers and coaches to assess performance and injury risk. One of the most popular movements to examine in athletes needing lower extremity strength and power is the vertical jump. Specifically, maximal effort countermovement and drop jumps performed on bilateral force plates provide a wealth of metrics. However, the expense of the equipment and expertise needed to interpret the results limits their use. Computer vision techniques applied to videos of such movements are a less expensive alternative for extracting complementary metrics. Blanchard et al. *[4] collected a dataset of 89 athletes performing these movements and showcased how OpenPose could be applied to the data. However, athlete error calls into question 46.2% of movements — in these cases, an expert assessor would have the athlete redo the movement to eliminate the error. Here, we augmented [4] with expert labels of error and established benchmark performance on automatic error identification. In total, 14 different types of errors were identified by trained annotators. Our benchmark models identified errors with an F1 score of 0.710 and a Kappa of 0.457 (Kappa measures accuracy over chance). All code and augmented labels can be found at https://blanchard-lab.github.io/apev.github.io/.*

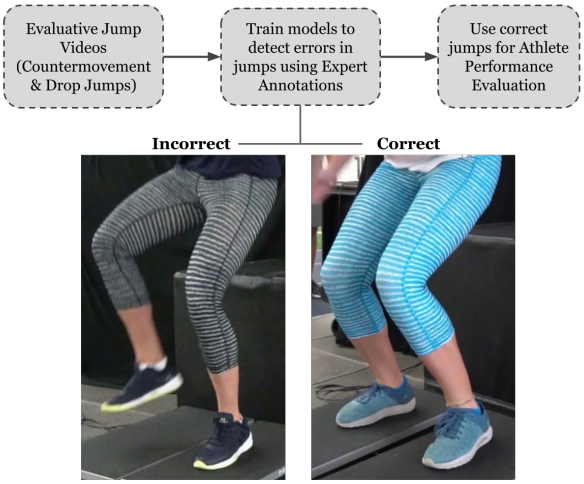

Figure 1. For athletes that rely on lower extremity strength and power, both countermovement and drop jumps have been used to assess metrics related to performance and injury risk. In the pursuit of a system to automatically provide feedback to athletes, we annotated a dataset of correct and incorrect jumps and trained machine learning models to automatically identify when jumps have been performed incorrectly. An 'Incorrect' technique includes landing on one foot, and the 'Correct' technique is landing simultaneously with both feet.

1. Introduction

Athletes of any caliber, from professionals to first time amateurs, try to limit injury while maximizing training/exercise. At the professional level, this balance is facilitated by trainers and coaches who perform expert-level assessments of athletes [5, 19] and customize training plans [34, 45]. Amateur athletes typically do not have access to the same equipment [20, 4] or personnel employed by professional athletes [9]. However, RGB cameras are ubiquitous. A computer vision system that performs personalized assessments using only RGB cameras would scale to athletes of all means. As an initial step toward such a system, it is necessary to assesses if the athlete performed the movement correctly.

RGB video has already been proven to be a valuable modality for athlete evaluation. Videos of movements are commonly used by clinicians to assess injury risk [40] and to perform performance evaluations [12, 29, 31, 11, 13]. Estimating pose and joint locations from video has become ubiquitous in computer vision [12, 29, 31, 11], and such information is even extractable from athletes in real time (for example, joints can be estimated while an athlete swims [29, 31]).

Blanchard *et al*. [4] published a video dataset of 89 athletes performing countermovement and drop jumps. Such movements are regularly used by professional trainers to evaluate athlete performance. However, performance evaluation was not the original intent of the dataset — Blanchard *et al*. [4] were explicitly interested in assessing jumps

for ACL injury risk. Since performance evaluation was not the intent, 46.2% of the jumps featured errors that make them unsuitable for evaluating performance [8]. We present APE-V — Athlete Performance Evaluation using Video — a performance-centric augmentation of [4] containing fine-grained annotations of errors found in the dataset. Specifically, we provide expert annotations of 14 error types for the countermovement and drop jump videos.

Finally, we establish an evaluative pipeline to automatically measure an athlete's performance without excess person-power, expertise, or machinery. We experiment with both pose estimates (from OpenPose [6]) and raw video frames, which we use to train three machine learning architectures (LSTM, ResNet-18, and TSM pretrained action recognition). We conduct hyperparameter optimization for architectures in all three scenarios and report results for the best models. Note, all evaluations are done using cross-validation, with no overlap between athletes in the training and test set, thereby establishing a person-independent assessment pipeline.

In summary, our work makes the following contributions:

1. This is the first work to provide expert annotations of 14 kinds of athlete errors found in countermovement and drop jumps.

2. We establish benchmarks showcasing the feasibility of automatic, person-independent assessment of athlete movements.

3. We investigate both raw-video and pose-centric methods, finding that pose-centric methods generalize better to unseen athletes.

2. Related Work

Athlete performance evaluations are conducted by trainers and coaches. However, traditionally, these evaluations require specialized equipment (e.g., force plates, leap measurement software, and agility ladders [2, 1, 43]). Force plates are one of the most common methods for such evaluations. From the force outputs, other metrics such as center of mass displacement and velocity as well as power output during the evaluative jumps [18, 23, 24, 42]. Studies have also been conducted using vertical jumps, to assess the Change of Direction Speed (CODS) along with inter-limb asymmetries [25, 35, 41].

Various screening methods have been used to analyze these jumps. For example, the Landing Error Scoring System (LESS) [48] has a trained practitioner evaluate a bilateral drop jump [38, 39, 40, 48], and the Functional Movement Screen (FMS) evaluates fundamental movement patterns in individuals with no pain complaints or musculoskeletal injury [27]. The LESS was developed to identify athletes with higher risk of injury, which can be utilized during a team screening session to save time and resources, and is shown to be accurate and reliable. Videos [40] for multiple angles of athlete jumps were recorded, and LESS scores were assigned in conjunction with the erroneous movement patterns across the multiple planes of motion. LESS evaluations showed that ACL injury risk could be visually identified from movement alone. This gives us confidence that injury risk and performance assessments can be done using only a visual medium.

Recent automated techniques used to analyze athlete motion include Einfalt *et al*.'s [12] technique of using 2D human pose sequences as a representation of the actual motion. They demonstrate two approaches for event detection in pose sequences — using multiple fixed cameras and domain information of sport, or using sequence recordings of athlete's motion from a single camera.

El-Sallam *et al*. [13] provide a markerless system for athlete performance optimization in the sports of pole vault, javelin throw, and jumping. They use multiple calibrated cameras for multiple view captures. This method segments the subject's body from video, and a 3D representation of the body is then reconstructed using silhouettes, which is then tracked over the frames in video. This method can be extended to other sports which do not explicitly need body joint detection, but can benefit from detection of the athlete body as a whole. Another markerless method of human motion tracking was developed by Saini *et al*. [44]. Their primary purpose was to detect the pose and position of the subject from video by comparing a rendered body model with the image in a video frame.

Elhayek *et al*. [15] suggest a method for capturing multiple 3D human skeletal movements, even with cluttered and moving backgrounds in videos captured from regular-use camera setups such as mobile phones. This method requires fewer cameras, and they may be unsynchronized. For single camera scenarios, Mehta *et al*. [37] combine a CNN-based pose regressor and kinematic skeleton fitting to propose a real-time 3D skeletal pose estimation method. This method is able to create a 3D representation of the real-time motion of the subject in a video by reconstructing a 3D skeleton based on joint predictions. In addition to 3D skeletal information, Cao *et al*. [7, 6] developed an efficient tool to detect 2D poses of multiple subjects in an image. They use Part Affinity Fields (PAFs) to establish pairwise relationships between body parts using their location and orientation. The technique of Elhayek *et al*. [14] combines a stable skeleton motion capture method and 2D joint detection using ConvNet for a kinematic skeleton model.

For identifying lower-body injury risk in athletes, Blanchard *et al*. [4] released a multi-angle video dataset of female athletes performing two specific athletic movements: countermovement and drop jumps. These evaluative jumps

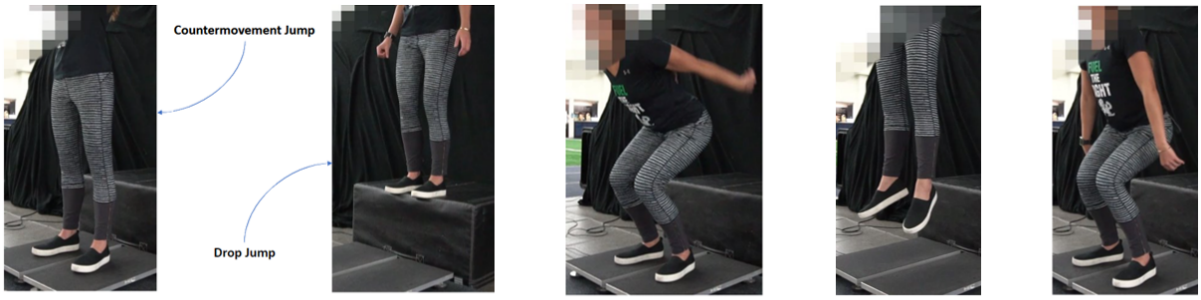

Figure 2. Important Events of an Evaluative Jump. The 'Start of Jump' is different for both jump types, while other events are similar.

are used for research in sports medicine for identifying athleticism and factors in an athlete's jump motion indicating ACL injury risk. The dataset is targeted towards Computer Vision researchers, who could build accurate models and track key movements in these jumps, for evaluating injury risk in the participants. One of the main features of this dataset is that the collection mechanism can be easily replicated, as it is inexpensive when compared with the high-end approaches that require non-portable setups.

Our work specifically looks at RGB video recordings of athletes performing two evaluative jumps: the countermovement and the drop jumps. The dataset consists of videos in which participants perform these evaluative jumps [4]. Unlike previous work, we note that athletes sometimes err when performing these movements; we provide expert-level annotation of such errors and train computer vision models to identify them. Long term, such models will be essential for ensuring that performance or risk assessments are accurate, and no measures are estimated from faulty data.

3. Experiments

3.1. Dataset

We augment the dataset from Blanchard *et al*. [4] with expert annotation of errors that would prevent accurate assessment of performance. Blanchard *et al*. [4] recorded videos from three different angles — center, left and right, as shown in Figure 3, of 89 participants performing two evaluative jumps — countermovement and drop jumps, as shown in Figure 2. Details on the original dataset can be found in [4] — however, note the dataset was expanded post-publication (from 55 athletes to 89). Summary of the extended dataset is presented in Table 1.

3.1.1 Annotation of Errors in jump motion

We found 14 types of errors in the videos — errors are broken down in Table 2. Note that each jump may have multiple errors. A correct evaluative jump performed by a participant has a few characteristics; we used these characteristics to identify pivotal errors in jumps. First, the participant assumes a straight posture while looking forward [33], jumping high enough after an initial squat, and then landing back in a similar position from which they started [28]. During these jump motions the participant might not start with the correct position, or, they might perform an irregular landing. Additionally, for the drop jump, the participant drops from a box onto the force plate [17]. A well executed drop

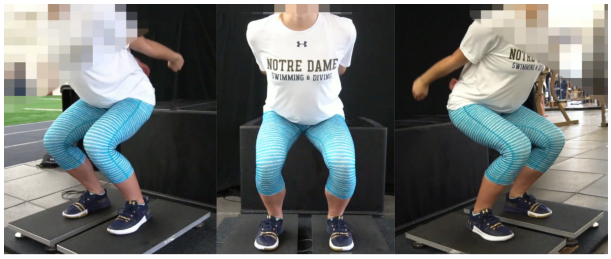

Figure 3. The dataset contains videos from multiple views.

Table 1. Video Dataset summary.

Total No. of Jumps	**582**
No. of Countermovement Jumps	346
No. of Drop Jumps	236
Total No. of Participants	**89**
No. of Participants Performing Countermovement and Drop Jumps	47
No. of Participants Performing only Countermovement Jumps	41
No. of Participants Performing only Drop Jumps	1
No. of Camera Views for each Video	**3**
Total No. of Videos	1746
Camera settings: No fixed camera angle and height.	

Table 2. Errors in jump motion. 14 annotated errors (sub-categories) in the evaluative jumps, along with the number of samples in the dataset. 10 errors are annotated for both jump types, while 4 errors are specific to the drop jump. Errors in bold are the 6 primary errors used to train the classification models [See "Errors in jump motion", Section 3.1].

ERRORS (OVERALL CATEGORIES)	JUMP TYPE	ERRORS (SUB-CATEGORIES)	NO. OF SAMPLES
START POSITION	BOTH	**FEET LESS THAN SHOULDER WIDTH APART**	30
INITIAL POSITION, AFTER START	DROP JUMP	**JUMPED UPWARD FROM BOX, RATHER THAN FORWARD**	22
		ASYMMETRIC LANDING AFTER JUMP	15
FIRST LANDING ON FORCE PLATE	DROP JUMP	**SQUAT TOO LOW**	37
		HEELS TOUCH FORCE PLATE	83
FIRST OR FINAL LANDING ON FORCE PLATE	BOTH	KNEE COLLAPSE	64
		BOTH FEET NOT ON RESPECTIVE PLATFORMS	5
		LAND OFF-BALANCE	49
DURING JUMP	BOTH	OFF-BALANCE	79
		BODY TWISTS, LANDING AT DIFFERENT ANGLE	74
FINAL LANDING ON FORCE PLATE	BOTH	**LANDED AT DIFFERENT POSITION FROM INITIAL LANDING**	133
		EXCESSIVE HIP AND KNEE FLEXION BEFORE RETURNING TO UPRIGHT STANDING POSITION	78
		TAKE ADDITIONAL STEPS TO MAINTAIN BALANCE	94
		FEET LESS THAN SHOULDER WIDTH APART	3

culminates with both feet touching simultaneously [46], followed by a quick reflex jump. Associated errors include jumping instead of dropping or lingering on the force plate for too long, rather than immediately jumping.

The most impactful errors are emphasized in bold in Table 2. The other eight errors tend to be more subtle deviations from the correct body movement, which provide supporting information regarding motion flaws in the jumps. In the long term, these annotations are essential for proper assessment, but for now, we do not consider jumps with only subtle deviations to be erroneous.

Two expert annotators labeled a subset (17%) of the dataset. Cohen's Kappa was used to assess inter-annotator agreement. Across all errors, the average Kappa was 0.89 (Max Kappa is 1.00), indicating very high agreement.

3.2. Model Training and Evaluation

We conducted baseline experiments to investigate the usability of the dataset and its corresponding annotations. We focused on two major questions: is video information enough to facilitate detection of errors during motion of evaluative jumps? And, if video information is good enough, what kinds of features (raw video or pose) provide optimal performance?

For our experiments, we used two types of video information — raw frames and pose information from OpenPose [6]. We also processed the videos to select a subset of frames for evaluating the models with less dense temporal information. The frames with differences in intensities larger than a data-defined threshold were retained. Different deep neural network architectures were employed on each type of input data to obtain view-specific detection results.

Figure 4. First Row: Key points from pose detection for a countermovement jump example; Second Row: Key points of a drop jump example with pose detection.

Table 3. Dataset sub-categories used for experiments.

Sr. No.	Dataset category	Features	Training
1	Clipped Athletic Data: every 10th frame, resize 256x256	Athlete jump video frames resized to 256x256. Includes every 10th frame from video. Video rate modified to 15 fps.	Train CNN and LSTM combination architectures, and TSM architectures using video frame features.
2	Clipped Athletic Data: keyframes, resize 256x256	Athlete jump video frames resized to 256x256. Includes only key-frames from video. Video rate modified to 15 fps.	
3	OpenPose [6] Skeleton Outputs: confidently detected frames	Skeleton outputs generated using OpenPose [6]. Frames stored in which hips, knees and ankles detected with confidence above 0.3.	Train LSTM models using x and y coordinates of the detected hips, knees and ankles.
4	OpenPose [6] Skeleton Outputs: keyframes	Skeleton outputs generated using OpenPose [6]. Frames stored in which hips, knees and ankles detected with confidence above 0.3, which are then filtered to keep only keyframes.	

Table 4. Hyperparameter Search Space for network architectures. Learning rate search range used for all three types is [0.0001 - 0.1].

Model Type	Pre-trained	Trainable Layers	Hidden Nodes	Batch Size	Epochs
LSTM	✗	1 - 4	10-200	[8, 16, 24, 32, 40, 48, 56, 64, 128, 256]	10-200
LSTM + ResNet-18 features	✓	Upto 5 end layers	NA	[8, 10, 12, 14, 16, 24, 32, 40, 48]	5-45
TSM [32]	✓	Upto 6 end layers	NA	[4, 8, 16]	15-75

3.2.1 Training procedure

A standard training procedure was followed across all experiments. The experiments were designed to run hyperparameter search from the given set of hyperparameters [Table 4]. Hyperopt [3] with the Tree Parzen Estimator (TPE) algorithm was used for this purpose. For each hyperparameter combination, the models were trained and evaluated using 5-fold cross validation. K-fold cross validation helps to evaluate a given model on the entire dataset, providing more robust measures of performance for small datasets. As features in the video frames might be similar for a particular participant, the data is distributed into folds based on participants and not jumps. Each model trained on (K-1) partitions, during cross validation, is then evaluated for validation loss at every epoch. The model corresponding to the lowest validation loss is saved. Note that not all participants performed the same number of jumps.

Models saved for each of these data folds at the end of the training cycle were then evaluated using Cohen's Kappa score and F1 score (positive error class). These values were used for selecting the best models during hyperparameter search. The data is unevenly distributed — 46% positive class, and 54% negative class. Cohen's Kappa represents how well a model performs when compared to a model that randomly predicts an output (i.e., accuracy above chance). A positive score for Kappa indicates that the model performs better than chance. After individual models are trained for each of the five data folds, metric scores were averaged across folds.

3.2.2 Network Architecture

Our experiments used three machine learning architectures. The first experiment [Section 4.1] used Long Short-Term Memory (LSTM) networks, which were trained from scratch on pose estimation joint data. This information was extracted from the evaluative jump videos using OpenPose [6], as discussed in Section 3.1. The LSTM-based architecture was chosen to learn the order dependence between items in a long data sequence, and is suitable for the task of detecting changes in athlete motion through the video frames. The models were trained on data subsets three and four [Section 3.1, Table 1].

The second and third set of experiments [Section 4.2] are end-to-end approaches, as they operated directly on video frames. The second type of architecture used a combination of a pretrained ResNet-18 and an LSTM architecture. The ResNet-18 [22], after fine-tuning, acted as a feature extractor for video frames in the training data. These features were fed into an LSTM architecture, which was trained from scratch. We chose ResNet-18 as the feature extrac-

Table 5. OpenPose Experiments: Comparison between best models trained on data from Confident Frames (CF) — which are selected based on confidence threshold of 0.3 — and Confident Keyframes (CKF).

Experiment		Test Acc. %	F1 Score: Error in Jump	F1 Score: No Error in Jump	Cohen Kappa
CF	Center	69.2 ±8.01	**0.629 ±0.10**	0.737 ±0.07	**0.374 ±0.16**
	Left	62.2 ±4.60	**0.508 ±0.21**	0.682 ±0.04	**0.224 ±0.13**
	Right	67.8 ±3.63	**0.635 ±0.06**	0.716 ±0.03	**0.356 ±0.08**
	Combined View	72.4 ±4.72	**0.710 ±0.05**	0.743 ±0.05	**0.457 ±0.09**
CKF	Center	64.6 ±6.80	**0.514 ±0.26**	0.698 ±0.10	**0.268 ±0.18**
	Left	63.4 ±5.73	**0.509 ±0.24**	0.692 ±0.06	**0.250 ±0.16**
	Right	67.0 ±3.61	**0.625 ±0.08**	0.710 ±0.02	**0.337 ±0.09**
	Combined View	70.8 ±6.83	**0.658 ±0.10**	0.744 ±0.06	**0.407 ±0.15**

Table 6. OpenPose experiments – Threshold comparison. We verify the use of Threshold 0.3 across all experiments for extracting the OpenPose skeleton outputs. Comparison is made based on the Cohen Kappa score. The values in bold signify the best results in that experiment, and the corresponding column gives the threshold used for pose data.

Experiment		Cohen Kappa				
		Threshold 0.1	Threshold 0.2	Threshold 0.3	Threshold 0.4	Threshold 0.5
CF	Center	0.117	0.248	**0.374**	0.285	0.239
	Left	0.153	0.155	**0.224**	0.153	0.175
	Right	0.130	0.314	**0.356**	0.224	0.325
CKF	Center	0.138	0.231	**0.268**	0.191	0.203
	Left	0.246	0.191	**0.250**	0.194	0.142
	Right	0.195	0.234	**0.337**	0.296	0.178

tor instead of deeper CNNs as we wanted to train simple and efficient architectures for the baseline models; ResNet-18 has good representation capacity as shown in recognition tasks on ImageNet [10], as well as other tasks like gender and activity recognition [49] and head pose estimation [16]. Finally, a fine-tuned version of the original TSM [32] action recognition model was used for the third set of experiments. It used a Temporal Shift Module (TSM) which shifted a subset of image features along the temporal dimension, providing information passage between successive frames. The architecture provides state-of-the-art performance on video related datasets [47, 26, 21], hence we use it as a baseline to evaluate jump videos with the curated error labels. We only show results from the second and third approaches using data subset one [Section 3.1, Table 1], as keyframe selection did not have a significant effect on performance. All three architectures were trained using multiple combinations of hyperparameters, as discussed in Table 4.

3.2.3 Multi-view Fusion

To understand if a model trained on multiple sources of information for the same task could perform better than the individual models, we perform experiments by combining the best models trained on individual view data. Specifically, the classification layer of the best models trained on each of these views was replaced with a separate classifier layer at the end of the combined architecture. This new layer was trained for a few epochs, and the trained model was then evaluated in the same manner as the individual models described in Section 3.2.1.

4. Results and Analysis

Using different data modalities and architectures described above, we obtain results on the novel annotated dataset for the task of erroneous jump evaluation. We focus on answering questions such as which type of data input and architecture performs better, and how does view information affect detection results.

4.1. Models trained on OpenPose [6] skeleton

This experiment evaluates if lower body joint coordinates detected on athletes performing the evaluation jumps are sufficient to train a machine learning model to distinguish between erroneous jumps and those useful for athlete evaluation. We used joint coordinates detected from OpenPose [6] for hips, knees, and ankles, for the full length of the videos or selected frames, to train a LSTM model.

Table 7. Experiments on confident frames extracted from Videos: ResNet-18 + LSTM models.

EXPERIMENT	TEST ACC. %	F1 SCORE: ERROR IN JUMP	F1 SCORE: NO ERROR IN JUMP	COHEN KAPPA
CENTER	62.9 ±4.15	**0.642 ±0.04**	0.611 ±0.06	**0.270 ±0.08**
LEFT	66.1 ±5.51	**0.593 ±0.13**	0.693 ±0.06	**0.298 ±0.13**
RIGHT	62.9 ±6.14	**0.627 ±0.07**	0.620 ±0.09	**0.266 ±0.11**
COMBINED VIEW	67.5 ±3.72	**0.690 ±0.04**	0.653 ±0.06	**0.360 ±0.07**

Table 8. Experiments on confident frames extracted from Videos: Fine-tuning TSM pretrained models.

EXPERIMENT	TEST ACC. %	F1 SCORE: ERROR IN JUMP	F1 SCORE: NO ERROR IN JUMP	COHEN KAPPA
CENTER	64.5 ±4.90	**0.649 ±0.08**	0.627 ±0.08	**0.293 ±0.09**
LEFT	61.1 ±5.94	**0.520 ±0.13**	0.652 ±0.10	**0.198 ±0.12**
RIGHT	62.5 ±8.11	**0.580 ±0.11**	0.658 ±0.07	**0.239 ±0.17**
COMBINED VIEW	66.0 ±5.89	**0.624 ±0.13**	0.675 ±0.04	**0.306 ±0.14**

Three individual models were trained on lower body joint data from the center, left, and right view videos using two types of input sequences. The first category of data were video frames for which OpenPose was highly confident in detecting lower body joints. The second category had frames which highlighted the key pose changes during the jump motion. About 100 models, with different hyperparameters, were trained during each of these experiments, and results from the best performing models were presented in Table 5. These results show that among the single view models, the center view provided the best model performance as it contained more details compared with the left and right view, which may occlude key points. Also, dropping some temporal information when using selected keyframes hurt the performance.

For the view fusion model, we combined the individual models trained on the different views for each type of data input. The results showed that the performance of the model that uses all available view information was better than the single view models for both the full set of confident frames and the selected subset. Table 9 shows the error-wise accuracies of both fusion models (see Supplemental Material for error-wise accuracies of all models presented in the paper). We see that both models perform similarly for the "Feet less than shoulder width apart" and the squat-related errors ("Squat too low", and "Excessive hip and knee flexion before returning to upright standing position"). This could mean that reducing the amount of frame data does not largely affect the performance of the models for errors related to a fixed position (Feet error), or related to the lowest point in a movement (squat-related errors). The other errors show a drop in performances when the number of frames are reduced; all such errors are motion-related, highlighting the need for additional transitioning frame data.

4.1.1 OpenPose alternatives: EvoSkeleton and evopose2d

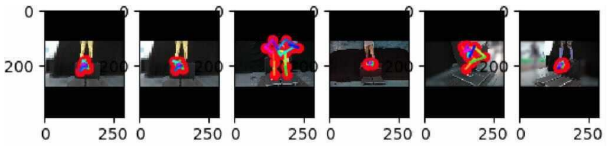

Figure 5. EvoSkeleton outputs for select frames from various participants, jump types, and angles.

We explored two alternatives to OpenPose — EvoSkeleton [30] and evopose2D [36] — to see if these pose estimating algorithms provided us with more accurate joint data than OpenPose. EvoSkeleton [30] showed incompatibility with the data because it was intended for full bodies. The videos from Blanchard *et al*. [4] focus on the lower body, with the upper body spanning out of frame during the flight phase. An example sequence is depicted in Figure 5.

The other alternative, evopose2d [36], also performed poorly on the data. Notably, evopose2D provides a means of fine-tuning on data. Unfortunately, our video data lacks the frame-wise keypoint annotations needed for the fine-tuning process. Thus, we determined neither EvoSkeleton nor evopose2d were acceptable alternatives to OpenPose [4].

4.1.2 Confident Frame Selection from OpenPose

We selected frames where the pose detection model had a confidence of c or higher in its predictions, as confident frames for our experiments. To obtain the threshold c, we performed additional experiments. We trained the same architecture with the subset of data extracted using different thresholds. The architecture and hyperparameters remained constant, and only the data changed. This helped us eval-

Table 9. Error-wise model accuracies (%) for the multi-view fusion scenario of presented architectures . F-SW: Feet less than shoulder width apart; JU: Jumped upward from box, rather than forward; SL: Squat too low; LDP: Landed at different position from initial landing; EF: Excessive hip and knee flexion before returning to upright standing position; AS: Take additional steps to maintain balance. Number of instances for each error are provided in parentheses.

EXPERIMENT	ACCURACY	F-SW (30)	JU (22)	SL (37)	LDP (133)	EF (78)	AS (94)
OPENPOSE (CF)	72.4	76.7	90.9	91.9	77.4	76.9	65.9
OPENPOSE (CKF)	70.8	80.0	68.2	86.5	66.2	82.1	44.7
RESNET-18 + LSTM	67.5	86.7	95.5	83.8	78.9	83.3	78.7
TSM	66.0	83.3	77.3	81.1	66.9	76.9	69.1

uate the effect of different thresholds used for obtaining a good set of video pose features from raw video frames.

From the comparison in Table 6, it is observed that models performed best with joint data extracted with a confidence threshold of 0.3. This threshold eliminates many noisy frames with fluctuating pose estimations for the lower body joints, while retaining ample information to train good models for any of the three views.

4.2. Models trained on Pixel Data from Video Frames

We evaluated two baseline models trained directly on pixel information [Section 3.1, Table 1]. We utilized all frames to train these models, since results in Table 5 indicated that training the models on keyframe information led to lower performing models.

The results for the model using fine-tuned ResNet-18 [22] features from video frames used with a LSTM model trained from scratch, are as shown in Table 7. We see a trend of improved model performance with the combined model in comparison to single view models, as witnessed with the models using pose data. When compared with the fusion model using the skeleton data, we see an overall dip in performance. This could be due to extra attributes in terms of image noise being presented to the model, or due to the limited ability of the pretrained ConvNet (ResNet-18) to extract the relevant features even after fine-tuning it on the available dataset.

The pretrained TSM [32] action recognition model, which was fine-tuned using the video frame data, was the third type of model evaluated in our experiments. The TSM model took as input a specific number of frames. We ran experiments for training these models on 8, 16, 32, and 64 frames, out of which the models trained on 16 frames performed best. The combined model uses the features from individual models trained on 16 frames. Results for this method are shown in Table 8. Individual models perform similar to the second set of experiments. This could be attributed to the absence of important temporal features with respect to the reduced number of frames presented to the model; experiments with input frames greater than 64 could provide better model performance.

Observing the error-wise accuracies of fusion models of experiments on pixel data in Table 9, we see an improvement in performance for the error related to a fixed position (Feet error). The ResNet-18 and LSTM combination's fusion architecture performs better for errors related to motion when compared to the models trained on OpenPose data [Discussion from Section 4.1], indicating that the additional pixel information improves the overall performance when identifying motion-related errors.

5. Conclusion

In this work, we presented expert-level error annotations for a jump video dataset [4] to facilitate fitness assessment from RGB video. Further, we provided baselines showcasing that, while these annotations include relatively fine-grained phenomena, it is feasible to identify them with computer vision techniques. We present automated approaches to detect and screen out improper techniques present in jumps performed for athlete evaluation, so that time and expertise can be allocated for assessing only the correctly performed jumps, and feedback can be provided for improving jump motion of those which are discarded.

Accurately evaluating athletes based on movements recorded with ubiquitous RGB cameras has a multitude of implications for fitness recommendations. Ideally, accurate evaluations can enable widespread access to state-of-the-art fitness recommendations. Although this work does not focus on injury prevention, an implicit side effect of appropriate fitness recommendations is the limitation of overextension, which can lead to injury. We anticipate that releasing the resources presented in this work is essential for future investigations into performance assessments.

References

[1] José Afonso, Israel Teoldo da Costa, Miguel Camões, Ana Silva, Ricardo Franco Lima, André Milheiro, Alexandre Martins, Lorenzo Laporta, Fábio Yuzo Nakamura, and Filipe Manuel Clemente. The effects of agility ladders on performance: A systematic review. *International journal of sports medicine*, 2020.

[2] George Beckham, Tim Suchomel, and Satoshi Mizuguchi. Force plate use in performance monitoring and sport science testing. *New Studies in Athletics*, 29(3):25–37, 2014.

[3] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR, 2013.

[4] Nathaniel Blanchard, Kyle Skinner, Aden Kemp, Walter Scheirer, and Patrick Flynn. "keep me in, coach!": A computer vision perspective on assessing acl injury risk in female athletes. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1366–1374. IEEE, 2019.

[5] Pitre C Bourdon, Marco Cardinale, Andrew Murray, Paul Gastin, Michael Kellmann, Matthew C Varley, Tim J Gabbett, Aaron J Coutts, Darren J Burgess, Warren Gregson, et al. Monitoring athlete training loads: consensus statement. *International journal of sports physiology and performance*, 12(s2):S2–161, 2017.

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[8] Carlo Castagna, Marco Ganzetti, Massimiliano Ditroilo, Marco Giovannelli, Alessandro Rocchetti, and Vincenzo Manzi. Concurrent validity of vertical jump performance assessment systems. *The Journal of Strength & Conditioning Research*, 27(3):761–768, 2013.

[9] Stephanie H Clines, Cailee E Welch Bacon, Christianne M Eason, Kelly D Pagnotta, Robert A Huggins, and Bonnie L Lunen. Athletic directors' perceptions regarding the value of employing athletic trainers in the secondary school setting. *Journal of Physical Education and Sports Management*, 2019.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[11] Moritz Einfalt, Charles Dampeyrou, Dan Zecha, and Rainer Lienhart. Frame-level event detection in athletics videos with pose-based convolutional sequence networks. In *Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*, pages 42–50, 2019.

[12] Moritz Einfalt and Rainer Lienhart. Decoupling video and human motion: towards practical event detection in athlete recordings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 892–893, 2020.

[13] Amar A. El-Sallam, Mohammed Bennamoun, Ferdous Sohel, Jacqueline A. Alderson, Andrew Lyttle, and Marcel Mourao Rossi. A low cost 3d markerless system for the reconstruction of athletic techniques. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 222–229. IEEE, 2013.

[14] Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin, Micha Andriluka, Chris Bregler, Bernt Schiele, and Christian Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3810–3818, 2015.

[15] Ahmed Elhayek, Carsten Stoll, Kwang In Kim, and Christian Theobalt. Outdoor human motion capture by simultaneous optimization of pose and camera parameters. In *Computer Graphics Forum*, volume 34, pages 86–98. Wiley Online Library, 2015.

[16] Ahmet Firintepe, Mohamed Selim, Alain Pagani, and Didier Stricker. The more, the merrier? a study on in-car ir-based head pose estimation. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1060–1065. IEEE, 2020.

[17] Eamonn P Flanagan and Thomas M Comyns. The use of contact time and the reactive strength index to optimize fast stretch-shortening cycle training. *Strength & Conditioning Journal*, 30(5):32–38, 2008.

[18] Kevin R Ford, Gregory D Myer, Rose L Smith, Robyn N Byrnes, Sara E Dopirak, and Timothy E Hewett. Use of an overhead goal alters vertical jump performance and biomechanics. *The Journal of Strength & Conditioning Research*, 19(2):394–399, 2005.

[19] Carl Foster, Jose A Rodriguez-Marroyo, and Jos J De Koning. Monitoring training loads: the past, the present, and the future. *International journal of sports physiology and performance*, 12(s2):S2–2, 2017.

[20] David H Fukuda. *Assessments for sport and athletic performance*. Human Kinetics, 2018.

[21] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. arxiv 2015. *arXiv preprint arXiv:1512.03385*, 2015.

[23] Naruhiro Hori, Robert U Newton, Naoki Kawamori, Michael R McGuigan, William J Kraemer, and Kazunori Nosaka. Reliability of performance measurements derived from ground reaction force data during countermovement jump and the influence of sampling frequency. *The Journal of Strength & Conditioning Research*, 23(3):874–882, 2009.

[24] Franco M Impellizzeri, Ermanno Rampinini, Nicola Maffiuletti, and Samuele M Marcora. A vertical jump force test for assessing bilateral strength asymmetry in athletes. *Medicine & Science in Sports & Exercise*, 39(11):2044–2050, 2007.

[25] Matthew J Jordan, Per Aagaard, and W Herzog. Lower limb asymmetry in mechanical muscle function: a comparison between ski racers with and without acl reconstruction. *Scandinavian journal of medicine & science in sports*, 25(3):e301–e309, 2015.

[26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola,

Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[27] Kornelius Kraus, Elisabeth Schütz, and Ralf Doyscher. The relationship between a jump-landing task and functional movement screen items: a validation study. *The Journal of Strength & Conditioning Research*, 33(7):1855–1863, 2019.

[28] Guillaume Laffaye, Phillip P Wagner, and Tom IL Tombleson. Countermovement jump height: Gender and sport-specific differences in the force-time variables. *The Journal of Strength & Conditioning Research*, 28(4):1096–1105, 2014.

[29] Haojie Li, Jinhui Tang, Si Wu, Yongdong Zhang, and Shouxun Lin. Automatic detection and analysis of player action in moving background sports video sequences. *IEEE transactions on circuits and systems for video technology*, 20(3):351–364, 2009.

[30] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[31] Rainer Lienhart, Moritz Einfalt, and Dan Zecha. Mining automatically estimated poses from video recordings of top athletes. *International Journal of Computer Science in Sport*, 17(2):94–112, 2018.

[32] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019.

[33] Nicholas P Linthorne. Analysis of standing vertical jumps using a force platform. *American Journal of Physics*, 69(11):1198–1204, 2001.

[34] Marcel Lopes Dos Santos, Melissa Uftring, Cody A Stahl, Robert G Lockie, Brent Alvar, J Bryan Mann, and J Jay Dawes. Stress in academic and athletic performance in collegiate athletes: A narrative review of sources and monitoring strategies. *Frontiers in Sports and Active Living*, 2:42, 2020.

[35] Marc Madruga-Parera, Chris Bishop, Azahara Fort-Vanmeerhaeghe, Maria R Beltran-Valls, Oliver G Skok, and Daniel Romero-Rodríguez. Interlimb asymmetries in youth tennis players: Relationships with performance. *The Journal of Strength & Conditioning Research*, 34(10):2815–2823, 2020.

[36] William McNally, Kanav Vats, Alexander Wong, and John McPhee. Evopose2d: Pushing the boundaries of 2d human pose estimation using neuroevolution, 2020.

[37] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017.

[38] Darin A Padua, Michelle C Boling, Lindsay J DiStefano, James A Onate, Anthony I Beutler, and Stephen W Marshall. Reliability of the landing error scoring system-real time, a clinical assessment tool of jump-landing biomechanics. *Journal of sport rehabilitation*, 20(2):145–156, 2011.

[39] Darin A Padua, Lindsay J DiStefano, Anthony I Beutler, Sarah J De La Motte, Michael J DiStefano, and Steven W Marshall. The landing error scoring system as a screening tool for an anterior cruciate ligament injury–prevention program in elite-youth soccer athletes. *Journal of athletic training*, 50(6):589–595, 2015.

[40] Darin A Padua, Stephen W Marshall, Michelle C Boling, Charles A Thigpen, William E Garrett Jr, and Anthony I Beutler. The landing error scoring system (less) is a valid and reliable clinical assessment tool of jump-landing biomechanics: the jump-acl study. *The American journal of sports medicine*, 37(10):1996–2002, 2009.

[41] Georgios Papaiakovou. Kinematic and kinetic differences in the execution of vertical jumps between people with good and poor ankle joint dorsiflexion. *Journal of sports sciences*, 31(16):1789–1796, 2013.

[42] Ricardo Peterson Silveira, Pro Stergiou, Felipe P Carpes, Flávio A de S Castro, Larry Katz, and Darren J Stefanyshyn. Validity of a portable force platform for assessing biomechanical parameters in three different tasks. *Sports biomechanics*, 16(2):177–186, 2017.

[43] Chris Richter, Enda King, Siobhan Strike, and Andrew Franklyn-Miller. Objective classification and scoring of movement deficiencies in patients with anterior cruciate ligament reconstruction. *PloS one*, 14(7):e0206024, 2019.

[44] Sanjay Saini, Nordin Zakaria, Dayang Rohaya Awang Rambli, and Suziah Sulaiman. Markerless human motion tracking using hierarchical multi-swarm cooperative particle swarm optimization. *PLoS One*, 10(5):e0127833, 2015.

[45] Sean Scantlebury, Kevin Till, Thomas Sawczuk, Padraic Phibbs, and Ben Jones. Navigating the complex pathway of youth athletic development: Challenges and solutions to managing the training load of youth team sport athletes. *Strength & Conditioning Journal*, 42(6):100–108, 2020.

[46] Randy J Schmitz, John C Cone, Amanda J Tritsch, Michele L Pye, Melissa M Montgomery, Robert A Henson, and Sandra J Shultz. Changes in drop-jump landing biomechanics during prolonged intermittent exercise. *Sports Health*, 6(2):128–135, 2014.

[47] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *ACM International Conference on Multimedia*, Mountain View, CA USA, October 2017.

[48] Helen C Smith, Robert J Johnson, Sandra J Shultz, Timothy Tourville, Leigh Ann Holterman, James Slauterbeck, Pamela M Vacek, and Bruce D Beynnon. A prospective evaluation of the landing error scoring system (less) as a screening tool for anterior cruciate ligament injury risk. *The American journal of sports medicine*, 40(3):521–526, 2012.

[49] Turker Tuncer, Fatih Ertam, Sengul Dogan, Emrah Aydemir, and Paweł Pławiak. Ensemble residual network-based gender and activity recognition method with signals. *The Journal of Supercomputing*, 76(3):2119–2138, 2020.