HAW HAMBURG

BACHELOR THESIS

From Pose Estimation to Coaching Feedback: Rule-Based Analysis of Amateur Tennis Serves

vorgelegt am 29/09/2026

Markus Epstein

Erstprüferin: Prof. Dr. Tessa Taefi
Zweitprüfer: Thorsten Wagner

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Department Medientechnik
Finkenau 35
22081 Hamburg

Zusammenfassung

Diese Arbeit untersucht, inwieweit sich kinematische Kriterien aus der biomechanischen Tennisliteratur in Regeln auf Basis monokularer Posenschätzung umsetzen lassen. Ziel ist Aufschlagsfeedback für Amateurspieler, das ohne Laboraufnahme auskommt. Das Video einer einzelnen Kamera wird mit MediaPipe BlazePose verarbeitet. Anschließend werden vier Kriterien mit Referenzbereichen aus der Literatur verglichen. Rumpfneigung und Beugung des vorderen Knies werden in der Trophy-Position gemessen, Ellenbogenbeugung und Schulterhebung beim Ballkontakt. Die Arbeit bewertet jedes Kriterium nach vier Fehlerquellen: Landmarkenrauschen, Ereigniserkennung, Projektion und abweichende Winkeldefinitionen. Der zentrale Beitrag ist eine Einstufung, die für jedes Kriterium zeigt, wie weit man den Hinweisen einer einzelnen Kamera trauen kann. Mangels externer Referenz ist die Einstufung argumentativ begründet, nicht gemessen. Die Validität hängt vom Kriterium ab und ist abgestuft. Das vordere Knie liefert den verlässlichsten Hinweis. Bei der Schulterhebung sagt der Hinweis mehr über die Messung aus als über den Spieler. Die Pipeline eignet sich damit als Hinweissystem, das die Kameraebene berücksichtigt. Eine Kamera erfasst nur eines der beiden Trophy-Kriterien, für beide braucht es eine zweite. Der entscheidende nächste Schritt ist ein Abgleich mit einer dreidimensionalen Referenz.

Abstract

This thesis asks to what extent kinematic criteria from tennis biomechanics literature can be turned into rules over monocular pose estimation. The goal is serve feedback for amateur players without laboratory capture. Single-camera video is processed with MediaPipe BlazePose. Four criteria are then compared against reference bands from the literature. Trunk inclination and front knee flexion are read at the trophy position. Elbow flexion and shoulder elevation are read at ball impact. The thesis grades each criterion by four error sources: landmark noise, event detection, projection and definitional mismatch. The main contribution is a per-criterion grading of how far single-camera flags can be trusted. With no external reference, the grading is argued rather than tested. Validity varies by criterion and by degree. Front knee flexion gives the most trustworthy flag. The shoulder elevation flag reflects the measurement more than the player. The pipeline works as a plane-aware attention flag. One camera carries only one of the two trophy criteria, so reading both takes a second camera. The decisive next step is a test against a three-dimensional reference.

Contents

List of Figures V

List of Tables VI

1 Introduction 1

2 State of the Art 3
- 2.1 Human Pose Estimation . . . 3
  - 2.1.1 From Marker-Based to Markerless Systems . . . 3
  - 2.1.2 MediaPipe BlazePose . . . 3
  - 2.1.3 Model Variants and Operating Conditions . . . 5
  - 2.1.4 Limitations in Sports Contexts . . . 6
- 2.2 Biomechanics of the Tennis Serve . . . 7
  - 2.2.1 The Serve as a Phased Motion . . . 7
  - 2.2.2 Key Kinematic Features at the Serve . . . 8
- 2.3 Automated Serve Evaluation Systems . . . 9
- 2.4 Research Gap and Investigatory Questions . . . 10

3 Methodology 11
- 3.1 Research Design . . . 11
- 3.2 Data and Recording Setup . . . 11
- 3.3 Pose Estimation . . . 13
  - 3.3.1 Monocular 2D Operating Point . . . 13
  - 3.3.2 BlazePose Heavy Variant . . . 13
- 3.4 Rule Base . . . 14
  - 3.4.1 Criteria Selection . . . 14
  - 3.4.2 Reference Values and Rule Semantics . . . 16
- 3.5 Processing Pipeline . . . 17
  - 3.5.1 Pose Extraction . . . 18
  - 3.5.2 Landmark Preprocessing . . . 20
  - 3.5.3 Key-Event Detection . . . 21
  - 3.5.4 Angle Computation . . . 22
  - 3.5.5 Rule Evaluation and Indicator Generation . . . 23

3.6 Design of the Feasibility Assessment . . . 23
3.6.1 Error Budget . . . 23
3.6.2 Projection and Noise Propagation . . . 24
3.6.3 The Decidability Criterion and Event-Detection Stability . . . 25
3.7 Methodological Limitations . . . 26

4 Results **28**
4.1 Sample Composition and Criterion Availability . . . 28
4.2 Key-Event Detection Accuracy . . . 29
4.3 Projection and Noise Sensitivity . . . 30
4.4 Decidability Under Projection and Noise . . . 32
4.5 Empirical Indicators on the Amateur Corpus . . . 33
4.6 Example on a Single Clip . . . 34

5 Discussion **35**
5.1 Criterion Availability Under a Monocular Setup . . . 35
5.2 Stability of the Event Proxies . . . 36
5.3 Stability Under Noise and Viewpoint . . . 37
5.4 Flag Rates and Their Separate Causes . . . 38
5.5 Systematic Limits on Indicator Validity . . . 40

6 Conclusion **42**

Bibliography **44**

Appendix **46**

List of Figures

2.1 BlazePose 33-landmark topology 4

3.1 Projection of an inclination angle 16
3.2 Processing pipeline 19

4.1 Monocular projection curves across camera angles 30
4.2 Landmark noise propagation across camera angles 31
4.3 Decidability map across parameter space 32
4.4 Empirical angle distributions against reference bands 33
4.5 Qualitative pipeline execution on a representative serve clip 34

List of Tables

2.1 BlazePose variants: model size, inference speed and accuracy (Bazarevsky et al., 2021). 5
2.2 Kinematic reference values at the three serve key points (Jacquier-Bret & Gorce, 2024a). 9

3.1 Outcome of the three-condition criteria selection. 17
3.2 The four rules of the rule base. 18

4.1 Criterion availability across the 52 clips. 29
4.2 Key-event detection performance against the manual reading across the 52 clips. 29
4.3 Empirical joint angle distributions and indicator outcomes on the amateur corpus. 34

1 Introduction

Automated movement analysis in racket sports conventionally relies on body-worn sensors or on multi-camera systems that are expensive and complex to set up (Ebner & Findling, 2020; Santos-Gago et al., 2019). In recent years, markerless Human Pose Estimation (HPE) frameworks, such as Google MediaPipe, have emerged as a non-invasive and accessible alternative that extracts skeletal keypoints directly from standard video sequences (Aulton et al., 2025; Bazarevsky et al., 2020). In sports like tennis, analysing movement kinematics is particularly valuable, as specific biomechanical variations are associated with differences in playing level (Baily et al., 2020).

While action recognition and stroke classification are established applications of pose estimation in tennis, far less work has turned biomechanical movement characteristics into feedback a player can act on (Aulton et al., 2025). The tennis serve is a natural target for such work. Because it is self-initiated and self-paced, it can be recorded under controlled and repeatable conditions. Its sequential kinetic chain, however, makes a full technical assessment demanding. A complete account would span joint angles, inter-segment timing and angular velocities across several execution phases (Kovacs & Ellenbecker, 2011). This kind of assessment is available in elite and research settings, but hard to reproduce for an amateur. The capture systems behind elite feedback stay costly and complex to deploy. The existing kinematic studies of the serve remain heterogeneous in methodology, sampled populations and reporting conventions, which complicates the derivation of consistent reference values (Jacquier-Bret & Gorce, 2024a).

This thesis builds a rule-based setup as its research instrument. It extracts kinematic features from monocular video using MediaPipe, which allows deployment without specialised hardware. Each rule encodes an explicit relationship between one kinematic feature and a corresponding reference value, so any flagged deviation can be traced back to a named joint angle at a defined phase. Where the literature gives a deviation a direction, it can also be mapped onto a specific instruction, such as more flexion of the front knee. The reference criteria are not taken from any single study but from a systematic review and meta-analysis of tennis serve kinematics that aggregates angle measurements across 27 studies at key points of the serve motion (Jacquier-Bret & Gorce, 2024a).

However reliable the references, a rule is only as good as the angle it reads. A single camera recovers each angle only through its projection onto the image plane. When the movement does not lie in the plane the camera faces, the angle in the image differs from the one the body actually forms. The pose estimator adds error of its own, since markerless motion capture retains substantial random error for segment-level quantities (Emmerson et al., 2026). Whether these errors stay small relative to the reference band is what decides whether a rule can be applied at all. Pose-estimation studies in sport mostly validate on datasets with technical metrics and rarely show whether a system is accurate enough for a practitioner to act on (Aulton et al., 2025). This question matters most in the amateur setting, because the camera is set up by eye rather than calibrated.

The central research question is to what extent kinematic criteria established in tennis biomechanics literature can be operationalised as rules over monocular pose estimation output and which systematic error sources constrain their validity as a basis for coaching feedback on tennis serves. Rather than answer this with a single accuracy figure, the thesis grades each criterion by the error sources that limit it, so the outcome is a per-criterion account of how far a monocular rule can be trusted.

2 State of the Art

2.1 Human Pose Estimation

HPE is a foundational computer vision task concerned with localising and tracking anatomical keypoints across video frames or static images. The development of convolutional neural networks and deep learning architectures has shifted the paradigm from handcrafted feature-based methods towards data-driven approaches that localise keypoints directly from raw image data. Depending on the architecture, the resulting joint configurations are mapped into either two-dimensional or three-dimensional coordinate spaces (Aulton et al., 2025).

2.1.1 From Marker-Based to Markerless Systems

Historically, capturing human movement with high fidelity required intrusive setups, such as optical systems with physical markers or body-worn inertial sensors. Although such systems remain a biomechanical benchmark, their equipment requirements and typical confinement to controlled settings limit their use for routine analysis in applied training contexts. Markerless computer vision alternatives address this by estimating body configurations non-invasively from consumer-grade video, trading a degree of precision for substantially increased accessibility (Aulton et al., 2025; Emmerson et al., 2026).

2.1.2 MediaPipe BlazePose

Earlier real-time HPE systems such as OpenPose established keypoint detection from monocular video, but were designed primarily for pose extraction rather than lightweight on-device deployment. MediaPipe BlazePose follows a different design emphasis, targeting real-time operation on commodity hardware while predicting 33 skeletal landmarks per frame at over 30 fps on a mobile phone (Bazarevsky et al., 2020).

The pipeline operates as a detector-tracker pair. A lightweight detector locates the person once and a tracker then takes over from there. For each frame, the tracker predicts the keypoint coordinates, a refined region of interest and a presence signal that reports whether

a person is still in view. This region is propagated to the next frame, which keeps the tracker running continuously and leaves the detector as a fallback, rerun only when the tracker reports that no person is present. To keep this first stage fast, the detector does not look for the whole body but for the face, which serves as a proxy for the person and assumes that the head stays visible throughout (Bazarevsky et al., 2020). The landmark model rests on two prediction principles that apply at different stages. During training, a heatmap and offset branch supervises a lightweight embedding. A regression encoder reads this embedding to predict the coordinates directly. The heatmap branch is then discarded before inference, so that only the regression path runs on the device (Bazarevsky et al., 2020). The resulting 33-landmark topology, shown in Figure 2.1, spans the whole body, from the trunk and shoulders across the elbows and wrists to the hips, knees and ankles. Its design therefore prioritises accessibility and broad landmark coverage over maximal joint-level accuracy.

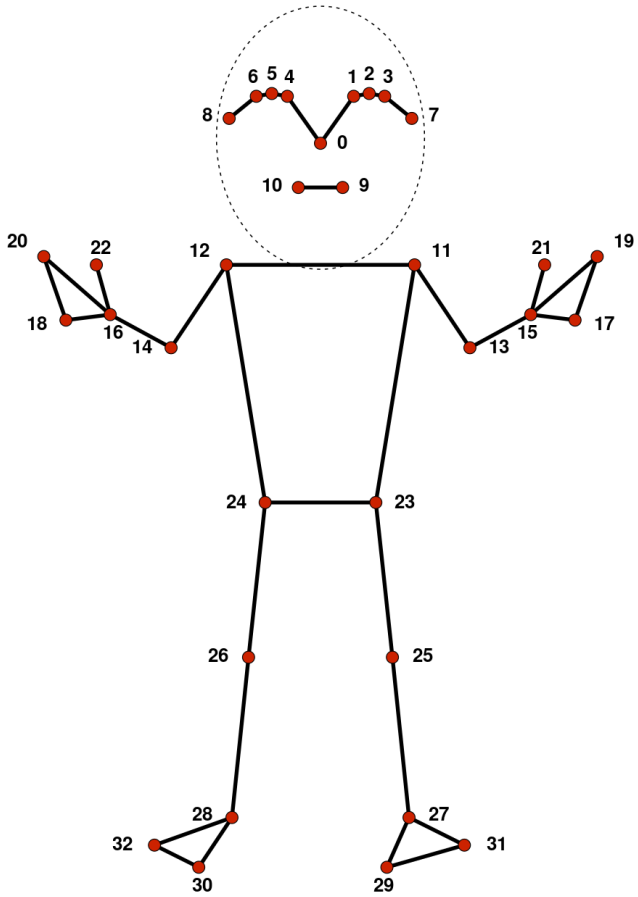


Figure 2.1: The 33-landmark topology predicted by BlazePose. Own illustration after (Bazarevsky et al., 2020).

2.1.3 Model Variants and Operating Conditions

The accuracy of BlazePose is reported as the percentage of detected joints (PDJ), also known as the percentage of correct keypoints at a tolerance of 20 % (PCK@0.2). A landmark counts as correct only when two conditions hold. The model must judge correctly whether the landmark is visible or occluded. The distance between the predicted and the annotated position must stay below 20 % of the projected torso diameter (Bazarevsky et al., 2020, 2021). The metric therefore records whether a landmark lands within a rough radius of its reference position, not how closely a joint angle derived from several such landmarks matches its true value.

BlazePose is distributed in three variants that differ in model capacity and trade accuracy against model size and inference speed, as summarised in Table 2.1. The PDJ values are averaged over 14 geographical subregions and measured against a human inter-annotator agreement of 97.5 PDJ (Bazarevsky et al., 2021). Higher accuracy comes at a consistent cost, since the heavy variant is roughly nine times larger than the lite variant and drops to about four frames per second on a CPU, a rate that bears on its suitability for continuous video.

Table 2.1: BlazePose variants: model size, inference speed and accuracy (Bazarevsky et al., 2021).

Variant	Size [MB]	FPS (CPU / GPU)	PDJ [range] [%]
Lite	3	44 / 49	87.0 [83.2–89.7]
Full	6	18 / 40	91.8 [89.2–94.0]
Heavy	26	4 / 19	94.2 [91.4–96.2]

Note. Frame rates measured on a Pixel 3 (released October 2018) via XNNPACK TFLite (CPU) and TFLite GPU. PDJ averaged over fourteen geographical subregions; the range gives the spread across those regions.

The reported figures rest on a validation set of 1400 smartphone photographs that the model card draws from the same source as the training images. Because this set stays close to the training distribution, the figures rank the three variants against one another but say nothing about the accuracy to be expected for movements outside that distribution. A second restriction, likewise stated in the model card, concerns what the figures actually measure. The z-coordinate returned by the model is not metric, since it is not observed but reconstructed by fitting a synthetic body model to the two-dimensional landmark projections. The developers accordingly evaluate accuracy on the two-dimensional coordinates alone, because the synthetic depth values cannot be compared against human annotation. The reported accuracy therefore characterises the projected landmark positions and offers no evidence for the quality of the depth estimate (Bazarevsky et al., 2021).

The same documentation fixes the conditions under which the model is meant to operate at all. Subjects further than approximately 4 m from the camera fall outside the intended range, the head has to remain visible for the face-based person detection described above to function, and only one person is tracked when several are present in the scene. Degraded lighting, noise and motion are expected to reduce landmark quality and to increase inter-frame prediction noise. These conditions apply to all three mentioned variants (Bazarevsky et al., 2021).

2.1.4 Limitations in Sports Contexts

Despite their accessibility, markerless systems carry measurement limitations that are particularly relevant for biomechanical analysis. Human pose estimation does not track fixed points the way marker-based motion capture does. It estimates joint locations from visual data, and this limits its accuracy (Aulton et al., 2025). A recurring source of error is that individual joints are frequently not visible in a two-dimensional image, which pose estimators address only indirectly through per-keypoint visibility scores rather than through actual observation (Baily et al., 2020). The landmarks also differ in kind from the joint centres of marker-based analysis. Many pose estimators are trained on annotations that are not anatomically accurate. They represent the whole torso by the hip and shoulder keypoints alone (Emmerson et al., 2026).

The consequence is most visible at the level of individual segments. A study on tennis strokes including serves quantifies this by evaluating a markerless system against two criterion methods, marker-based motion capture and ground reaction forces. For whole-body external mechanical work, agreement with these criteria was strong, with limits of agreement of roughly 6 to 10 %. At the finer segment level, agreement deteriorated. High-frequency artifacts in the markerless kinematics produced substantially higher random errors, with limits of agreement of about 8 to 13 % for HRNet and 12 to 21 % for OpenPose, and large differences between individuals (Emmerson et al., 2026). This gap appeared even though the evaluation used a multi-camera three-dimensional configuration with established pose estimators, a more favourable measurement case than a single monocular camera.

A second limitation concerns robustness across recording conditions. Tracking quality in sport depends on the recording and on the athlete. Occluded body parts, camera resolution, clothing and unusual individual poses can all degrade it (Aulton et al., 2025). The skill level of the performer may matter too, since it shapes how fast and how far the body moves. In elite sprinting, for example, occlusions caused by fast hand movements have been reported to lower accuracy (Aulton et al., 2025). The evidence base adds to this uncertainty. Most studies train and test their models on private data, so their accuracy figures are hard to reproduce and may not hold under other recording conditions (Aulton et al., 2025).

A third limitation concerns the viewpoint. Two-dimensional video recovers a joint angle validly only when the movement lies in the plane the camera faces. Validity is low for angles outside that plane and a review of two-dimensional running analysis advises particular caution for the trunk, pelvis and hip (Leporace et al., 2023). How well a serve angle can be read therefore depends not only on the estimator but on how the recording is aligned to the plane of that angle.

2.2 Biomechanics of the Tennis Serve

2.2.1 The Serve as a Phased Motion

The serve is a full-body movement in which segmental rotations transfer force from the ground through the kinetic chain and into the ball, so that the outcome depends on the coordination of the whole sequence rather than on any single joint (Kovacs & Ellenbecker, 2011). To make this continuous motion tractable for analysis, Kovacs and Ellenbecker describe it through an eight-stage model grouped into three phases. The preparation phase covers the start, the ball release, the loading of the lower body and the cocking of the racket arm. It stores the potential energy that the later phases release. The acceleration phase covers the upward drive and the contact with the ball. The follow-through phase covers the deceleration of the arm and the finish at landing (Kovacs & Ellenbecker, 2011). The phases thus reflect three distinct functions, storing energy, releasing it and absorbing it after contact.

This staged description explains why serve kinematics are read at a small number of discrete instants rather than continuously. Jacquier-Bret and Gorce anchor their analysis to three such instants (Jacquier-Bret & Gorce, 2024a). The trophy position corresponds to the loading stage of the model and is defined equivalently as the first peak of vertical racket movement and as the moment of lowest vertical elbow position and maximum knee flexion in a fully loaded lower body. The racket low point marks the instant of maximum external shoulder rotation with the racket tip pointing towards the ground and falls at the transition from cocking into acceleration. Ball impact marks the moment of contact between racket and ball and corresponds to the contact stage. The reference values read at these instants, reported in the following section, are in all cases angles between body segments rather than properties of the racket. The racket enters only the definition of the instants and not the quantities measured at them.

2.2.2 Key Kinematic Features at the Serve

The systematic review and meta-analysis of Jacquier-Bret and Gorce consolidates the kinematic values reported for the three serve key points across 27 studies (Jacquier-Bret & Gorce, 2024a). Those studies do not always describe the serve in the same terms. They differ in terminology, rotation sequences and reference frames. So before pooling, the authors bring all data to the conventions of the International Society of Biomechanics (ISB). Where a study reports an intersegmental angle, they recompute its complementary or supplementary angle. Where its axes of rotation depart from the ISB definition, they decompose the movement again.

Jacquier-Bret and Gorce judge each pooled value by its dispersion and treat a value below 20° as usable in training (Jacquier-Bret & Gorce, 2024a). The review gives no definition of this measure. Only trunk inclination and front knee flexion at the trophy position fall below the threshold across all studies. For the other parameters the dispersion lies between 20° and 50°. The authors trace this to one or two studies with a strong effect on the dispersion. Shoulder elevation and elbow flexion at ball impact reach the threshold only after two such studies are removed. That removal, which the authors present only in their discussion as a conditional adjustment rather than as a result of the meta-analysis, shifts the pooled shoulder elevation from 110.7° ± 16.9° to 104.6° ± 6.1° and elbow flexion from 30.1° ± 15.9° to 29.2° ± 9.9° (Jacquier-Bret & Gorce, 2024a). Shoulder external rotation at the racket low point stays outside this set, with the widest pooled standard deviation of all reported parameters and no comparable exclusion offered.

Table 2.2 brings together three quantities of different kinds. The all-studies column gives the mean over the included studies with its standard deviation (Jacquier-Bret & Gorce, 2024a). The review does not state how this deviation is computed. Since it is taken over study results, it most likely describes how the study means scatter rather than how individual players vary. The dispersion column gives the authors' verdict of whether a parameter stays below the 20° threshold they treat as usable and the excluded-deviation column reports the recomputed mean for the two parameters that reach that threshold only after the removal of the deviating studies (Jacquier-Bret & Gorce, 2024a).

Two further properties of the pooled values follow from the review. The first concerns the sample. Jacquier-Bret and Gorce exclude studies on beginners, so the means describe players who have already acquired the movement. The remaining heterogeneity stems from the populations and the test conditions. It spans mean sample ages from 9.3 to 34.4 years, levels from International Tennis Number (ITN)-rated juniors to professionals and flat, kick, slice and topspin serves performed with both foot-up and foot-back technique. The second concerns the direction of a deviation, which carries meaning for only one of the parameters, front knee flexion at the trophy position. Drawing on Hornestam et al., Jacquier-Bret and Gorce report a

Table 2.2: Kinematic reference values at the three serve key points (Jacquier-Bret & Gorce, 2024a).

Key point	Parameter	All studies [°]	Excl. dev. [°]	Disp. $< 20°$
Trophy position	Trunk inclination	25.0 ± 7.1	—	yes
Trophy position	Front knee flexion	64.5 ± 9.7	—	yes
Trophy position	Back knee flexion	67.4 ± 16.4	—	no
Racket low point	Shoulder external rotation	130.1 ± 26.5	—	no
Ball impact	Shoulder elevation	110.7 ± 16.9	104.6 ± 6.1	yes[a]
Ball impact	Elbow flexion	30.1 ± 15.9	29.2 ± 9.9	yes[a]

Note. The fourth column gives the value recomputed after removing the studies the authors identify as deviating without stated reason, Brocherie and Dinu for shoulder elevation and Wang et al. for elbow flexion. [a] reached only after this exclusion.

group flexing to $55.6° \pm 8.47°$, roughly 9° below the pooled mean, with a racket velocity lower by $3.33\,\mathrm{km\,h^{-1}}$ (Jacquier-Bret & Gorce, 2024a). A further study associates greater flexion with greater racket velocity and on this basis the authors recommend pronounced knee flexion at this key point (Jacquier-Bret & Gorce, 2024a). For the other parameters the evidence is weaker. At ball impact the review notes that shoulder elevation and low elbow flexion help to reach a high contact point and that a slightly flexed elbow seems to favour racket velocity. It reports no measured effect of a deviation for these parameters or for trunk inclination, so deviations in them carry no established direction.

2.3 Automated Serve Evaluation Systems

Automated analysis of the tennis serve has so far concentrated on extracting and validating kinematic quantities rather than on returning technique feedback to the player. The closest existing work is the markerless evaluation of tennis strokes (Emmerson et al., 2026). Its multi-camera laboratory setup recovers geometry that a single consumer camera cannot. More broadly, markerless pose estimation is increasingly applied across sports for movement measurement (Aulton et al., 2025), and pose-based stroke analysis has been demonstrated on a mix of online and self-recorded video (Baily et al., 2020). These systems show that usable kinematic estimates can be obtained outside the laboratory, but they stop at describing movement and do not translate it into criterion-referenced statements about technique.

In principle, two routes exist for turning such kinematic estimates into an assessment of technique. A learning-based route infers quality directly from labelled examples and a rule-based route compares measured quantities against reference values drawn from the literature.

Which of these routes is viable given the data available for tennis serves, and how it is operationalised, is a question of system design taken up in Chapter 3.

What the existing systems do not provide is an interpretable, criterion-referenced assessment of serve technique obtained from a single consumer camera. Whether such an assessment is attainable remains open, because the accessible monocular setting is precisely the one in which the measurement reliability examined in Section 2.1.4 is weakest.

2.4 Research Gap and Investigatory Questions

HPE makes serve kinematics accessible from a single consumer camera, but its reliability falls off for the segment-level joint angles read from individual frames and its robustness across realistic serve-recording conditions is not established (Section 2.1.4). The biomechanics literature supplies reference values for the serve, yet only a subset of them show a dispersion narrow enough to serve as a stable reference, with trunk inclination and front knee flexion clearing the threshold across all studies and the two ball-impact parameters only after a post-hoc exclusion (Section 2.2.2). Whether these values can in addition be recovered from point landmarks under a single camera plane is a matter of system design and is examined in Chapter 3. Existing automated systems extract and validate kinematics but do not return criterion-referenced technique feedback. The accessible monocular setting in which such feedback would be useful is the one in which measurement is least certain (Section 2.3). What remains open is therefore not whether reference values exist or whether pose can be estimated, but whether the established criteria can be recovered from monocular 2D video reliably enough to support statements about technical deviation.

This gap motivates the research question, namely to what extent the kinematic criteria established in the tennis biomechanics literature can be operationalised as rules over monocular pose estimation output and which systematic error sources constrain their validity as a basis for coaching feedback on tennis serves. Because the underlying measurement is itself uncertain, the question is treated as a feasibility analysis and it is examined through the following three investigatory questions.

1. Which of the kinematic serve-technique criteria established in the literature can be extracted from monocular 2D pose estimation under the measurement and camera-plane constraints?
2. How stable are the extracted criteria across the recording conditions of amateur video, in particular under landmark noise, camera viewpoint and the automatic detection of the key points such as the trophy position?
3. To what extent do the extracted criteria support interpretable deviation indicators for coaching and under which conditions do the measurement limits render such indicators unreliable?

3 Methodology

3.1 Research Design

This thesis is a feasibility analysis of a rule-based pipeline that applies kinematic serve criteria to the output of monocular pose estimation. It reads each criterion from estimated body landmarks and compares it against reference values from the biomechanics literature. The work is exploratory, because the monocular measurement of joint angles that every later step rests on is itself of uncertain reliability. Its purpose is to locate where along the pipeline the criteria still hold and where the measurement makes them unrecoverable. The object of the analysis is therefore the pipeline as a whole rather than the raw accuracy of the pose estimator. The estimator is judged as it would be used in practice, with its landmark output first cleaned by the preprocessing and then read by the rules.

Each serve is examined once, at the instants that define the key events of the motion. The quantities read there are individual joint angles rather than aggregate scores. Neither a second observer nor an annotated reference dataset is available and the sample is small. Therefore, the analysis maps what the pipeline recovers and loses. It claims no statistical generalisation and passes no judgment on any individual player. Accordingly its checks are all internal. The steps that need verifying, above all the automatic detection of the key events, are held against the recorded frames themselves rather than an external ground truth.

The assessment follows the pipeline it examines. The recording setup, the pose estimation, the rules and the path to a deviation indicator are fixed first. Only then does the assessment judge what the pipeline recovers, organised around the three investigatory questions raised before.

3.2 Data and Recording Setup

The analysis draws on two sources of serve recordings, both restricted to the recreational and club players the research question concerns rather than tour professionals or elite juniors. The first is a single amateur serve recorded for this study on a consumer smartphone, which reproduces the target case. The second is publicly available online footage of amateur players

serving. This footage is retrieved from public tennis community feeds such as `r/10s` on Reddit with targeted queries for serve-specific terminology such as trophy pose, serve or similar. The admitted corpus holds 52 clips. One is self-recorded. The other 51 are online footage, 45 from Reddit and 6 from YouTube.

Both sources are held to the operating range the pose estimator depends on. Each serve shows a single player within roughly four metres of a static camera, with the head visible, no other person in the foreground and lighting sufficient to limit landmark noise. The self-recorded serve meets this range by control at the point of filming. As the online footage cannot be controlled at recording, it is brought into the same range by selection.

A clip is admitted only when it also frames the whole body at a size that resolves the limbs, holds the camera still through the serve and shows one complete serve from the loading phase to contact without a cut. The joints each criterion needs must be clear at the trophy position and at ball impact. A technical floor requires a minimum resolution of 720p and a clip duration of at least 1.5 s. Slow-motion recordings at 60, 120, or 240 fps are preferred, since ball impact is a brief instant that a higher frame rate localises more precisely. Clips with heavy motion blur or compression at the relevant joints are discarded.

The admitted online clips are normalised with FFmpeg into standardised MP4 containers with H.264 encoding and their native slow-motion frame rates, where present, are preserved without downsampling. The audio track is stripped for privacy. Dataset provenance, trim timestamps and the per-clip annotation are exported to CSV and JSON, so a run can be traced and reproduced.

Each admitted clip is annotated by viewpoint. A frontal or rear view faces the frontal plane and carries the trunk inclination. A sagittal view faces the sagittal plane and carries the front knee flexion. An oblique view is discarded, since it can serve no criterion. Each clip is further annotated with the dominant serving arm, the front leg and the viewing direction, which fix the side every later angle is read from. The residual variation this screening cannot remove, above all the differing frame rates across sources, is carried to the methodological limitations (Section 3.7).

Each admitted clip is finally trimmed to begin at the toss and after any pre-serve ball bouncing. The trophy detector takes the lowest pelvis position before ball impact. A bounce crouch can push the pelvis below the loading posture and pull the detected frame outside the serve. Trimming to the toss onset keeps the pelvis minimum within the loading phase it marks.

3.3 Pose Estimation

3.3.1 Monocular 2D Operating Point

The rules read joint angles from the two image-plane coordinates of each landmark and leave the estimated depth aside. BlazePose returns a third coordinate for every landmark, but it is not a metric depth measurement. It is recovered by fitting a synthetic body model to the two-dimensional landmark projections and is defined only up to scale, so it adds no independent observation of how far a joint lies from the camera (Bazarevsky et al., 2021). The reported accuracy points the same way, because the BlazePose model is evaluated against human annotation on the projected positions alone and no comparable figure is given for the depth estimate. An angle built from that coordinate would therefore rest in part on a value whose error the developers do not report. First evaluations point to large depth errors for MediaPipe on single-camera video (Bini et al., 2026). Restricting the rules to the two image-plane coordinates keeps each measurement within what the accuracy figures actually validate.

This choice comes with a consequence. An angle recovered in the image plane is a projection of the true spatial angle, so its value depends on the position of the camera relative to the plane of the motion. That dependence decides which criteria a given viewpoint can support and is taken up where the rule base is selected (Section 3.4.1).

3.3.2 BlazePose Heavy Variant

The BlazePose estimator fits the recording setup on four counts. It runs from a single RGB camera and needs no depth sensor, so it matches the monocular footage. It returns every joint the four serve angles are built from, so no criterion is lost to a missing landmark. It attaches a per-landmark visibility score, which is the signal the preprocessing gating reads to decide what counts as trustworthy. Finally, it is openly available, so the setup can be reproduced. The estimator comes in three variants that trade landmark accuracy against model size and inference speed (Table 2.1) and this setup uses the heavy variant.

The recordings are analysed from stored video rather than live on the phone, so the pipeline faces no real-time or on-device constraint and the speed advantage of the lighter variants carries no weight. The heavy variant leads with the highest PDJ across the evaluated regions, at a spread no wider than the other variants (Bazarevsky et al., 2021). That consistency matters, because each rule reads its angle at a single key frame, where the landmark quality sets the angle quality. The choice therefore reduces to accuracy.

The PDJ figures rank the variants on images close to the model's training distribution (Section 2.1.3). They also date from the 2021 model card. The current Pose Landmarker documentation still links this card for all three variants and publishes no newer figures (Google AI Edge, 2026b), so it remains the best available description of the models MediaPipe loads. The figures justify the choice between variants but do not predict accuracy on the high-velocity serve. How far the estimates hold under serve conditions is examined in the feasibility assessment (Section 3.6).

3.4 Rule Base

The reference values retained in the State of the Art (Section 2.2.2) enter this section as fixed inputs. What follows turns them into rules over the pose-estimation output and sets out the conditions each rule needs to be read from a single camera.

3.4.1 Criteria Selection

A parameter becomes a rule only if it clears three conditions. Its reference must be stable enough to mark a target, which is a property of the data. It must be computable from point landmarks, which is a property of the setup. Finally, if it is read in a fixed plane, that plane must be one a single camera can face, which is a property of the geometry. Each condition rules out a different kind of error and only a parameter that clears all three enters the rule base.

The first condition inherits the dispersion screening carried out in the State of the Art (Section 2.2.2). Trunk inclination and front knee flexion at the trophy position stay below the dispersion threshold across all studies and form the stable core. Shoulder elevation and elbow flexion at ball impact reach it only after the post-hoc exclusion and are carried forward as conditional criteria. Back knee flexion and shoulder external rotation both remain too dispersed to mark a target. Back knee flexion is set aside on this ground alone. Shoulder external rotation fails the second condition as well and is taken up there (Jacquier-Bret & Gorce, 2024a).

The second condition asks what the landmarks can express. A point landmark fixes an estimated joint position and positions fix the direction of every limb segment, hence any angle between two segments. Three of the four remaining criteria are exactly such angles. The knee lies between thigh and shank, the elbow between upper arm and forearm and the shoulder elevation between trunk and upper arm. The fourth, trunk inclination, is the angle of the hip-to-shoulder line to the vertical. Each one can be built from what the estimator returns. What positions do not express directly is the rotation of a segment about its own

long axis. Even from three-dimensional keypoints, a segment keeps only five of its six degrees of freedom. Its axial twist can be inferred only through the neighbouring segment and not at all once the elbow nears full extension (Emmerson et al., 2026). Shoulder external rotation is such a twist of the upper arm. At the racket low point the elbow is flexed, so it could be inferred in principle, but only from the three-dimensional orientation of the arm. The two-dimensional operating point discards this (Section 3.3.1). It is therefore not recoverable in this setup, which adds a second ground to its dispersion. The same limit removes the instant it marks. The racket low point ends the cocking stage, which the phase model fixes by maximum external rotation and racket orientation (Kovacs & Ellenbecker, 2011) and this setup observes neither. None of its published definitions can be applied to the landmarks (Jacquier-Bret & Gorce, 2024a), so excluding it leaves the loading and contact stages the setup addresses.

Two of the retained criteria pass the second condition only approximately. Trunk inclination is taken over the hip-to-shoulder axis rather than the trunk segment itself. Pose models do not track the spine and largely treat it as one rigid segment between the pelvic and shoulder girdles (Bini et al., 2026). The resulting definitional offset cannot be measured without a joint-centre ground truth. Elbow flexion is read at ball impact with the arm near full extension. Shoulder, elbow and wrist then lie almost on one line, so a small landmark error shifts the angle by a large share of its value. How far this holds is quantified in the feasibility assessment (Section 3.6).

The third condition turns the two-dimensional operating point into a requirement on the recording. Because an angle is observed only through its projection, a criterion is measurable only from a viewpoint that faces its plane. For a distant camera, perspective is approximated by scaled orthographic projection, whose uniform scale cancels in any angle. The approximation is close when the depth relief of the body is small against its distance from the camera and the player stays near the image centre (Hartley & Zisserman, 2004). The camera is taken as level and turned about the vertical axis, so that its image plane meets the plane of the motion at an angle θ (Figure 3.1). A segment moving within that plane then keeps its vertical component, while its horizontal component is scaled by $\cos\theta$. Since the tangent of an inclination is the ratio of the horizontal to the vertical component, the projected angle follows as

$$\tan\alpha_{\text{proj}} = \tan\alpha_{\text{true}}\cos\theta. \tag{3.1}$$

The relation is exact in this model. Under a real lens it is an approximation, since perspective shifts each landmark by an amount that depends on its depth and can move an angle either way. At the distances of at most about four metres the recordings admit, the depth relief of a serving arm is not small, so the model indicates the size of the projection effect rather than fixing it exactly.

The two core criteria occupy orthogonal planes, front knee flexion in the sagittal plane and trunk inclination in the frontal plane. A single camera aligned to one foreshortens the

other. Example on a true trunk inclination at its reference mean of 25.0°. By Equation 3.1, it reads below 17.9° once the camera sits about 46° off the frontal plane. With the camera facing the sagittal plane, it reads 0°, so the lean vanishes from the image entirely. A review of two-dimensional running analysis points the same way. There, validity against a three-dimensional reference is highest for sagittal knee and ankle angles and low in the frontal and transverse planes. The authors name out-of-plane movement under a single camera as one likely cause (Leporace et al., 2023). The condition is therefore a measurable tolerance on camera placement rather than a caution and within a single recording at most one of the two core criteria can be read cleanly.

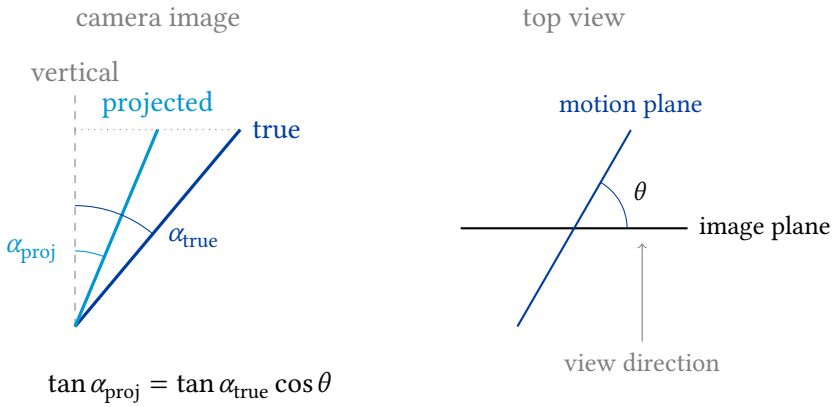


Figure 3.1: Foreshortening of an in-plane inclination under a monocular view. Left: a segment as a camera facing the motion plane sees it (true) and as it appears once the motion plane is turned by θ (projected). The vertical extent is kept and the horizontal extent shrinks by $\cos \theta$, so the inclination reads smaller, following Equation 3.1. Right: the same rotation seen from above. Own illustration.

The rule base that clears all three conditions is therefore compact. Table 3.1 sets out the outcome. Trunk inclination and front knee flexion enter as core criteria at the trophy position, one per recording according to the camera plane. Elbow flexion and shoulder elevation enter as conditional criteria at ball impact. These two are not read in one fixed plane, so the third condition places no alignment requirement on them. The table marks this with “any”. Back knee flexion and shoulder external rotation are excluded. This set is the scope that the remaining sections operationalise.

3.4.2 Reference Values and Rule Semantics

Turning a retained value into a rule fixes two things, what the value is a target for and what a departure from it may signal. The target is the mean. The pooled studies exclude beginners and describe players who have already acquired the movement, so their average marks a

Table 3.1: Outcome of the three-condition criteria selection.

Parameter	Key point	Disp. $< 20°$	From landmarks	Plane	Status
Trunk inclination	Trophy position	yes	yes	frontal	core
Front knee flexion	Trophy position	yes	yes	sagittal	core
Elbow flexion	Ball impact	after excl.	yes	any	conditional
Shoulder elevation	Ball impact	after excl.	yes	any	conditional
Back knee flexion	Trophy position	no	yes	sagittal	excluded
Shoulder external rotation	Racket low point	no	no (axial)	—	excluded

Note. A parameter enters the rule base only if it clears all three conditions. Core criteria clear the dispersion threshold across all studies, conditional criteria only after the post-hoc exclusion of deviating studies. A criterion that is not tied to a fixed plane is marked any and imposes no admission constraint on camera placement. Shoulder external rotation is excluded twice over, on dispersion and because an axial rotation is not recoverable from two-dimensional landmarks.

technique worth aiming at (Section 2.2.2). A departure from the mean is read as an indicator rather than a verdict. It flags a criterion that warrants attention, not a serve declared faulty.

Each band spans one standard deviation on either side of the mean. The review reports no other spread for its pooled values, so the band stays tied to the source instead of to an arbitrary cut-off. A factor of one also adds no free multiplier. The band remains an operationalisation, not a claim that sound technique ends at its edge. The pooled standard deviation describes how the study results spread, not how far a sound serve may stray before it becomes faulty.

The band is applied to both sides unless the evidence points in a direction. For trunk inclination, elbow flexion and shoulder elevation no measured direction of deviation is reported, so a departure on either side is flagged alike. Front knee flexion is the single exception. Greater flexion is associated with greater racket velocity and the review recommends pronounced flexion at this instant (Hornestam et al., 2021; Jacquier-Bret & Gorce, 2024a), so only insufficient flexion carries a technical reading. Its rule therefore uses a one-sided lower bound at the mean minus one standard deviation, which flags too little flexion. Both conditional criteria take their recomputed values, since only these clear the dispersion threshold (Section 3.4.1). For the elbow this also narrows the band from about 14°–46° to 19°–39°.

3.5 Processing Pipeline

The processing pipeline is the chain that turns a single recording into the deviation indicators. It runs in five stages. The recording is reduced to landmark trajectories, the trajectories are conditioned, the two key events are located within them, the rule angles are computed at

Table 3.2: The four rules of the rule base.

Rule	Key point	Reference [°]	Flag when [°]	Landmarks
Trunk inclination	Trophy position	25.0 ± 7.1	outside 17.9–32.1	hip–shoulder vs. vertical
Front knee flexion	Trophy position	64.5 ± 9.7	below 54.8	hip–knee–ankle
Elbow flexion	Ball impact	29.2 ± 9.9[a]	outside 19.3–39.1	shoulder–elbow–wrist
Shoulder elevation	Ball impact	104.6 ± 6.1[a]	outside 98.5–110.7	elbow–shoulder–hip

Note. Reference values from (Jacquier-Bret & Gorce, 2024a). Each flag range is the reference mean plus or minus one standard deviation. Applied symmetrically except for front knee flexion, where directional evidence supports a one-sided lower bound. [a] value recomputed after the post-hoc exclusion of deviating studies. Angles are computed from the listed landmarks as defined in Section 3.5.4.

those instants and each angle is compared against its band to yield an indicator. Figure 3.2 shows this chain. The stages are described in turn.

3.5.1 Pose Extraction

Each recording is decoded frame by frame with OpenCV and passed to the heavy Pose Landmarker through the MediaPipe Tasks API.[1] The extractor operates in video mode and tracks a single pose. The detection, tracking and presence thresholds are left at their default of 0.5 and the segmentation output is disabled (Google AI Edge, 2026c). In video mode the extractor derives the tracking region from the previous frame's landmarks and runs the detector only when no pose is tracked (Google AI Edge, 2026a). The landmark coordinates pass through a One Euro filter and the visibility scores pass through an exponential low-pass (Google AI Edge, 2026a). Both filters are switched on by default by the video mode. The preprocessing therefore receives landmarks that are already filtered and its own low-pass runs on top of the extractor's.

Frames without a detected pose are marked as such and still carried, so the series keeps one entry per frame across all landmarks. Following the two-dimensional operating point, the pipeline keeps each landmark's two image-plane coordinates and its visibility score and discards the depth and world coordinates. The serve becomes a set of two-dimensional landmark trajectories and the retained visibility supplies the gating in the preprocessing.

[1]The pipeline is implemented on a 2021 MacBook M1 Pro in Python with MediaPipe 0.10.35, OpenCV 5.0.0.93, NumPy 2.2.6, SciPy 1.15.3 and Matplotlib 3.10.9 for the figures. The versions are pinned so a run reproduces exactly.

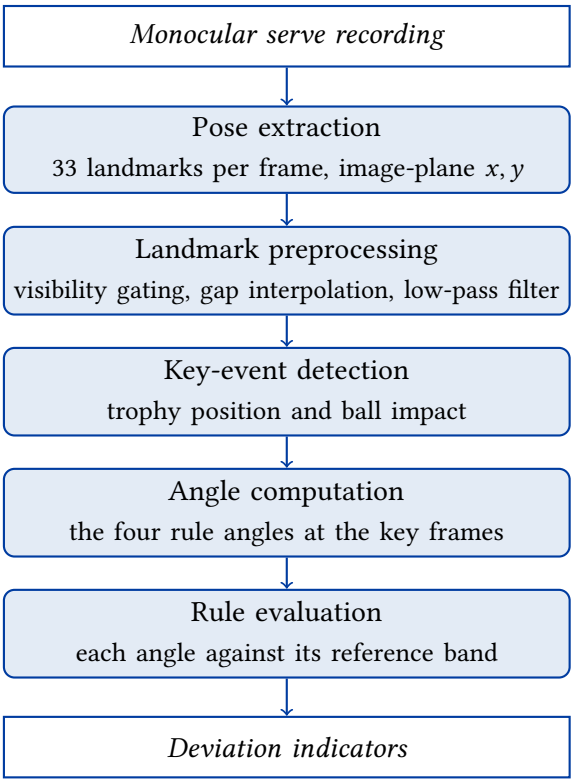


Figure 3.2: The processing pipeline from a single recording to the deviation indicators. Own illustration.

3.5.2 Landmark Preprocessing

The extraction returns one landmark estimate per frame and each rule reads its angle at a single key frame. Two properties of that per-frame estimate would therefore pass into the angle unchanged. A landmark drops out where the joint is occluded, which the estimator reports as a low visibility score. Where the joint is visible, its estimate still jitters slightly from frame to frame. The preprocessing conditions the trajectories against dropout and jitter in a simple sequence of visibility gating, interpolation of short gaps and low-pass smoothing. The parameters are set to conventional values rather than individually tuned. The steps are implemented with NumPy and SciPy.

Gating fixes what counts as trustworthy. At every frame a landmark is marked reliable when its visibility score reaches the threshold of 0.5 and unreliable below it. This matches the pose estimator's other default confidence thresholds (Google AI Edge, 2026c). The mark is stored next to the coordinate rather than the sample being discarded, so no data is removed and every rejection stays traceable to its cause: an undetected pose or a landmark below the threshold.

Short interruptions of a reliable trajectory are then bridged, so that the low-pass filter can run over a continuous signal and a briefly occluded landmark stays usable at a key frame. A gap is filled only when reliable samples enclose it on both sides and it stays within a set bound. Its missing samples are interpolated linearly between the two edges and count as reliable. Linear interpolation keeps the step simple. The bound is set at 120 ms and converted per clip into a frame count, so the same physical gap length applies from 30 to 240 frames per second. It lies within the gap lengths examined for gap filling in camera-based motion analysis, where interpolation error grows with how much of a local extremum the gap covers (Mohammadzadeh Gonabadi et al., 2022). The value is a practical setting rather than a measured occlusion length. A longer occlusion stays unreliable and a criterion that needs the landmark at the key frame is then not evaluated.

To remove frame-to-frame jitter, the gated and interpolated trajectories undergo a final smoothing. This matters especially for the following key-event detection, which reads its two events as extrema of a landmark coordinate. A jittery trajectory can settle the extremum on a false frame. Following standard practices for biomechanical motion data (Robertson & Dowling, 2003), the setup uses a second-order Butterworth low-pass filter. The cut-off frequency of 8 Hz is taken from a tennis serve study (Jacquier-Bret & Gorce, 2024b). Each clip uses a filter designed specifically for its own frame rate. It is applied forward and backward, so it adds no time shift and the key frames keep their alignment (Robertson & Dowling, 2003). The double pass also lowers the effective cut-off, in this setup to about 6.4 Hz to 7.0 Hz depending on the frame rate (Robertson & Dowling, 2003). Because the filter needs a

continuous signal, it runs over each unbroken stretch of reliable and interpolated samples separately.

3.5.3 Key-Event Detection

The rules read their angles at two instants, the trophy position and ball impact. The pipeline locates these two frames within the filtered trajectories. The serve literature names each instant through the racket (Jacquier-Bret & Gorce, 2024a). In this setup, the racket never enters the measured angle, so both events are located from body landmarks alone. The trophy position ends the loading stage of the serve. In the literature it is the first vertical peak of the racket over a fully loaded lower body at maximum knee flexion (Jacquier-Bret & Gorce, 2024a). Neither cue is robust as a landmark signal in both camera planes. The racket is not tracked and a camera facing the frontal plane foreshortens the knee angle. Loading is therefore captured through the vertical position of the pelvis, taken as the midpoint of the two hip landmarks. Its height survives in both planes, since the orthographic model behind Equation 3.1 preserves the vertical image component.

The trophy frame is the one at which the pelvis reaches its lowest point. Only frames before ball impact are searched, so the low pelvis of the landing is not taken for the loading. The lowest pelvis still only approximates maximum knee flexion. Its height also depends on the back leg and on the stance, whose width limits how deep the player can squat (Kovacs & Ellenbecker, 2011). Since the hip landmarks feed both the proxy and the two angles read at the trophy frame, an error in them would shift the frame and those angles together.

Ball impact has no such body-based definition. It is the moment of racket-ball contact, which the landmarks do not show. It is therefore located through a kinematic proxy, the frame at which the racket-arm wrist reaches its highest point. This frame is taken to coincide with the extended reach at contact. The serving arm is recorded as a per-clip parameter, so the wrist is taken from the racket arm. This is a deliberate choice over a racket or ball detector. Such a detector would add an independent error source and a tracking problem outside the monocular landmark setup.

The two key-event detectors run in sequence. Ball impact is found first, as the minimum of the racket-arm wrist y-coordinate over the clip. The trophy is then the maximum of the pelvis y-coordinate over the frames that go before it. Image coordinates grow downward, so the highest point is a minimum and the lowest point a maximum. Each event is an extremum of one coordinate series. They are computed with NumPy over the filtered trajectory.

Two conditions guard the detection. The first requires the event frame to fall on an originally reliable sample rather than an interpolated one, so no event is placed on a filled position. It also rejects an extremum beside an unfilled gap, where the true extremum may lie in

the unobserved gap. A filled gap is not covered. A linear fill holds no interior extremum, so a peak hidden in a filled gap is flattened onto the gap edge and the event lands there. This bears most on ball impact, where a gap of up to 120 ms can hide the brief contact peak. The resulting offset enters the event error the manual frame check measures. The second condition requires at least one frame between the trophy and ball impact, which rejects a wrist peak so early in the clip that no loading phase precedes it. The wrist height is also compared at both events. Because ball impact is the wrist's highest point in the clip, this comparison in effect only requires the wrist to be tracked at the trophy frame. Where either condition fails, both events are reported as not locatable.

3.5.4 Angle Computation

With the key frames located, the pipeline reads the four rule angles from the filtered landmark trajectories. Trunk inclination and front knee flexion are read at the trophy frame. Elbow flexion and shoulder elevation are read at the ball-impact frame. Each angle is computed only when the landmarks it needs are reliable at that frame.

The MediaPipe Pose Landmarker returns x normalised by the frame width and y by the frame height (Google AI Edge, 2026c), so the pipeline first rescales the coordinates to pixels. Without this step the aspect ratio of the frame would distort every angle. Each is then measured between two planar vectors u and v formed from landmark positions or, for the trunk, from the image vertical[2].

$$\alpha = \operatorname{atan2}(\|u \times v\|,\ u \cdot v)\,, \tag{3.2}$$

This returns a value between 0° and 180°. The computed α is the projected angle of Equation 3.1, not the true spatial one.

Front knee flexion and elbow flexion are turning angles between two consecutive segments. For the knee, u runs from the hip to the knee and v from the knee to the ankle, so a straight leg reads as zero and a bent one as a positive flexion. For the elbow, u runs from the shoulder to the elbow and v from the elbow to the wrist, on the same convention. The annotated front leg and serving arm fix the side each landmark triplet comes from.

Shoulder elevation and trunk inclination are each read as the angle between two vectors. At the serving shoulder, elevation lies between the upper-arm vector from shoulder to elbow and the trunk vector from shoulder to hip on the same side. An arm hanging along the body reads zero and the value grows as the arm is raised. For the trunk, the axis runs from the mid-hip to the mid-shoulder, each the midpoint of its left and right landmarks. The axis is

[2]Because u and v are planar, the cross product $u \times v$ is the scalar $u_x v_y - u_y v_x$ and $\|u \times v\|$ denotes its absolute value, not a three-dimensional vector norm.

measured against the image vertical. A level camera turned about the vertical axis keeps that vertical intact, as shown in Section 3.4.1. Because image coordinates grow downward, the vertical is taken as $(0, -1)$. An upright trunk then reads close to zero and a lean reads as its angle to the vertical. The reference value is instead a relative inclination, taken against the pelvis (Jacquier-Bret & Gorce, 2024a). A pelvis reference would rest on the line between the two hip landmarks. That line is far shorter than the trunk axis, so the same landmark error tilts it much more (Section 3.6.2). The image vertical carries no error of its own. The two readings agree only while the pelvis stays level. Powerful servers tilt it during loading (Kovacs & Ellenbecker, 2011).

3.5.5 Rule Evaluation and Indicator Generation

The final stage compares each computed angle against the band fixed for its rule and returns whether the angle lies inside or outside the band. The output is a set of deviation indicators, each naming its criterion, with the knee flag marking insufficient flexion. An indicator is produced only where the pipeline's conditions hold together, the camera plane supporting the criterion, the key frame being locatable and the needed landmarks being reliable at that frame. Where any of these fails, the criterion is reported as unavailable rather than pushed to a verdict, so the indicator set is what the recording can actually support.

3.6 Design of the Feasibility Assessment

This section sets out how the feasibility of the rules is judged. Two internal references stand in for a ground truth. The projection analysis prescribes the true angle and computes its projection. The automatic event detection is held against a manual check of the selected frames. Of the investigatory questions (Section 2.4), the first is settled by the criteria selection and the pipeline, so only its outcome is reported. The second is taken up in the projection and noise analysis and the event-detection check. The third rests on the decidability criterion and on the flags the pipeline returns on the corpus.

3.6.1 Error Budget

A reported angle can depart from the true joint angle at four points and the assessment that follows is organised around them. The first is the pose estimation error, the deviation of the estimated landmarks from their true image positions. Even exact landmarks show only a projection of the body, so a projection error separates the observed angle from the true spatial one. The angle is then read at a single frame and the event error is that frame's offset

from the true instant of the trophy position or ball impact, including the offset built into the pelvis proxy (Section 3.5.3). Finally, the landmarks follow the estimator's own keypoint definition, whereas the reference values were formed from anatomical joint centres. This definitional mismatch remains even when every other step is exact.

The first three sources can each be brought into the analysis, though each by a different means. The projection error needs no recording at all, since it follows from the projection geometry alone. The event error does need the recordings. It is measured on them against a manual reading of the key frames. The pose estimation error cannot be measured even there, because the clips carry no hand-annotated landmark positions. The model card gives no pixel-level error either. It reports the estimator's accuracy only as a torso-normalised detection rate (Bazarevsky et al., 2021). The error magnitude σ is therefore swept over a plausible range.

The definitional mismatch stays outside the analysis. Quantifying it would need the joint centres from a three-dimensional motion capture of the same serves, which the setup does not provide, so it is carried to the limitations (Section 3.7).

3.6.2 Projection and Noise Propagation

The projection analysis isolates the projection error using geometry alone, building on the projection relation of the third selection condition (Equation 3.1). For a given camera orientation the true angle is prescribed, projected through a level orthographic camera and compared with the value the camera would report. A single inclination such as the trunk follows that closed form. For the two-segment joints the segments are placed symmetrically about the vertical at the prescribed angle. The plane they share is turned by θ and the projected angle is re-read numerically.

Since the pose estimation error cannot be measured on the recordings, it enters as controlled landmark noise instead. This is how geometric computations are commonly tested on synthetic data (Hartley & Zisserman, 2004). The landmarks are placed at representative body-segment proportions (Winter, 2009) on a stature of 600 px, which sets the pixel length of every segment. Each exact landmark position p is perturbed independently of the others by zero-mean Gaussian noise in both image coordinates,

$$\tilde{p} = p + \eta, \qquad \eta \sim \mathcal{N}(0, \sigma^2 I_2), \tag{3.3}$$

where I_2 is the 2×2 identity, so the offset has a standard deviation of σ pixels in each image axis and no preferred direction.

The model leaves out the anisotropy of real estimator error and its correlation between landmarks, which are taken up in the limitations (Section 3.7). A fixed pixel error subtends a

larger angle across a shorter segment, so its effect depends on segment length and the noise levels below are read against this stature. Each criterion is evaluated separately at its own reference mean.

The induced spread is estimated by Monte Carlo simulation, which completes the synthetic test described above. For each criterion, each viewpoint θ and each noise level σ, the landmarks are perturbed and the angle recomputed $N = 10\,000$ times from a fixed seed (42). The standard deviation of those readings is the induced spread. The two parameters are swept rather than fixed. The noise level σ runs from 2 px to 6 px, about 0.3 % to 1 % of the stature. Two pixels stands for clean footage. Six pixels is the largest error still plausible under the admission criteria, which require the whole body to be framed at a size that resolves the limbs (Section 3.2). The viewpoint θ runs from 0° to 45° in 5° steps. At 0° the camera faces the plane of the motion head-on. That ideal cannot be held in practice, because the player's lean direction is not known before the serve. The upper bound stays just below the 46° at which projection alone moves a trunk inclination at its reference mean out of its band (Section 3.4.1). This is the range the decidability criterion reads against.

3.6.3 The Decidability Criterion and Event-Detection Stability

The system emits a decision, not an angle, so what matters is not the angular error itself but its effect on the rule. A criterion is decidable only when the noise in its input stays small against the band the decision is made on. The question is therefore whether the induced spread stays inside that band or grows large enough to push a serve across its edge.

The decidability criterion makes this precise. For each criterion and each cell of the sweep, the induced spread is held against the band half-width. That is the distance from the reference mean to the band edge, one reference standard deviation. The definition covers the one-sided knee band as well. A cell is decidable where the spread stays below the half-width and unreliable where it reaches it. A criterion counts as decidable when every cell of the swept range is decidable. The threshold uses the same factor of one as the bands (Section 3.4.2). Because σ is swept, the noise level and viewpoint at which a criterion first turns unreliable are reported as the finding.

The automatic event detection is checked against a manual reading of the same clips. A single observer steps frame by frame through each clip's pose overlay and marks one trophy and one impact frame, without seeing the detected frames. The trophy is marked at the racket's first vertical peak over the loaded legs, ball impact at the visible racket and ball contact. These are the racket-based definitions that the landmark proxies only approximate (Section 3.5.3). The manual review carries the pipeline's frame index, so both readings count frames alike. The event error is the offset between the detected and the manual frame. It is reported as the share of locatable clips whose offset exceeds 1, 3 or 5 frames and as the share of large errors

of 30 frames or more. The detected frames are not corrected. The angles are read at them, so the event error passes into the flags.

3.7 Methodological Limitations

The design carries limitations that follow from its own choices. Each is named here with the residual that remains.

The first is geometric. Every measured angle is read from a monocular projection, so it depends on how the camera is aligned to the plane of the motion. That misalignment cannot be driven to zero, because it depends in part on the player's lean direction. For the two trophy-position criteria, the third selection condition contains it by requiring the camera to face the target plane. Their residual is quantified rather than assumed (Section 3.6.2). No such condition exists for the two ball-impact criteria, since they are tied to no fixed plane. Nothing therefore holds their viewpoint within the swept range. A real lens also adds perspective distortion, which the orthographic model omits. Because it can move an angle either way, the reported projection error is an approximation.

The processing pipeline adds further limitations. The trophy frame and the two angles read there all rest on the hip landmarks, so one hip-landmark error shifts them together (Section 3.5.3). Gating and smoothing reduce such errors but cannot remove them completely. The remaining limitations act mainly on the key frames. Both events rest on kinematic proxies, the highest wrist for ball impact and the lowest pelvis for the trophy. Each can drift from the instant it stands for. The preprocessing can shift them further and its parameters are conventional rather than tuned. Gaps of up to a practical 120 ms are filled linearly, so a fill across an extremum places the event on the gap edge. A comparison of gap-filling methods recommends cubic interpolation for peak values (Mohammadzadeh Gonabadi et al., 2022). The bound and the 8 Hz cut-off refer to the clip's playback rate. For slow-motion footage without a capture-rate tag they therefore cover a shorter real interval and a higher real cut-off. Slow motion was preferred but not required, so the temporal resolution of the brief ball-impact instant is bounded by each clip's frame rate. At a standard 30 fps, one frame spans about 33 ms. The extractor also adds a lag of its own. Its One Euro filter can delay the landmarks slightly and its smoothed visibility score can delay the gating by several frames (Google AI Edge, 2026a). None of these residuals is corrected. Their effect on the key frames is part of the measured event error.

The extractor filters before the preprocessing does. In video mode it smooths landmarks and visibility, and it drops the whole pose when its presence score falls below 0.5. It does not gate single landmarks, so the pipeline's gating adds a step rather than repeating one. The smoothing does overlap. The Tasks API offers no switch for it other than image mode, which

would also give up the tracking (Google AI Edge, 2026a). The Butterworth filter therefore runs on top of the One Euro filter. The two stages do different work. The One Euro filter adapts to speed and barely smooths the fast phases, while the Butterworth filter acts evenly and adds no lag. The Butterworth stage has two costs of its own. Its cut-off was not raised to offset the double pass, so it acts at about 6.4 Hz to 7.0 Hz rather than at the nominal 8 Hz. It also overshoots slightly at rapid transitions, which a critically damped filter would avoid (Robertson & Dowling, 2003). Their combined response was not characterised, so the effective cut-off given in Section 3.5.2 describes the Butterworth stage alone. Its effect on the key frames is again part of the measured event error.

The models that feed the feasibility assessment simplify in two ways. The landmark noise is treated as isotropic and independent between landmarks. No measured error covariance exists for the estimator on serves, so it is unknown whether the error favours one direction or links neighbouring landmarks. The real error is likely also larger in the fast phases of the serve (Emmerson et al., 2026), while the model applies one noise level to both key events. So the noise figures are indicative rather than exact. The swept range of 2 to 6 pixels is itself an assumption. The decidability result holds only within it.

The definitional mismatch between the landmarks and the joint centres behind the reference values (Section 2.1.4) cannot be simulated without a joint-centre reference, so it stays an unquantified offset. This is present wherever a landmark stands in for a joint centre, as in the trunk inclination read over a connecting axis rather than the trunk segment itself. Trunk inclination carries a second offset. It is read against the vertical, while its reference is taken against the pelvis. The offset equals the lateral pelvis tilt and stays unquantified.

The reference values carry limitations of their own. They describe trained players of mixed ages, levels and serve types (Section 2.2.2). Their transfer to recreational amateurs is an assumption rather than an established fact. The bands also depend on the pooled standard deviation, whose limits Section 3.4.2 sets out. The elbow and shoulder elevation bands rest on the post-hoc exclusion of two deviating studies, so both rules carry that reservation into their flags.

Finally, the evaluation is internal and small. There is no external technique rater and no three-dimensional reference. The assessment can therefore show whether a criterion is recovered and decided, but not whether a flag marks a real fault. The synthetic true angle and the manual event check stand in as internal references. The latter is a single reading with no second rater to bound its own error. The recorded material is a small, non-statistical sample.

4 Results

The pipeline ran over the 52 admitted amateur clips. For each clip it returned a deviation indicator for every criterion the recording supported and the pipeline could measure. The two trophy-position criteria divided strictly along the camera plane. Trunk inclination was formed on the frontal clips and front knee flexion on the sagittal ones. The two ball-impact criteria were each decided on 49 of the 52 clips. Automatic event detection placed ball impact close to the manual reference frame. It placed the trophy position with a much wider scatter.

The geometric assessment ran on synthetic landmarks rather than on the clips. Monocular projection foreshortened all four angles as the camera turned off the plane. The simulated landmark noise increased the angular spread with its magnitude. Three of the four criteria stayed decidable across the whole swept range of viewpoint and noise. Only shoulder elevation crossed the decidability threshold, at the extreme corner of the grid. On the clips the flagged share ran from 73.5 % of decided elbow-flexion clips to 98.0 % of decided shoulder-elevation clips. The two ball-impact criteria read differently in frontal and sagittal clips.

4.1 Sample Composition and Criterion Availability

The admitted sample held 52 clips from the two sources set out in Section 3.2. Twenty-seven were recorded in the frontal plane and 25 in the sagittal plane. Fifty clips showed a right-handed serve with a left front leg and 2 showed a left-handed serve with a right front leg. The frame rate was 30.0 fps on 39 clips and 59.94 fps on one. The remaining 12 clips ran between 29.1 and 29.97 fps.

Table 4.1 gives the counts per criterion. All losses came from three sagittal clips. On two of them neither key event could be located, which removed front knee flexion and both ball-impact criteria. On the third, a landmark needed at impact was unreliable, which removed only the ball-impact criteria.

Table 4.1: Criterion availability across the 52 clips.

Criterion	Plane	Eligible	Decided	Lost (event or landmark)
Trunk inclination	frontal	27	27	0
Front knee flexion	sagittal	25	23	2
Elbow flexion	any	52	49	3
Shoulder elevation	any	52	49	3

Note. *Eligible* clips are those whose camera plane supports the criterion; elbow flexion and shoulder elevation are supported in either plane. *Decided* clips are those where a value was formed and evaluated against the reference band. *Lost* clips are eligible clips where no value could be formed, from an unlocatable key event or a gated landmark at the key frame, so that Eligible − Lost = Decided.

4.2 Key-Event Detection Accuracy

The detector's frames were compared against a manual reading on all 52 clips. Table 4.2 gives the detection rate, the temporal tolerances (±1, ±3 and ±5 frames), the large-error rate (≥ 30 frames) and the offset distribution for the trophy position and ball impact.

Table 4.2: Key-event detection performance against the manual reading across the 52 clips.

Metric	Trophy position (pelvis dip)	Ball impact (wrist peak)
Locatable clips	50 / 52 (96.2 %)	50 / 52 (96.2 %)
Off by > ±1 frame	36 / 50 (72.0 %)	15 / 50 (30.0 %)
Off by > ±3 frames	30 / 50 (60.0 %)	6 / 50 (12.0 %)
Off by > ±5 frames	22 / 50 (44.0 %)	4 / 50 (8.0 %)
Large errors (≥ 30 frames)	9 / 50 (18.0 %)	3 / 50 (6.0 %)
Median offset [frames]	0.0	−1.0
Interquartile range (IQR) [frames]	23.25	2.0
Mean offset [frames]	−14.14	−8.06
Maximum absolute error [frames]	155.0	173.0

Note. Offsets are computed as detected frame minus manually read frame; negative values indicate that the automatic detector placed the event earlier than the manual reading. Tolerances and large errors indicate the fraction of the 50 locatable clips where the detected frame deviated from the manual reading by more than the tolerance, or by 30 frames or more for large errors.

Ball impact stayed within ±3 frames on 44 of the 50 locatable clips. The trophy position did so on only 20, with a far wider interquartile range. Almost all large errors fell on the early side. All three at impact and eight of the nine at the trophy placed the event before the manual frame. This pulled both means well below their medians.

4.3 Projection and Noise Sensitivity

On synthetic landmarks each criterion's prescribed true angle was projected across camera misalignment $\theta \in [0°, 45°]$ and perturbed by Monte Carlo landmark noise ($\sigma \in [2, 6]$ px, $N = 10\,000$ per condition). Each angle was evaluated at the reference mean of its criterion, so the foreshortening and spread reported below held at that prescribed angle and not across the measured range of the corpus. Figure 4.1 shows the projected angles against the camera angle and Figure 4.2 shows the induced spread.

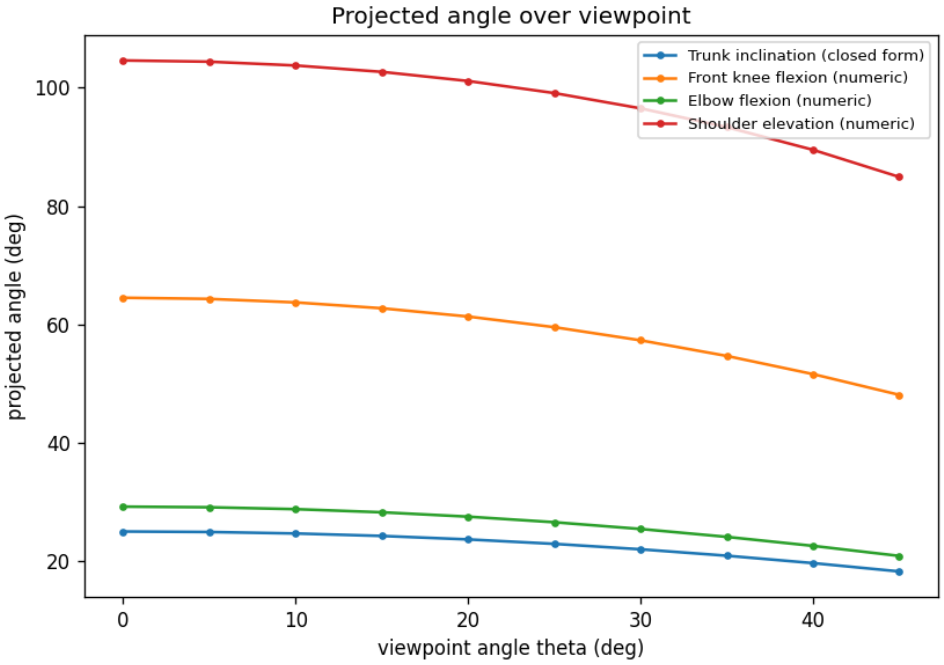


Figure 4.1: Projected joint angle α_{proj} as a function of camera misalignment $\theta \in [0°, 45°]$ for the four criteria evaluated at their reference mean values. Trunk inclination follows the analytical model of Equation 3.1; front knee flexion, elbow flexion and shoulder elevation are evaluated numerically.

Projection reduced all four angles monotonically as θ grew. At $\theta = 45°$ the foreshortening was 19.69° for shoulder elevation, 16.41° for front knee flexion, 8.33° for elbow flexion and 6.75° for trunk inclination.

The induced spread scaled linearly with σ and varied by at most 0.7° across θ. Elbow flexion carried the largest spread, from 2.81° to 8.56° across the $\sigma \in [2, 6]$ px sweep. Trunk inclination carried the smallest, from 0.94° to 2.94°. Front knee flexion ran 1.72° to 5.85° and shoulder elevation 1.84° to 6.19°.

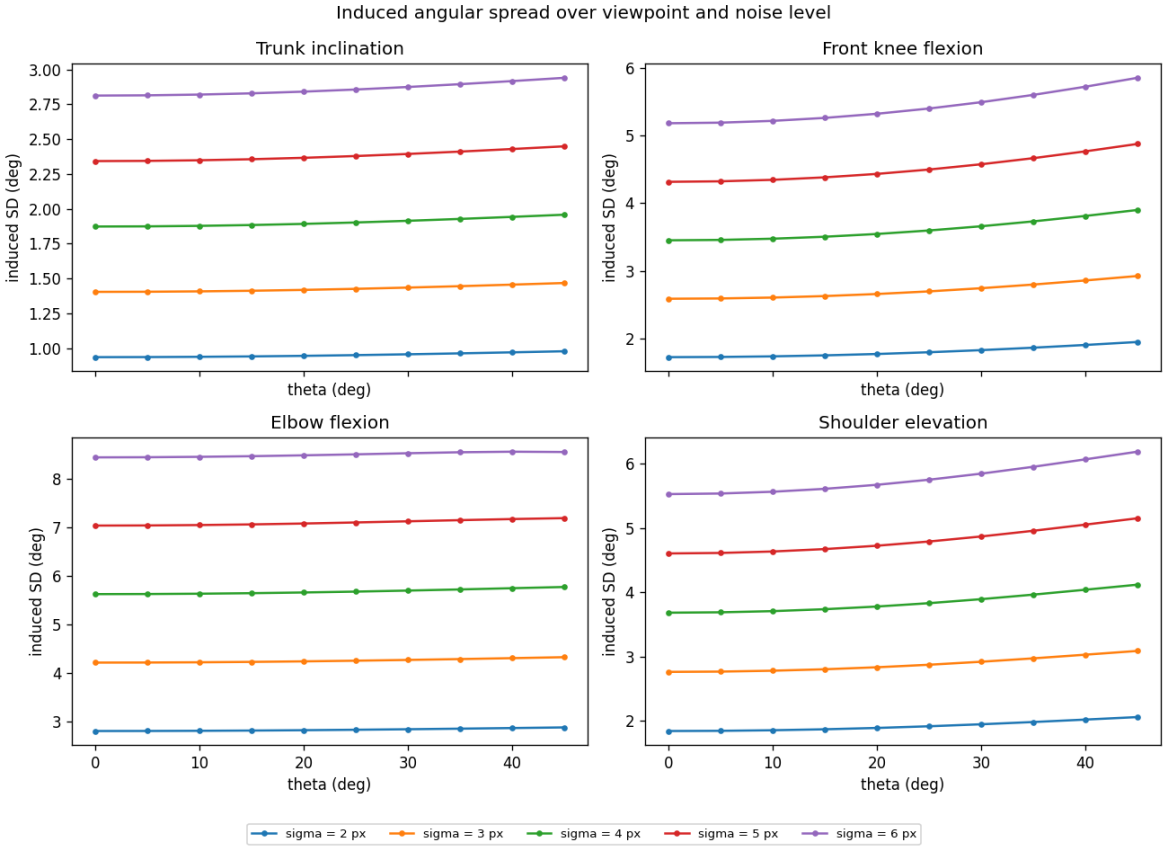


Figure 4.2: Induced angular standard deviation (SD_{noise}) as a function of camera angle $\theta \in [0°, 45°]$ across landmark noise levels $\sigma \in [2.0, 6.0]$ px from Monte Carlo simulation ($N = 10\,000$ samples per condition).

4.4 Decidability Under Projection and Noise

Each criterion's induced spread was expressed as a ratio to its band half-width, one reference standard deviation. A ratio below 1.0 was decidable and a ratio at or above 1.0 unreliable. Figure 4.3 shows the ratio across the swept grid. A threshold boundary traces the cell edges that separate decidable from unreliable cells and a point marks the onset condition at which a criterion first turns unreliable. The boundary appears only where the grid crosses 1.0.

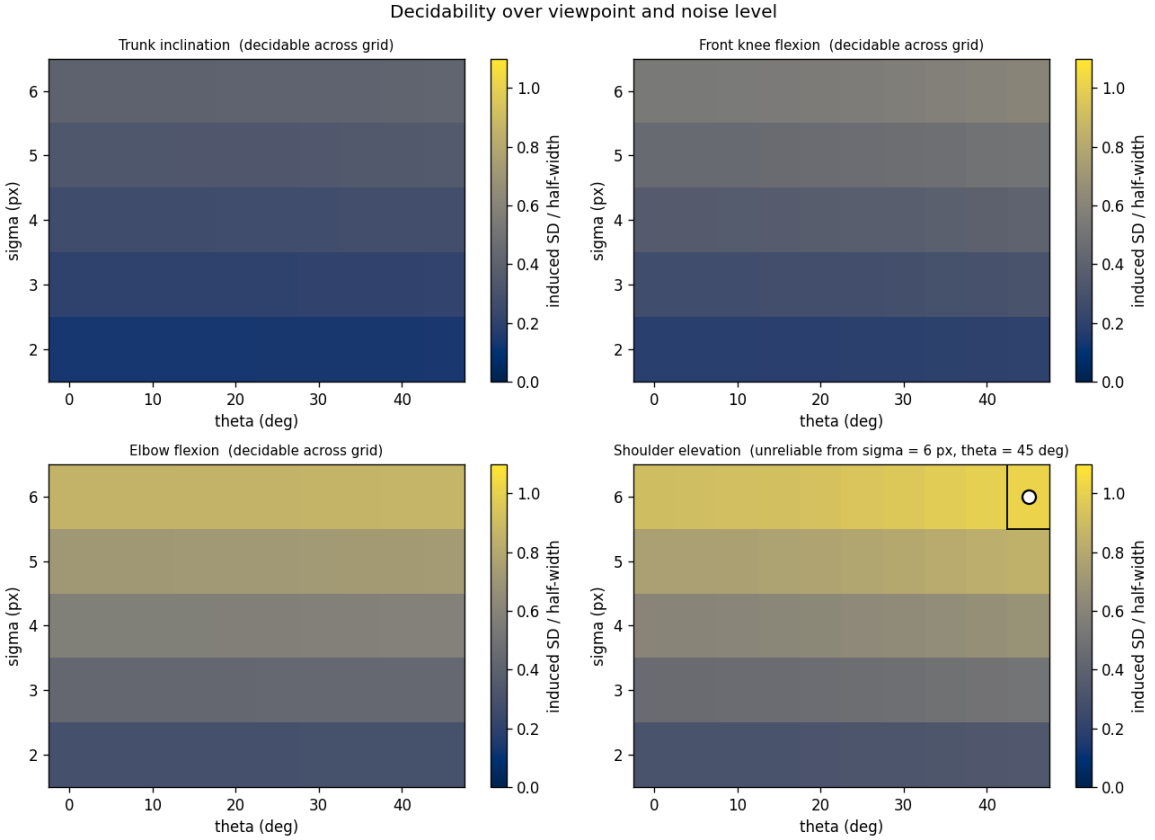


Figure 4.3: Ratio of induced noise standard deviation to reference band half-width across camera angle $\theta \in [0°, 45°]$ and landmark noise $\sigma \in [2.0, 6.0]$ px, for the four criteria. The black outline marks the decidable/unreliable boundary and the white point the onset condition.

Three of the four criteria stayed decidable across the whole grid. The peak ratio across the grid reached 0.414 for trunk inclination, 0.603 for front knee flexion and 0.864 for elbow flexion. Shoulder elevation crossed the threshold only at the extreme corner. At $\sigma = 6.0$ px and $\theta = 45°$ its spread reached 6.19° against the 6.1° half-width, a ratio of 1.014. For all $\sigma \leq 5.0$ px across the whole θ range, shoulder elevation stayed decidable with a peak ratio of 0.844.

4.5 Empirical Indicators on the Amateur Corpus

On the decided clips the pipeline compared each measured angle against its reference band and returned an indicator. These outcomes came from the corpus, not the simulation. Figure 4.4 shows the measured angles against their bands and Table 4.3 gives the distribution and the indicator counts per criterion. Three criteria carry a two-sided band and flag a departure on either side. Front knee flexion carries a one-sided lower bound, so only its lower edge at 54.8° acts as a threshold and a value above the edge is compliant. Its band is therefore drawn open towards higher flexion. The dashed line at 150° that closes it marks an anatomical plausibility limit for readability and sets no upper threshold.

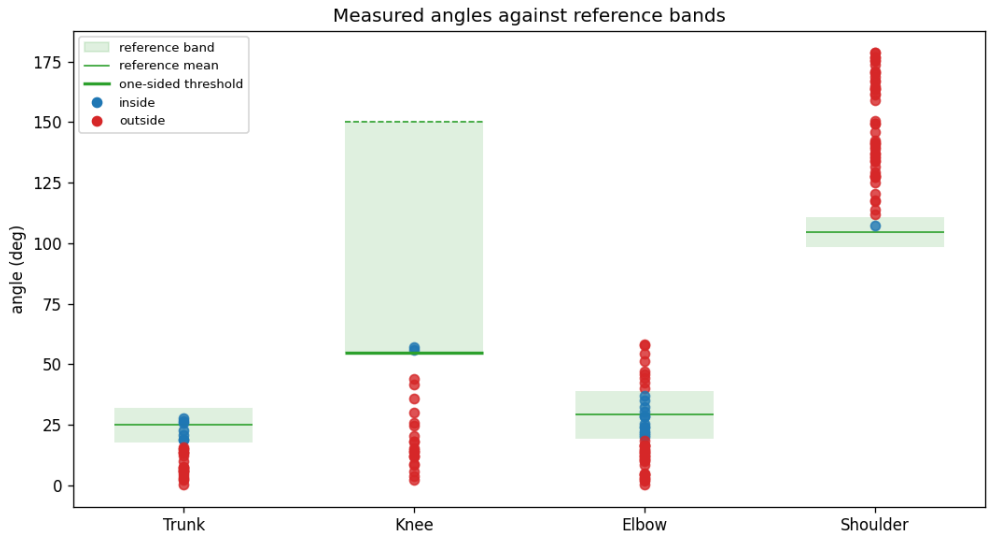


Figure 4.4: Measured angles across the decided clips against the reference bands (shaded regions) defined in Table 3.2. Blue markers lie inside the band, red markers outside it. The front knee flexion band is one-sided.

Every criterion flagged a large majority of its decided clips, from 73.5 % for elbow flexion to 98.0 % for shoulder elevation. Shoulder elevation lay above its band on every flagged clip and trunk inclination below it. Front knee flexion can only be flagged below its bound. Elbow flexion alone split across both bounds, with 27 clips below and 9 above. Of the nine clips with a large trophy error (Section 4.2), eight were flagged below the band, six on trunk inclination and two on front knee flexion.

The two ball-impact criteria were decided in both planes, so their outcomes can be split by camera plane. Shoulder elevation averaged 132.95° on the 27 frontal clips and 165.33° on the 22 sagittal clips and it lay above the band on 26 frontal and all 22 sagittal clips. Elbow flexion averaged 28.16° on the frontal clips, with 9 below the band, 10 inside and 8 above. On the sagittal clips it averaged 13.86°, with 18 below, 3 inside and 1 above.

Table 4.3: Empirical joint angle distributions and indicator outcomes on the amateur corpus.

					Outcome			
Criterion	N_{dec}	Mean (SD)	Median	Min–Max	Inside	Below	Above	Flagged
Trunk inclination	27	12.27 ± 7.94	12.33	0.23–27.69	7	20	0	74.1 %
Front knee flexion	23	21.57 ± 15.71	15.30	2.18–57.08	2	21	—	91.3 %
Elbow flexion	49	21.74 ± 16.01	16.57	0.48–58.30	13	27	9	73.5 %
Shoulder elevation	49	147.49 ± 20.78	142.29	107.33–178.75	1	0	48	98.0 %

Note. All angle values reported in degrees. N_{dec} is the number of decided clips. Reference bands: Trunk inclination [17.9°, 32.1°], Front knee flexion ≥ 54.8°, Elbow flexion [19.3°, 39.1°], Shoulder elevation [98.5°, 110.7°]. Flagged is the percentage of decided clips classified outside the reference band.

4.6 Example on a Single Clip

Figure 4.5 shows the end-to-end pipeline output on a representative clip. The clip was recorded from behind the player. The pipeline located the trophy position at frame 53 and ball impact at frame 61. Front knee flexion was unavailable, unsupported by the used plane. The three decided criteria were all flagged. Trunk inclination read 7.2° and elbow flexion 11.2°, both below the band. Shoulder elevation read 137.2°, above the band. These matched the three directions that dominated the corpus (Section 4.5).



Figure 4.5: End-to-end pipeline output on `video_000003`. Left: the detected trophy position at the mid-pelvis minimum, with the trunk inclination overlaid. Right: the detected ball impact at the maximum wrist height, with the elbow flexion and shoulder elevation overlaid.

5 Discussion

A rule-based flag is easy to over-read. When the pipeline marks a serve outside its reference band, the natural conclusion is that the player is at fault. Yet a flag can arise in three ways. The player may genuinely depart from the reference technique. The measurement chain may move the reading. Or the band may not mark a fault at all (Section 3.4.2). Under a single camera the second is a real possibility. The estimated landmarks carry noise and deviate from the joint centres behind the reference values. Monocular projection distorts the angles they form. The key event can also be mislocated. A flag therefore marks a departure from the pooled reference. On its own, it cannot say whether the player, the measurement or the band produced it.

Separating these causes is the work of what follows. Each criterion is read on its own terms, against the design that produced its measurement, the corpus it was measured on and the literature its reference value comes from. The study had no external reference for the angles, so this separation is argued rather than tested. On that reading the flags range between two poles. At one end a flag plausibly tracks technique. At the other it is largely produced by the instrument. No criterion sits cleanly at either end, so the chapter grades each by how far its flag can be trusted. The band's limitations bear on all four, but not equally.

5.1 Criterion Availability Under a Monocular Setup

One recording sees only part of the loading posture. The third selection condition leaves each clip with at most one of the two trophy criteria (Section 3.4.1). Both describe the loading at the trophy position. A single clip can therefore judge the loading only in part.

Within the planes the design admits, the pipeline lost little. Only 3 of the 52 clips lost a criterion their plane supported (Table 4.1). All 3 were sagittal, against none of the 27 frontal clips. Three cases are too few to tie the losses to the plane. What a recording yielded was therefore decided almost entirely by the plane split.

The ball-impact criteria cover both planes but give up plane control. Unlike the trophy criteria, they are tied to no fixed plane. No selection condition holds their viewpoint and their values shifted between frontal and sagittal clips (Section 4.5). The frontal and sagittal

clips were different serves by different players, so the corpus cannot fully separate viewpoint from player. The viewpoint is still a plausible cause. The shift agrees with two-dimensional running analysis, which names out-of-plane movement as one possible source of incorrect joint angles (Leporace et al., 2023). For these criteria, being available is therefore not the same as being usable.

A located key frame is not necessarily a correct one. The pipeline drops a criterion only when no event is found or a landmark is unreliable (Section 3.5.5). A mislocated event passes and its angles are read all the same. An available criterion can therefore rest on the wrong frame, which the next section takes up.

The answer to the first investigatory question is therefore narrow. Four of the six parameters the State of the Art retained can serve as criteria under a single camera (Table 3.1). At the trophy position these are trunk inclination and front knee flexion. At ball impact they are elbow flexion and shoulder elevation. One clip yields at most three of the four, one trophy criterion and both ball-impact criteria. Only the trophy criterion is read in a plane the design controls and even it can be read at the wrong frame.

5.2 Stability of the Event Proxies

How far a criterion's frame strayed from the manual reading depended on its key event. Ball impact clustered tightly around the manual frame, while the trophy position spread far wider (Table 4.2). The gap suggests that a body-based proxy is only as reliable as its link to the moment it replaces.

The two proxies differ in how closely they track their moment. Locating ball impact through the wrist worked because the wrist is at its highest when the racket meets the ball. Locating the trophy position through the pelvis rests on a weaker link. The lowest point of the pelvis depends on more than the front knee (Section 3.5.3), so it need not fall at maximum knee flexion. The pelvis trough is also likely broad, since the pelvis moves slowly at the bottom of the loading. A broad trough offers many near-minimal frames and small shifts in the signal move the detected one. This scatter therefore probably lies in the signal, not in how the extremum is found. A better detector on the same coordinate could remove large errors but is unlikely to sharpen the trough itself.

The markerless literature notes that fast movement can cost accuracy, for instance when fast hand movements cause occlusion (Aulton et al., 2025). Whether the landmarks at impact lost accuracy in this way stays open, since the event figures do not measure it. The timing of impact held up. It was located far more consistently than the trophy position at the end of the slow loading, apart from a few large errors. For locating an event, speed was therefore not the limiting factor.

The large errors did not move the medians. Both stayed near zero, so the detections were centred on the manual frame. The large errors pulled the means below the medians instead (Section 4.2). The trophy detector was therefore centred but imprecise. Some of its large errors may follow from the design. The trophy is searched only before the detected impact, so an early impact error can pull the trophy early as well (Section 3.5.3). Beyond that, this study did not trace why the errors fell early. None of them tripped the pipeline's guards. The pipeline therefore cannot flag them on its own.

The large trophy errors do more than add noise to the trophy criteria. They fell almost all early, before the loading was complete (Kovacs & Ellenbecker, 2011), when the knee is still straighter and the trunk more upright. An early error therefore pushes both trophy criteria towards the low side the corpus flagged. Eight of the nine clips with a large trophy error were flagged below the band (Section 4.5). Most clips without one were flagged below as well, so the corpus cannot show how many of these flags the errors caused. At most they account for 6 of the 20 trunk flags and 2 of the 21 knee flags. The trophy criteria are also tightly coupled to their event. The trophy frame and the trophy angles rest on the same hip landmarks, so one hip-landmark error moves them together (Section 3.5.3). The ball-impact criteria are coupled more loosely. Shoulder elevation shares no landmark with the wrist-height detector and inherits only its frame. Elbow flexion shares the wrist, but its event is sharp and has few large errors. The impact angles are therefore much less exposed to large event errors than the trophy angles.

All event figures rest on a single manual reading (Section 3.6.3). For ball impact this likely matters little, since a rater can place the visible contact to within about a frame. For the trophy position it matters more. The manual frame follows the racket's first vertical peak, which is likely harder to pin to one frame than contact, so part of the trophy offset may be the rater's own. The reference review treats this racket peak and maximum knee flexion as the same instant (Jacquier-Bret & Gorce, 2024a). The manual reading follows the racket, while the pelvis proxy aims at maximum knee flexion. The offset therefore measures how far the proxy strayed from the reference's instant, including any gap between its two definitions.

5.3 Stability Under Noise and Viewpoint

How far landmark noise moved an angle depended on the joint. In the simulation the spread hardly changed with the viewpoint (Section 4.3). Between the criteria, it differed with the length of the segments, since a fixed pixel error turns a short segment further (Section 3.6.2). Elbow flexion, with its short segments, spread the most. Trunk inclination, a single long segment read against the vertical, spread the least. These spreads were simulated at the

reference angles only. Most corpus readings lay outside the bands, away from those angles (Section 4.5). How far noise spread them on these serves was therefore not simulated.

Projection acts differently within one recording and across recordings. With one fixed camera, it shifts the angle in a similar way on every serve. The reading stays reproducible but moves away from what the reference describes. In the model, projection lowered every angle as the camera turned off the plane (Section 4.3). The model covered only the reference angles, ignored perspective and assumed one posture. Across amateur recordings, the camera does not stand in one place. The player's lean direction is not known in advance (Section 3.6.2). The ball-impact criteria are also read from either side. Between clips, projection therefore adds variation as well. For the shoulder it can even raise the angle. At impact the trunk points down and the arm points up. In this posture, turning the camera away from the plane of trunk and arm opens the angle instead of closing it. The corpus fits this. Shoulder elevation read about 32° higher in sagittal than in frontal clips. Elbow flexion read about 14° lower, in the direction the model predicts. Different players made up the two groups, so these shifts support the geometry but do not prove it (Section 4.5). The camera position can therefore change both the size and the direction of the projection error.

Taken together, the three recording conditions give an uneven answer to the second investigatory question. How stable a reading was depended on the condition and on the criterion. Event detection was measured on the clips themselves. Ball impact held, apart from a few large errors. The trophy position did not. Its early errors can push the trophy criteria towards the flagged side (Section 5.2). Landmark noise was only simulated and only at the reference angles. Across the simulated range it spread the angles by at most about 9°, most at the elbow. The camera viewpoint shifted the readings as well. For the trophy criteria the plane selection is meant to limit this shift, but how well it did so on the corpus was not measured. The ball-impact criteria have no such limit and their readings differed between frontal and sagittal clips (Section 4.5). The study can therefore vouch only for the timing of impact. The trophy timing and the camera viewpoint remain the unstable parts and the noise on the actual serves stays open.

5.4 Flag Rates and Their Separate Causes

The four flag rates do not share a cause. Trunk inclination and front knee flexion are the criteria where the flag most plausibly reflects genuine amateur deviation. Trunk inclination was flagged only below its band, although the band is two-sided. Front knee flexion nearly always fell below its bound (Section 4.5). Across the corpus, the two deficits point towards a shallow loading posture, with less knee bend and less trunk lean at the trophy position than the reference. Two measurement effects push the same way. In the model, residual

projection lowered both angles (Section 5.3). Early trophy errors read them before the loading is complete (Section 5.2). Within the simulated viewpoints, projection alone could explain a substantial part of the trunk deficit but not the whole of either. The direction of the two flags is therefore consistent with a shallower loading than the reference but does not establish one on its own.

The knee is the stronger of the two. Its angle is read directly between thigh and shank and its one-sided rule rests on evidence that more flexion goes with a faster racket (Hornestam et al., 2021; Jacquier-Bret & Gorce, 2024a). Its deficit is also large. The corpus mean sat about 43° below the reference, far below the band and not at its edge. Landmark noise has no preferred direction in the model, so it scatters the readings rather than shifting them (Section 3.6.2). Even at the largest simulated viewpoint, projection closed less than half of the gap. Event errors also lower the reading. The trophy is defined at maximum knee flexion, so any other frame reads less. The lowest pelvis need not fall at that instant (Section 5.2). How often the detector missed it was not measured, because the manual reading followed the racket. The corpus mean also lay about 34° below the slower group the review reports (Jacquier-Bret & Gorce, 2024a). The knee flag therefore reads as genuine, although projection and event errors may make the deficit look larger than it is. Trunk inclination lacks these strengths. Its angle rests on the hip-to-shoulder axis and the image vertical. Each adds an offsets the knee does not carry and neither was quantified (Section 3.7). Its deficit is smaller and large trophy errors may account for up to 6 of its 20 flags. The trunk flag is therefore the weaker of the two core signals.

Elbow flexion is the criterion where instrument and technique are hardest to separate. It was the only criterion flagged on both sides of its band and its flags split by camera plane (Section 4.5). Frontal clips centred inside the band and departed on both sides, which looks like scatter around the reference. Sagittal clips read about 14° lower, which fits projection but is confounded with the player (Section 5.3). Noise also weighs most on the elbow. At the noisiest simulated setting, its spread came close to the band half-width (Section 4.4). Noise alone could then push even a serve that matches the reference outside the band. The corpus spread of 16.01° was about twice the largest simulated spread. The simulation covered only the reference angle, so how much of this spread is noise stays open (Section 5.3). The elbow band also rests on a post-hoc exclusion of one deviating study, which narrowed it and raised its flag rate (Section 3.4.2). The review reports no measured effect of an elbow deviation. It notes only that low flexion helps to reach a high contact point and that a slightly flexed elbow seems to favour racket velocity (Jacquier-Bret & Gorce, 2024a). Even a genuine flag says little more than that the elbow differs from the average. The elbow flag rate therefore mixes viewpoint, noise and the band and possibly a genuine difference, with no way to separate these on this corpus.

Shoulder elevation is the criterion whose flag rate says least about the player. Nearly every

clip read above the band (Section 4.5). Part of this may be the viewpoint, since for the impact posture projection can raise the angle (Section 5.3). Yet even the frontal clips sat about 22° above the band. Most of that gap comes from the band itself. It rests on the post-hoc exclusion and is the narrowest of the four (Section 3.4.2). Against the band from all studies, the frontal clips would sit only about 5° above its edge (Jacquier-Bret & Gorce, 2024a). The rest may be the definitional mismatch. The reference follows an anatomical convention. The pipeline reads a two-dimensional angle between two image vectors. The offset between the two was not quantified (Section 3.7). The shoulder flag rate therefore reflects the viewpoint, the narrowed band and the definitional mismatch more than the player. It should not be read as a technical finding about players.

Beneath all four flag rates lies the band itself. Its half-width is one pooled standard deviation (Section 3.4.2). The knee case shows where that line falls. The review reports a group that flexes to 55.6° ± 8.47°, roughly 9° below the pooled mean of 64.5°. That group's racket velocity is 3.33 km h^{-1} lower than that of the group that flexes more (Jacquier-Bret & Gorce, 2024a). Its mean still sits just above the 54.8° lower bound, so fewer than half of its serves would be flagged. The band thus draws its line beneath a group the velocity evidence already marks as slower. At the same time it is narrow. Its width reflects how the study results spread (Section 3.4.2). The review notes a wide dispersion between players within the same study (Jacquier-Bret & Gorce, 2024a). A band of one pooled standard deviation can therefore flag serves that lie within normal variation. The high flag rates thus mark departures from a population average, not crossings of a validated fault boundary. Only the knee clears this reservation, since its readings fell far below even this low bound.

The single clip in the Results shows these causes side by side (Section 4.6). All three of its decided criteria were flagged, in the directions that dominated the corpus. The low trunk flag may reflect a shallow loading, although projection and event errors can lower the trunk as well. The low elbow flag may reflect a genuine difference, noise or both and it carries no established direction. The high shoulder flag mostly reflects the narrowed band and the definitional mismatch, not the player. A single serve can thus carry three flags that mean three different things.

5.5 Systematic Limits on Indicator Validity

The four sources of the error budget differ in what could reduce them (Section 3.6.1). Landmark noise depends on the footage. Simulated at the reference angles, it turned only shoulder elevation unreliable and only at the noisiest and most oblique corner of the grid (Section 4.4). The trophy error depends partly on the detector. A better detector could remove the large errors. It would probably not narrow the wide scatter of the trophy frame (Section 5.2). Noise

and the large event errors are therefore the limits that cleaner footage or a better detector could lower.

Projection and definitional mismatch set the hard boundaries. A better camera position can shrink projection but not remove it, because the player's lean is not known before the serve (Section 3.7). A large residual can still change a decision. In the model at 45° off plane, projection lowered front knee flexion and shoulder elevation by more than their band half-width, but not trunk inclination or elbow flexion (Section 4.3). At the edge of the swept range, projection alone could therefore flag a knee or shoulder serve that matches the reference. Definitional mismatch is the harder boundary. Unlike projection, it could not be modelled, because that needs joint-centre reference data the setup lacks (Section 3.7). Neither boundary yields to cleaner footage or a better detector, which is what makes both hard.

Every criterion meets at least one hard boundary, since definitional mismatch applies to all four (Section 3.7). Cleaner footage or a better detector would therefore leave every criterion with an error of unknown size. How far each flag can be trusted still differs by criterion, as graded in Section 5.4. The measurement limits make a flag unreliable under three conditions: a trophy frame that misses the loading, a ball-impact criterion read from an uncontrolled viewpoint and a noisy or oblique recording near the edge of the swept range. Together with the graded flags, these conditions answer the third investigatory question. They also bear on a question that technical validation metrics leave open (Aulton et al., 2025). Whether a system is accurate enough to act on depends here on the criterion, not on the pipeline as a whole.

The hard boundaries found here are not particular to this pipeline. Both persist in setups more favourable than a single amateur recording. Two-dimensional running analysis places the camera perpendicular to the plane it measures. Out-of-plane movement still distorts the angles, because human movement does not stay in one plane (Leporace et al., 2023). A markerless analysis of tennis strokes uses several cameras and three-dimensional reconstruction. It still names pose estimators trained on anatomically inaccurate annotations as a source of error (Emmerson et al., 2026). A single camera without laboratory control meets both boundaries at once. It is thus a harder case of known limits, not a new one.

Markerless validation in tennis names random error as the main weakness at the level of single segments (Emmerson et al., 2026). The present results point to a different constraint. Noise turned only one criterion unreliable, while projection alone could flag a reference serve for two. The two findings do not contradict each other. The noise here was only simulated and only at the reference angles (Section 5.3). Markerless error also appears largest in fast movements (Emmerson et al., 2026) and ball impact is the fastest instant this pipeline reads. The contrast is therefore one of emphasis. Within the simulated noise range, the systematic sources mattered at least as much as random scatter, which does not make random scatter harmless.

6 Conclusion

A single phone could bring serve feedback to amateur players who have no motion-capture laboratory. That promise holds only if the joint angles behind the rules can be trusted from one recording. This thesis asked how far established serve criteria translate into rules over monocular pose estimation and which systematic error sources limit their validity for coaching feedback. Four of the six parameters retained from the review became rules and each recording could carry only one of the two trophy criteria. Where the camera plane supported a criterion, the pipeline almost always returned a value. How far that value could be believed differed by criterion. Front knee flexion gave the most trustworthy flag, although its measured deficit may overstate the true one. Shoulder elevation gave the least, since its flag rate reflected its narrow band and the measurement more than the player. Trunk inclination and elbow flexion fell in between. The hard limits were projection and the definitional mismatch, which cleaner footage cannot remove. The trophy criteria were also limited by the detection of their event. Simulated landmark noise mattered less and stayed below the decidability threshold almost everywhere. Apart from the knee, even a correct reading would mark only a departure from the pooled average of trained players, not a fault. The pipeline therefore serves as a plane-aware attention flag. It shows a coach where to look, but not yet what is wrong.

Pose estimation already describes how a serve moves, but it has rarely been used to check a serve against established criteria from one phone. The thesis contributes a feasibility analysis of that check, grading each criterion by the error sources that limit it. No single statement of validity fit the whole pipeline, so each criterion needed its own. Most sports applications report technical metrics on datasets and these do not tell a practitioner whether a system is accurate enough to act on (Aulton et al., 2025). The grading asks that question of each criterion. Markerless validation in tennis finds random error the main weakness for single segments (Emmerson et al., 2026). Under one camera, the systematic sources deserve at least the same attention.

In coaching practice, a flag is most useful when it points to a correction. Only the knee flag does, since only more knee flexion is backed by evidence of a faster racket. A coach can turn it into a cue to bend the front knee more. The trunk and elbow flags carry no such direction and mark only a phase worth a second look. The shoulder flag is better left aside until its band and its measurement are settled. The knee flag needs a side view, so reading

the trunk as well takes a second camera from the front or behind. The elbow and shoulder read differently in side and frontal clips, so they should only be compared within one plane. For developers, the pipeline's guards do not catch a frame that is simply wrong. None of the worst trophy frames tripped them. A tool should therefore show the frame it read, so that the coach can reject a clearly wrong one.

The largest limit of the study itself is that it had no external reference. Its grading therefore separated player from measurement by argument rather than by test. The event frames also rested on a single reading. The small sample was limited to the light, resolution and joint visibility the admission criteria allowed. The next study belongs on the court, where the tool would be used. Several synchronised phones would film each serve from different sides. Their combined view would give the serve in three dimensions, against which the projection error of a single phone could be measured on real postures. The same recordings would separate viewpoint from player, which the present corpus could not fully do. Independent raters would bound the error of the event frames, and a larger sample under everyday light, resolution and occlusion would show how far the findings hold. Two errors would still need a laboratory. All phones read the same landmarks, so none of them can size the definitional mismatch or the true landmark error. A smaller subset recorded with markers could size both, including at the fast instant of impact. Together, the court and laboratory recordings would show whether the errors of a single phone are stable enough to be corrected on court.

The trophy position is the weak link inside the pipeline itself. The trophy frame missed the manual reading by more than three frames on most clips, while the impact frame rarely did. The pelvis signal behind the trophy frame is only loosely tied to maximum knee flexion, so the remedy likely lies in a different signal rather than a finer reading of this one. Since each clip reads only one trophy criterion, that signal need not work in both planes. In side views the knee angle itself could mark maximum flexion and in frontal views a tracked racket could mark its first vertical peak. A second signal could also check the first, since two cues that disagree mark a doubtful frame. The court study's three-dimensional recordings would give the instant of maximum knee flexion to test against. They would also show how far a wrong frame moves the trophy angles, which this study could not measure. The outcome would show whether the trophy frame can become as dependable as the impact frame.

The same court recordings could also redraw the bands. Each serve would be linked to its racket velocity and rated by independent coaches, which would show where amateur performance drops. Bands drawn there could give every criterion a direction and let each flag carry an instruction. All three next steps rest on the court recordings. They would test, for each of the four criteria, what this thesis could only argue: how far a single phone can be trusted on an amateur's serve.

Bibliography

Aulton, C., Wakili, L., Strafford, B. W., Davids, K., & Chiu, C. Y. (2025). The application of deep learning human pose estimation in sport: A systematic review. *Sports Medicine - Open*, *11*(1), 155. https://doi.org/10.1186/s40798-025-00953-3

Baily, L., Truong, N., Lai, J., & Nguyen, P. (2020). Stroke comparison between professional tennis players and amateur players using advanced computer vision. *Proceedings of the 8th International Conference on Sport Sciences Research and Technology Support (icSPORTS 2020)*, 44–52. https://doi.org/10.5220/0010145800440052

Bazarevsky, V., Grishchenko, I., & Bazavan, E. G. (2021, April 16). *Model card: MediaPipe BlazePose GHUM 3D*. Google. Retrieved August 5, 2026, from https://storage.googleapis.com/mediapipe-assets/Model%20Card%20BlazePose%20GHUM%203D.pdf

Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M. (2020). BlazePose: On-device real-time body pose tracking. https://arxiv.org/abs/2006.10204

Bini, R. R., Colyer, S., Moura, F. A., Santiago, P., & Vanicek, N. (2026). Markerless motion analysis in sports and exercise [Editorial]. *Journal of Sports Sciences*, *44*(10), 1255–1260. https://doi.org/10.1080/02640414.2026.2661158

Ebner, C. J., & Findling, R. D. (2020). Tennis stroke classification: Comparing wrist and racket as IMU sensor position. *Proceedings of the 17th International Conference on Advances in Mobile Computing and Multimedia*, 74–83. https://doi.org/10.1145/3365921.3365929

Emmerson, J., Needham, L., Williams, S., & Colyer, S. (2026). Evaluation of a markerless motion capture system for measuring mechanical work during tennis strokes. *Journal of Sports Sciences*, *44*(10), 1348–1360. https://doi.org/10.1080/02640414.2025.2555096

Google AI Edge. (2026a, April 28). *MediaPipe v0.10.35 source code: Pose landmarker*. Retrieved September 26, 2026, from https://github.com/google-ai-edge/mediapipe/tree/v0.10.35/mediapipe/tasks/cc/vision/pose_landmarker

Google AI Edge. (2026b, August 17). *Pose landmark detection guide*. Retrieved September 28, 2026, from https://developers.google.com/edge/mediapipe/solutions/vision/pose_landmarker

Google AI Edge. (2026c, August 17). *Pose landmark detection guide for Python*. Retrieved September 26, 2026, from https://developers.google.com/edge/mediapipe/solutions/vision/pose_landmarker/python

Hartley, R., & Zisserman, A. (2004). *Multiple view geometry in computer vision* (2nd ed.). Cambridge University Press.

Hornestam, J. F., Souza, T. R., Magalhães, F. A., Begon, M., Santos, T. R. T., & Fonseca, S. T. (2021). The effects of knee flexion on tennis serve performance of intermediate level tennis players. *Sensors*, *21*(16), 5254. https://doi.org/10.3390/s21165254

Jacquier-Bret, J., & Gorce, P. (2024a). Kinematics characteristics of key point of interest during tennis serve among tennis players: A systematic review and meta-analysis. *Frontiers in Sports and Active Living*, *6*, 1432030. https://doi.org/10.3389/fspor.2024.1432030

Jacquier-Bret, J., & Gorce, P. (2024b). Kinematics of the tennis serve using an optoelectronic motion capture system: Are there correlations between joint angles and racket velocity? *Sensors*, *24*(11), 3292. https://doi.org/10.3390/s24113292

Kovacs, M., & Ellenbecker, T. (2011). An 8-stage model for evaluating the tennis serve: Implications for performance enhancement and injury prevention. *Sports Health*, *3*(6), 504–513. https://doi.org/10.1177/1941738111414175

Leporace, G., Metsavaht, L., Gonzalez, F. F., Arcanjo de Jesus, F., Machado, M., Guadagnin, E. C., & Gomes-Neto, M. (2023). Validity and reliability of two-dimensional video-based assessment to measure joint angles during running: A systematic review and meta-analysis. *Journal of Biomechanics*, *157*, 111747. https://doi.org/10.1016/j.jbiomech.2023.111747

Mohammadzadeh Gonabadi, A., Cesar, G. M., Buster, T. W., & Burnfield, J. M. (2022). Effect of gap-filling technique and gap location on linear and nonlinear calculations of motion during locomotor activities. *Gait & Posture*, *94*, 85–92. https://doi.org/10.1016/j.gaitpost.2022.02.025

Robertson, D. G. E., & Dowling, J. J. (2003). Design and responses of Butterworth and critically damped digital filters. *Journal of Electromyography and Kinesiology*, *13*(6), 569–573. https://doi.org/10.1016/S1050-6411(03)00080-4

Santos-Gago, J. M., Ramos-Merino, M., Valladares-Rodríguez, S., Álvarez-Sabucedo, L. M., Fernández-Iglesias, M. J., & García-Soidán, J. L. (2019). Innovative use of wrist-worn wearable devices in the sports domain: A systematic review. *Electronics*, *8*(11), 1257. https://doi.org/10.3390/electronics8111257

Winter, D. A. (2009). Anthropometry. In *Biomechanics and motor control of human movement* (4th ed., pp. 82–106). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470549148.ch4

Appendix

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit mit dem Titel

From Pose Estimation to Coaching Feedback: Rule-Based Analysis of Amateur Tennis Serves

selbstständig und nur mit den angegebenen Hilfsmitteln verfasst habe. Alle Passagen, die ich wörtlich aus der Literatur oder aus anderen Quellen wie z. B. Internetseiten übernommen habe, habe ich deutlich als Zitat mit Angabe der Quelle kenntlich gemacht.

Hamburg, den 29. September 2026