Vision Fit: An Ai-Powered Virtual Fitting Assistant For Personalized Clothing Size Recommendation Using Dual-Anchor Anthropometric Calibration

Vinodh VS[1], Anusha Lakshmi[2]
[1]Dept of Computer Applications
[2]Assistant professor, Dept of Computer Applications
[1, 2] Dr. M.G.R Educational And Research Institute (Deemed to be University)
Chennai, Tamil Nadu ,India

Abstract- *There has been a significant shift in clothing shopping habits over the past decade, with most purchases now made online. A major challenge remains consumers' inability to try on products before purchase, leading to return rates of 30–40% due to incorrect sizing. This paper presents Vision Fit, a system that resolves this problem using an ordinary laptop webcam. Vision Fit employs a Dual-Anchor Anthropometric Calibration pipeline using Blaze Pose to detect 33 body landmarks, then converts pixel coordinates into centimetres without physical reference objects. Two anatomical proportionality ratios (Nose-Hip ≈ 48% and Nose-Ankle ≈ 82% of standing height) are fused with a 0.6:0.4 weight to derive a robust scale factor. A 30-frame temporal stabilization pipeline with jitter rejection reduces RMSE from 2.15 cm to 0.71 cm — a 67% improvement. The brand advisory module achieves 94% size-label accuracy across 10 major apparel brands with no missed recommendations over 50 subjects. The system operates fully offline via FastAPI and PyWebView.*

Keywords: pose estimation, anthropometric calibration, clothing size recommendation, BlazePose, computer vision, VisionFit

I. INTRODUCTION

Online clothing retail faces a persistent and costly problem: consumers cannot physically try on garments before purchasing. This limitation leads to return rates of 30–40% for online apparel, with each return incurring logistical, financial, and environmental costs. Conventional solutions include manual measurement guides, virtual try-on overlays, and 3D body scanners — all of which require either user effort, expensive equipment, or controlled environments.

VisionFit addresses this gap by combining real-time human pose estimation with novel anatomical calibration. Using a standard 720p webcam and a user-entered height, the system extracts seven key skeletal landmarks from a live video feed, converts pixel-space distances to physical centimetres using two anatomical anchors, and maps the derived measurements to size charts across ten major apparel brands. The entire pipeline runs offline without cloud dependencies, network connectivity, or calibration objects.

This paper makes the following contributions: (1) a novel Dual-Anchor Anthropometric Calibration method that exploits two human body proportionality constants to derive scale without external references; (2) a temporal stabilization pipeline using 30-frame averaging combined with a jitter rejection gate that reduces RMSE by 67%; (3) a two-pass brand advisory engine achieving 94% correct size-label prediction across 50 subjects and 10 brands; and (4) a fully offline, privacy-preserving desktop deployment architecture.

II. RELATED WORK

Body measurement from monocular video has been studied extensively. Weng et al. [1] demonstrated MediaPipe-based auto-capture for garment production, achieving acceptable accuracy for shoulder and bust measurements. Moreira et al. [2] evaluated multiple pose models for clinical anthropometry, noting that BlazePose Level 1 offers the best speed–accuracy trade-off for non-clinical use. Chen et al. [3] introduced a focused body model for anthropometric extraction at CVPR 2025, achieving sub-0.5 cm accuracy but requiring GPU inference.

Kaur and Kukreja [7] applied CNNs with XAI for interpretable clothing size prediction, while Sricharan et al. [8] proposed a virtual fit trail system integrating recommendation logic with an e-commerce interface. Jain et al. [9] explored image-based clothing detection and shopping recommendations using FashionAI. Timmins et al. [11] used low-cost depth sensors for under-clothing anthropometry, achieving high accuracy but requiring specialized hardware. Bazarevsky et al. [12] originally introduced BlazePose with its sub-15ms inference capability for on-device real-time tracking.

In contrast to prior work, VisionFit requires no GPU, no depth sensor, no calibration card, and no cloud connectivity. Its key novelty lies in the Dual-Anchor Calibration approach and the empirically optimized 0.6:0.4 scale fusion, enabling reliable measurements with consumer-grade webcams under typical indoor conditions.

III. SYSTEM ARCHITECTURE

A. Pose Estimation Framework

VisionFit uses MediaPipeBlazePose at complexity level 1, yielding 25–35 ms per frame on CPU while maintaining 30+ fps. The pipeline detects 33 anatomical landmarks per frame; seven are used for measurement: nose (0), left/right shoulders (11, 12), left/right hips (23, 24), and left/right ankles (27, 28). All inference runs in the browser via the MediaPipe JavaScript SDK within a PyWebView window.

B. Dual-Anchor Calibration

The primary scale factor S_1 uses the nose-to-hip vertical distance, representing ~48% of standing height H:

$$S_1 = (0.48 \times H) / d(nose, hip\_mid)\_px$$

When both ankles are visible (visibility > 0.50), a second scale factor S_2 is computed from the nose-to-ankle distance (≈82% of H):

$$S_2 = (0.82 \times H) / d(nose, ankle\_mid)\_px$$

The final scale factor fuses both anchors:

$$S\_final = 0.6 \cdot S_1 + 0.4 \cdot S_2 \quad (ankles\ detected); \quad S\_final = S_1 \quad (otherwise)$$

Anatomical bias corrections account for 2D projection compression: shoulders are multiplied by 1.025 and hips by 1.045, derived empirically from 50-subject tape measurements.

C. Temporal Stabilization

Single-frame shoulder width exhibits a standard deviation of ±2.3 cm due to pose jitter and neural network noise. VisionFit accumulates 30 consecutive valid frames and computes mean landmark positions. A Jitter Rejection Gate discards batches where the average RMS spatial shift across seven key landmarks exceeds 0.01 normalized units (~7–10 pixels at 720p). This reduces RMSE from 2.15 cm (single frame) to 0.71 cm — a 67% reduction without hardware changes.

D. Brand Advisory Engine

Derived measurements are matched against size charts for 10 brands (Nike, Adidas, Zara, H&M, Allen Solly, and five others). A two-pass algorithm first attempts exact interval match; if the measurement falls between two tiers, a nearest-neighbour fallback ensures a recommendation is always returned. No 'not found' response was observed across all 50 test subjects.

IV. EXPERIMENTAL RESULTS

A. Measurement Accuracy

Table I compares calibration methods across 50 subjects measured at distances of 1.0–3.0 m.

TABLE I. SHOULDER MEASUREMENT ERROR BY CALIBRATION METHOD

Method	MAE (cm)	RMSE (cm)	Max Error (cm)
Single-Anchor Baseline	1.84	2.15	4.60
Dual-Anchor Fusion	0.58	0.71	1.52

Dual-anchor fusion reduced MAE by 68.5% and RMSE by 67% versus the baseline. Table II presents per-dimension accuracy.

TABLE II. MEASUREMENT ACCURACY BY BODY DIMENSION

Measurement	MAE (cm)	RMSE (cm)	95th pct (cm)
Shoulder Width	0.58	0.71	1.28
Hip Width	0.74	0.89	1.51

B. Temporal Stabilization Impact

Table III demonstrates the progressive benefit of temporal stabilization and jitter gating on measurement repeatability.

TABLE III. IMPACT OF TEMPORAL STABILIZATION

Pipeline Config	Shoulder SD (cm)	RMSE (cm)
Single-frame, no stabilization	1.90	2.15
10-frame average, no jitter gate	0.95	1.12
30-frame average, no jitter gate	0.81	0.93
30-frame + jitter gate (VisionFit)	0.61	0.71

C. Brand Recommendation Accuracy

Across 50 subjects tested against all 10 brands, the advisory engine achieved 94% correct size-label accuracy (47/50 subjects). The 3 errors occurred at measurement values within 0.3 cm of a size boundary. Table IV summarizes recommendation outcomes.

TABLE IV. RECOMMENDATION DISTRIBUTION ACROSS 50 SUBJECTS

Outcome	Count	Percentage
Exact Interval Match (Pass 1)	41 / 50	82%
Nearest-Neighbour Fallback (Pass 2)	9 / 50	18%
No Recommendation Returned	0 / 50	0%
Correct Size Label vs. Tape	47 / 50	94%

D. Computational Latency

The dominant latency component is BlazePose inference at 25–48 ms per frame. Total pipeline latency from session start to result display (excluding the 30-frame accumulation interval of ~0.93 s) is 42–117 ms. End-to-end user experience is under 2 seconds on a standard dual-core CPU laptop, validated through usability testing.

E. Environmental Adaptability

Under standard illumination (300 lux) all 35 trials succeeded. Under low illumination (80 lux), 11/15 trials succeeded; 4 failed at the landmark visibility threshold. High illumination (800 lux) had no adverse effect. Participants wore typical indoor clothing (t-shirts, light tops) without impact on accuracy, as clothing offsets are absorbed into the anatomical bias correction.

V. DISCUSSION

VisionFit demonstrates that webcam-based clothing size recommendation can reach clinically acceptable accuracy (±2 cm tolerance) without external hardware, cloud infrastructure, or user calibration steps. The Dual-Anchor approach outperforms single-anchor baselines by exploiting complementary perspective distortion characteristics of torso and full-body anchors. The 0.6:0.4 fusion weight was empirically optimized via grid search on 50 subjects and reflects the superior reliability of the torso anchor at close camera distances.

The 30-frame temporal stabilization is the second key contribution. The combination of averaging and jitter gating brings a statistically significant RMSE reduction that cannot be achieved by either technique alone. The jitter gate's threshold of 0.01 normalized units was chosen to balance user patience (rejection rate ~12%) against measurement quality.

Limitations include the inability to measure circumferences (chest, waist), sensitivity to user-entered height inaccuracy (~0.9 cm error per 1 cm height error), and degraded performance below 80 lux. Multi-person environments can disrupt BlazePose's person detector, requiring session restart.

VI. CONCLUSION

This paper presented VisionFit, a fully offline, privacy-preserving virtual fitting assistant that achieves 0.71 cm RMSE in shoulder-width measurement and 94% brand size-label accuracy using only a standard webcam and user-reported height. The three architectural elements — Dual-Anchor Anthropometric Calibration, temporal stabilization with jitter rejection, and a two-pass brand advisory engine — together enable practical, consumer-grade clothing recommendations without specialized hardware. Future work will extend VisionFit to circumference measurements via side-profile capture, adaptive ML-based bias correction, and mobile browser deployment.

REFERENCES

[1] S. Y. Weng, T. L. D. Tran, and D. X. Do, "Human body measurement using MediaPipe pose auto-capture for garment production," J. Fashion Technology & Textile Engineering, vol. 11, no. 2, 2023.

[2] R. Moreira et al., "Assessing human pose estimation models for clinically relevant body segment measurements," in Proc. 2025 IEEE 38th Int. Symp. CBMS, 2025, pp. 835–840.

[3] S. Chen et al., "A focused human body model for accurate anthropometric measurements extraction," in Proc. IEEE/CVF CVPR, 2025, pp. 22658–22667.

[4] E. Duraiarasu et al., "Predicting anthropometric dimensions using machine learning for biomedical wearable devices," in Proc. 2025 Int. Conf. ICESA, 2025, pp. 1–6.

[5] R. AL Dajani, "Multimodal AI for body fat estimation: Computer vision and anthropometry with DEXA benchmarks," in Proc. 2025 IEEE CASCON, 2025, pp. 240–245.

[6] M. Abdalla and M. Ashour, "Enhancing IMU-based posture estimation through multi-modal training," in Proc. 2025 IMSA, 2025, pp. 250–255.

[7] A. Kaur and V. Kukreja, "Interpretable clothing size prediction: CNN and XAI for personalized

recommendations," in Proc. 2025 ICTMIM, 2025, pp. 1781–1786.
[8] Sricharan R. et al., "Virtual fit trail: Revolutionizing size recommendations," in Proc. 2025 ICISS, 2025, pp. 581–586.
[9] D. Jain et al., "FashionAI: Image-based clothing detection and shopping recommendation," in Proc. 2024 INOCON, 2024, pp. 1–6.
[10] A. Kaur and R. Kaur, "Product recommendations using body measurement and 3D body shape reconstruction," IJRASET, vol. 11, no. 2, pp. 832–836, 2024.
[11] S. Timmins, P. Heise, and R. Knoll, "Efficient model-based anthropometry under clothing using low-cost depth sensors," Sensors, vol. 24, no. 3, p. 842, 2024.
[12] V. Bazarevsky et al., "BlazePose: On-device real-time body pose tracking," arXiv:2006.10204, 2020.
[13] Z. Cao et al., "OpenPose: Realtime multi-person 2D pose estimation," IEEE Trans. PAMI, vol. 43, no. 1, pp. 172–186, 2021.
[14] F. Zhang et al., "Distribution-aware coordinate representation for human pose estimation," in Proc. IEEE/CVF CVPR, 2020, pp. 7093–7102.
[15] G. Lugaresi et al., "MediaPipe: A framework for perceiving and processing reality," in Proc. Third Workshop on CV for AR/VR at CVPR, 2019.