

Recognition of Abnormal Head Behaviors in Online Examinations Using Pose Geometry and Contour-Assisted Features

Tianzhen Chen^{1,*}

¹EIT Data Science and Communication College, Zhejiang Yuexiu University, Shaoxing, Zhejiang, China

*Corresponding author: terrychen17@zyufl.edu.cn

Abstract: Automated recognition of abnormal head behaviors is a practical component of computer-vision-based examination monitoring. This study proposes a lightweight method in which pose geometry provides the primary discriminative information and contour or edge morphology serves as auxiliary evidence. YuNet automatically localizes the facial region, from which 20 contour and edge descriptors and 13 five-point pose-geometric features are extracted. Logistic regression, support vector machine, random forest, and K-nearest neighbors are evaluated using three-fold leave-one-participant-out validation with group-aware inner splitting to reduce identity and adjacent-frame leakage. Experiments use 12,553 images from three classes in the Mendeley Cheating Scenario Dataset in Online Exam. Random forest achieves the highest observed performance, with an accuracy of 0.9766 ± 0.0170 and a macro-F1 score of 0.9653 ± 0.0322 . Pose Only reaches 0.9636 ± 0.0238 macro-F1, Contour Only reaches 0.2545 ± 0.0779 , and the combined feature set reaches 0.9653 ± 0.0322 . The small and participant-dependent gain of the combined representation indicates that contour information is auxiliary rather than universally beneficial. The conclusions are limited to three participants and require validation on larger independent cohorts and real examination environments.

Keywords: Online Examination, Abnormal Head Behavior Recognition, Pose Geometry, Contour Features, YuNet, Cross-Participant Validation

1. Introduction

Online examinations extend assessment beyond conventional test rooms and introduce integrity risks related to identity, gaze deviation, abnormal head rotation, unauthorized materials, and environmental interference. Recent reviews report that automated proctoring can assist human supervision by screening suspicious video, audio, and endpoint signals, while reducing the scalability limitations of fully manual remote monitoring ^[1-3]. A practical challenge is therefore to extract stable and interpretable visual cues from ordinary camera images without excessive model complexity.

Recent systems increasingly combine face analysis, gaze or head-pose cues, suspicious-object detection, and multimodal signals ^[4-9]. These studies consistently indicate that facial orientation and eye-nose-mouth geometry are informative for identifying behavior states that warrant review. However, end-to-end deep models can require substantial annotation and computation, and video-derived datasets create an additional risk: random image-level splitting may place adjacent frames or the same participant in both training and test sets, yielding optimistic estimates of generalization ^[10].

Accordingly, this study emphasizes a lightweight and reproducible design. YuNet provides automatic facial localization; five-point facial geometry supplies the primary pose descriptors; contour and edge morphology is treated as auxiliary information; and conventional classifiers recognize three visual states, normal, head_down, and head_turn. These outputs are risk cues for review and are not independent evidence of academic misconduct. Yet the cross-participant value of such lightweight descriptors remains underexamined.

The contributions are threefold. First, a compact 33-dimensional representation combines 13 pose-geometric and 20 contour or edge descriptors. Second, four conventional classifiers are compared on 12,553 target images using strict leave-one-participant-out evaluation with group-aware inner validation. Third, ablation, cluster bootstrap, paired comparison, feature-importance, and error analyses are used to characterize both performance and failure modes across participants.

2. Methodology

2.1. Overall Framework

Each examination frame is processed by YuNet, which returns a face box and five landmarks. The box is padded by 20% on each side to form a head and face region of interest (ROI). Twenty contour or edge descriptors and thirteen pose-geometric descriptors are then extracted and concatenated into a 33-dimensional feature vector for classification as normal, head_down, or head_turn. Ground-truth YOLO boxes, class identifiers, scenario identifiers, participant identifiers, frame identifiers, and file names are excluded from the model input. Figure 1 illustrates the localization, landmark, ROI, and contour-extraction stages of this pipeline.

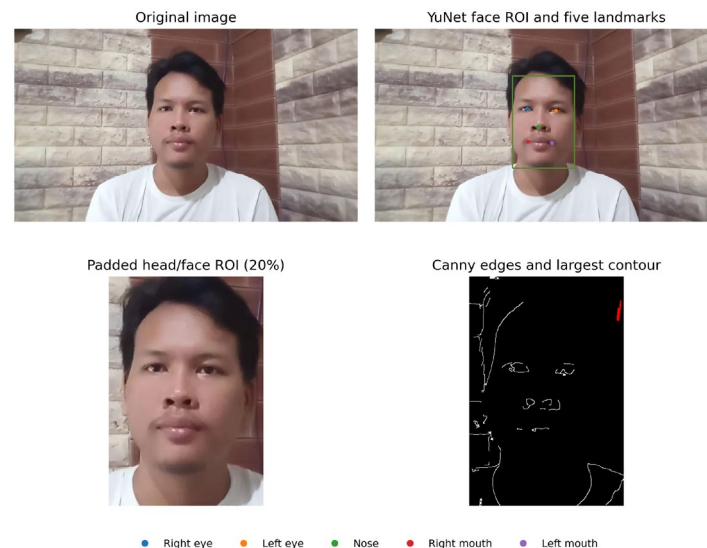


Figure 1: Example of YuNet facial localization, five-point landmarks, padded region of interest, and contour and edge extraction.

2.2. YuNet-Based Facial Region Localization

YuNet is a compact face detector designed for efficient inference ^[11]. Recent efficient face-detection studies have further emphasized computational efficiency and robustness ^[12-14]. The implementation uses face_detection_yunet_2023mar.onnx through OpenCV FaceDetectorYN, with a score threshold of 0.6, an NMS threshold of 0.3, and top_k of 5000. For multiple detections, a fixed label-independent rule prioritizes larger boxes, higher confidence, and smaller distance to the image center.

A fixed fallback sequence is used for difficult profile views: direct detection, horizontal mirroring, light enhancement, and enhanced mirroring. Mirrored detections are mapped back to the original coordinates with left-right landmark semantics restored. Of 12,553 formal images, 12,543 are detected directly and 10 require mirroring; no final detection failures occur. This result is dataset-specific and should not be interpreted as universal face-detection reliability.

2.3. Pose-Geometric Features

YuNet returns five two-dimensional landmarks: right eye, left eye, nose tip, right mouth corner, and left mouth corner. Related facial-landmark and head-pose research has emphasized robust geometric modeling under pose variation ^[15-17]. Geometric quantities are normalized by ROI size or interocular distance. Thirteen descriptors are used: eye_distance, nose_x_offset, nose_y_offset, nose_left_eye_distance, nose_right_eye_distance, nose_lr_ratio, eye_line_slope, mouth_line_slope, nose_to_eye_mid_distance, nose_to_mouth_mid_distance, upper_lower_face_ratio, mouth_width, and eye_mouth_geometry_ratio.

The horizontal nose offset and left-right nose-to-eye ratio capture asymmetry associated with head rotation, whereas vertical nose displacement and relative eye-nose-mouth distances characterize downward posture. Equation (1) defines the normalized horizontal nose offset, where x_n is the nose x-

coordinate, x_l and x_r are the left- and right-eye x-coordinates, and d_{eye} is the interocular distance. The resulting value is used as one of the pose-geometric descriptors for characterizing horizontal head rotation. This low-dimensional representation is computationally efficient and remained stable on the extreme profile samples in the present dataset.

$$\delta_x = [x_n - (x_l + x_r) / 2] / d_{eye} \quad (1)$$

2.4. Contour and Edge-Based Auxiliary Features

The contour branch characterizes local structure inside the automatically detected ROI rather than treating the rectangular ROI itself as a body contour. The ROI is converted to grayscale, smoothed with a 5 by 5 Gaussian filter, processed by Canny edge detection^[18], and closed morphologically with a 3 by 3 elliptical element. From the largest external contour, the method extracts area ratio, perimeter ratio, aspect ratio, compactness, normalized centroid, solidity, and extent. Seven Hu invariant moments^[19] and four edge-distribution descriptors complete the 20-dimensional contour feature set.

Equation (2) defines contour compactness κ , where A and P are the area and perimeter of the largest contour, respectively. This descriptor measures how closely the extracted contour approaches a compact shape and is included in the 20-dimensional contour feature set. Left-right and upper-lower edge-count ratios are stabilized with a bounded $\tanh(\log \text{ratio})$ transform. These descriptors provide structural evidence but are not assumed to be sufficient for posture recognition on their own.

$$\kappa = 4\pi A / P^2 \quad (2)$$

2.5. Classifiers and Hyperparameter Selection

Four classifiers are compared: logistic regression (LR), support vector machine (SVM), random forest (RF), and K-nearest neighbors (KNN). LR, SVM, and KNN use StandardScaler within a scikit-learn Pipeline, while RF is trained without standardization. Balanced class weights are used for LR, SVM, and RF.

Hyperparameters are selected only from each outer fold's development data using macro-F1 on the group-aware inner validation set. The search sets are C in $\{0.1, 1, 10\}$ for LR; linear or RBF SVM with C in $\{0.1, 1, 10, 100\}$ and γ in $\{\text{scale}, 0.01, 0.1\}$ for RBF; $n_{\text{estimators}}$ in $\{100, 200, 300, 500\}$ and max_depth in $\{6, 8, 12, 16, \text{None}\}$ for RF; and $n_{\text{neighbors}}$ in $\{3, 5, 7, 9, 11\}$ for KNN. After selection, the complete development set is refitted and the held-out participant is evaluated once.

3. Experimental Design

3.1. Dataset and Behavior Categories

Experiments use the public Cheating Scenario Dataset in Online Exam from Mendeley Data, Version 1, DOI 10.17632/mjrfmvsh7d.1, under CC BY 4.0^[20]. The dataset contains 15,051 images from 60 scenarios and three participants. The local archive does not provide official English strings for all 13 class IDs, so the present three-class mapping is based on scenario grouping, class IDs, and visual verification rather than claiming those working labels as original archive strings. The resulting class mapping and sample counts are summarized in Table 1.

Table 1: Mapping of the three target behavior classes and sample counts.

Scenario range	Verified working semantics	Study label	Images
Scenario 1-12	Frontal head pose	normal	3,665
Scenario 13-24	Upward head pose	excluded	2,498
Scenario 25-36	Downward head pose	head_down	2,531
Scenario 37-48	Left head rotation	head_turn	2,948
Scenario 49-60	Right head rotation	head_turn	3,409

The formal task contains 12,553 images: 3,665 normal, 2,531 head_down, and 6,357 head_turn. The head_turn class combines left and right head poses; head_up is excluded because it is outside the pre-specified three-class task, not because of model performance.

3.2. LOPO Split and Leakage-Control Strategy

The three participants are denoted S1, S2, and S3. Formal evaluation uses three-fold leave-one-participant-out (LOPO): one participant is held out for testing and the other two form the development set. This design prevents participant identity from crossing the outer train-test boundary and is more appropriate than random frame splitting for video-derived data^[10].

Within each development set, approximate clip groups are constructed from participant, scenario, gaze suffix, and contiguous frame runs. Whole groups, including exact-duplicate components, are assigned to training or validation. Formal auditing confirmed zero overlap of outer participants, inner clip groups, and exact duplicates across splits. The fold composition and group counts are reported in Table 2.

Table 2: Leave-one-participant-out folds and group-aware inner splits.

Fold	Test ID	Train	Val.	Test	Train groups	Val. groups	Test groups
1	S1	6,912	1,829	3,812	225	70	154
2	S2	6,354	1,896	4,303	222	59	168
3	S3	6,300	1,815	4,438	243	79	127

3.3. Experimental Environment and Evaluation Metrics

Experiments were run on Windows 10 64-bit with Python 3.8.8. Major dependencies included NumPy 1.20.1, Pandas 1.2.4, scikit-learn 0.24.1, SciPy 1.6.2, and OpenCV 4.9.0.80. The random seed was fixed at 42, and the YuNet model and frozen experiment artifacts were verified by SHA-256 checksums.

Accuracy, balanced accuracy, macro-precision, macro-recall, and macro-F1 are reported, with macro-F1 used as the primary metric because head_turn is the largest class. In this multiclass setting, balanced accuracy equals the arithmetic mean of class recall. Uncertainty is estimated by 2,000-repetition clip-group cluster bootstrap within each held-out participant. Resampling complete clip groups preserves within-clip frame dependence better than image-level bootstrap, but the resulting intervals do not constitute population-level confidence intervals because only three independent participants are available.

4. Results and Discussion

4.1. Comparison of Classification Models

Table 3: Cross-participant performance of the four evaluated classifiers.

Model	S1 F1	S2 F1	S3 F1	Acc.	Macro-P	Macro-R	Macro-F1
LR	0.5495	0.4708	0.8062	0.6901 ± 0.0678	0.8000 ± 0.0327	0.6899 ± 0.1265	0.6088 ± 0.1432
SVM	0.6893	0.8173	0.9130	0.8469 ± 0.0794	0.8710 ± 0.0523	0.8326 ± 0.0763	0.8065 ± 0.0916
RF	0.9850	0.9199	0.9909	0.9766 ± 0.0170	0.9761 ± 0.0181	0.9599 ± 0.0388	0.9653 ± 0.0322
KNN	0.8674	0.4778	0.9262	0.8107 ± 0.1308	0.8567 ± 0.0762	0.7794 ± 0.1679	0.7571 ± 0.1990

As summarized in Table 3, RF provides the highest observed mean performance among the four evaluated classifiers, with accuracy of 0.9766 ± 0.0170 and macro-F1 of 0.9653 ± 0.0322 . Its macro-F1 values are 0.9850, 0.9199, and 0.9909 for S1, S2, and S3, respectively. The lower S2 result indicates participant-dependent variation; accordingly, the comparison supports RF as the best-performing tested model but not a universal superiority claim.

4.2. Feature Ablation

As shown in Table 4 and visualized in Figure 2, ablation reveals that the two feature groups play different roles. Contour Only yields 0.2545 ± 0.0779 macro-F1, indicating weak stand-alone cross-

participant discrimination. Pose Only reaches 0.9636 ± 0.0238 , showing that five-point geometry contains most discriminative information. Fusion reaches 0.9653 ± 0.0322 , only 0.0017 higher on average.

Table 4: Random forest ablation results for contour and pose features.

Feature branch	S1 F1	S2 F1	S3 F1	Accuracy	Macro F1
Contour Only	0.3370	0.1500	0.2766	0.3414 ± 0.0461	0.2545 ± 0.0779
Pose Only	0.9332	0.9660	0.9915	0.9717 ± 0.0196	0.9636 ± 0.0238
Contour + Pose	0.9850	0.9199	0.9909	0.9766 ± 0.0170	0.9653 ± 0.0322

The gain is not consistent across participants: Fusion improves S1 from 0.9332 to 0.9850, reduces S2 from 0.9660 to 0.9199, and is nearly unchanged on S3. Contour information is therefore treated as auxiliary rather than as a consistently beneficial component.

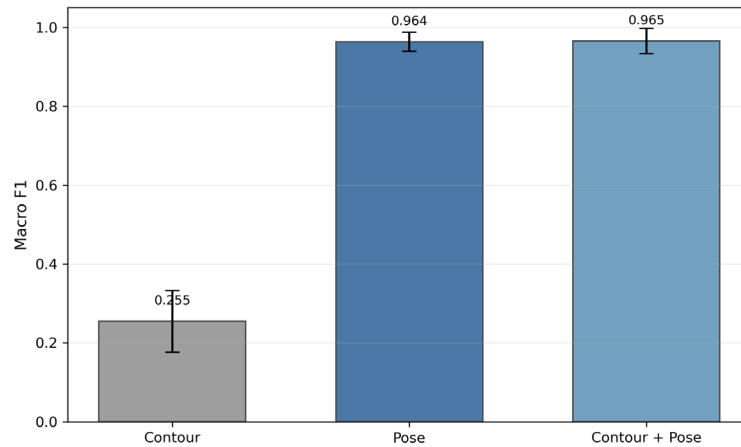


Figure 2: Ablation results for Contour Only, Pose Only, and the combined feature set.

4.3. Class-Wise Performance and Confusion Analysis

Table 5: Class-wise performance of the random forest model.

Class	Precision	Recall	F1	Support
normal	0.9420	0.9801	0.9607	3,665
head_down	0.9896	0.9356	0.9618	2,531
head_turn	0.9923	0.9907	0.9915	6,357

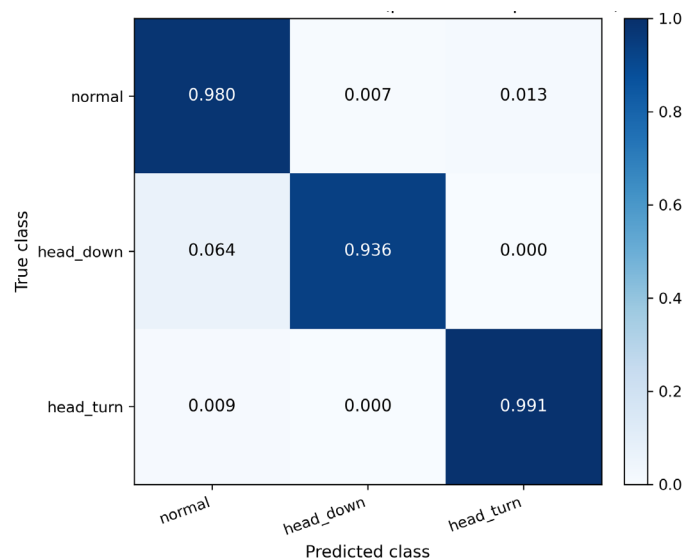


Figure 3: Normalized confusion matrix for the random forest classifier across the three outer test folds.

As shown in Table 5, the RF classifier achieved F1 scores of 0.9607, 0.9618, and 0.9915 for normal, head_down, and head_turn, respectively. The mean balanced accuracy across the three LOPO folds was 0.9599 ± 0.0388 . Figure 3 further illustrates the class-wise confusion patterns.

The main weakness occurs for S2 head_down: 148 of 568 images are classified as normal, an error rate of 26.06%. Post hoc analysis shows that nose_y_offset, upper_lower_face_ratio, and mouth_width for these errors shift toward S2's normal distribution. This is an association rather than evidence that glasses, illumination, facial shape, or another specific factor is causal.

4.4. Cross-Participant Uncertainty and Differences between Pose and Fusion

Table 6: Clip-group cluster bootstrap confidence intervals for random forest.

Test participant	Clip groups	Accuracy	Accuracy 95% CI	Macro F1	Macro F1 95% CI
S1	154	0.9843	[0.9696, 0.9939]	0.9850	[0.9719, 0.9939]
S2	168	0.9531	[0.9223, 0.9765]	0.9199	[0.8648, 0.9605]
S3	127	0.9926	[0.9819, 0.9993]	0.9909	[0.9753, 0.9993]

As reported in Table 6, cluster bootstrap intervals remain high for S1 and S3, whereas S2 has a wider macro-F1 interval of [0.8648, 0.9605]. Figure 4 provides a direct comparison of the participant-level RF macro-F1 scores. This pattern agrees with the S2 head_down errors and indicates participant-dependent uncertainty. Because resampling is performed at the clip-group level, these intervals are preferred to image-level intervals for the present video-derived data.

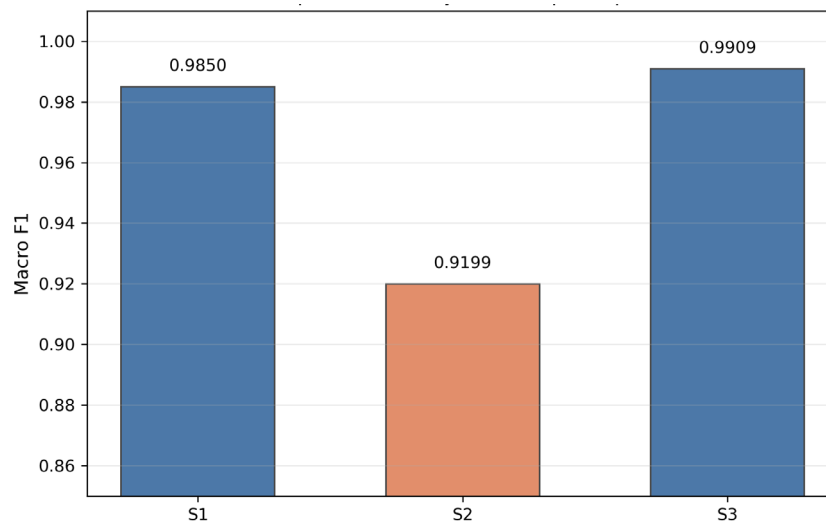


Figure 4: Random forest macro-F1 scores on the three held-out participants.

As reported in Table 7, group-level paired analysis shows inconsistent effects of adding contour features. Here, *b* denotes clip groups correctly classified by Pose Only but incorrectly classified by Fusion, whereas *c* denotes the opposite case. On S1, eight discordant groups favor Fusion; on S2, seven favor Pose Only; and S3 has no discordant groups. Exact McNemar *p*-values are 0.0078 for S1 and 0.0156 for S2, but the small numbers of discordant groups and opposing directions do not support a general claim that Fusion significantly outperforms Pose Only.

Table 7: Group-level paired comparison between Pose Only and Contour + Pose.

Participant	Groups	<i>b</i>	<i>c</i>	<i>p</i>	Observed direction
S1	154	0	8	0.0078	Fusion
S2	168	7	0	0.0156	Pose
S3	127	0	0	1.0000	No difference

4.5. Feature Importance

RF impurity importance assigns 0.8828 ± 0.0231 to the Pose group and 0.1172 ± 0.0231 to Contour. Outer-test permutation importance shows the same direction, with group sums of 0.2515 for Pose and -

0.0193 for Contour. Figure 5 presents the 15 highest-ranked features according to RF impurity importance. The agreement supports the interpretation that pose geometry is the primary discriminative information, although importance measures describe model association rather than causality.

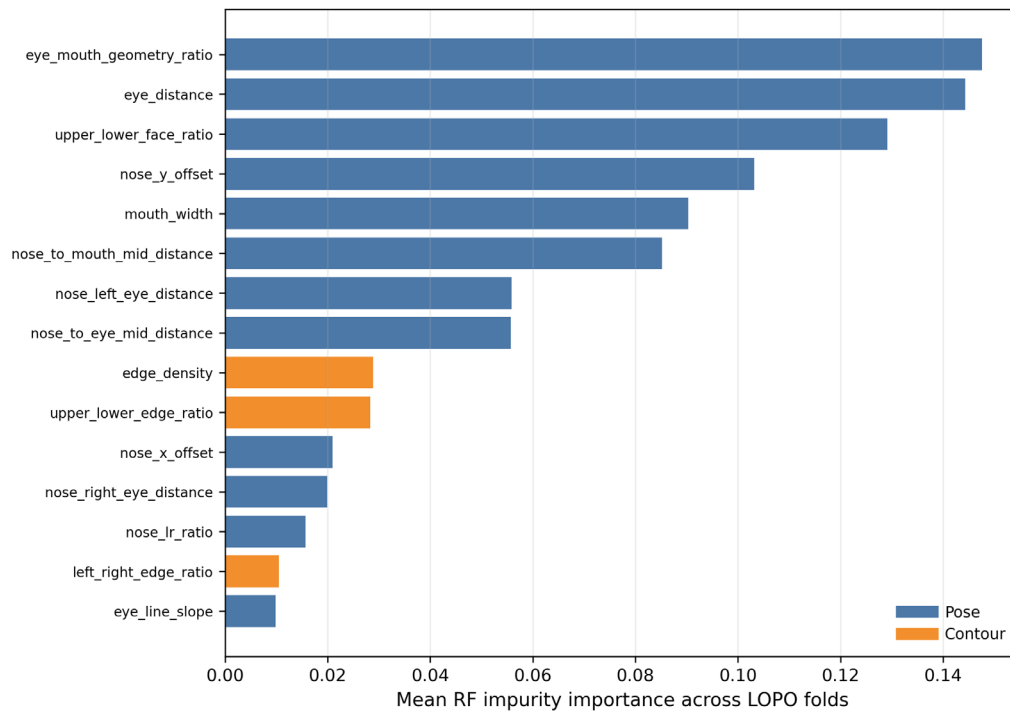


Figure 5: Top 15 random forest features ranked by impurity importance.

Among the top-ranked variables, `eye_mouth_geometry_ratio`, `eye_distance`, `upper_lower_face_ratio`, and `nose_y_offset` are prominent pose descriptors. Contour variables such as `edge_density` and `upper_lower_edge_ratio` also appear among highly ranked features, but their group-level contribution is substantially smaller.

4.6. Computational Efficiency and Discussion

YuNet requires 53.704 ms per image on average, contour extraction 2.905 ms, and pose geometry 0.0779 ms; detection plus all 33 features requires 56.851 ms per image. Final RF training across three folds takes 1.160 s, and prediction from an extracted feature vector takes about 0.0130 ms per sample. These are component timings and do not represent end-to-end system FPS.

The main strength of the framework is its combination of low-dimensional interpretable geometry with strict cross-participant evaluation. High RF performance on S1 and S3 is offset by the lower S2 result, and the weak Contour Only performance shows that head-region edge morphology is strongly dependent on participant characteristics. Larger independent cohorts and external datasets are needed to determine which auxiliary contour cues remain stable across identities and acquisition conditions.

5. Conclusions and Future Work

This study presents a lightweight approach for recognizing normal, head_down, and head_turn in online examination images using automatic YuNet localization, 13 pose-geometric features, and 20 contour or edge descriptors. Under strict three-fold LOPO evaluation, RF achieves the highest observed performance among the four tested classifiers, with macro-F1 of 0.9653 ± 0.0322 . Pose Only already reaches 0.9636 ± 0.0238 , whereas Contour Only reaches 0.2545 ± 0.0779 ; the combined representation provides only a small and participant-dependent average gain. These findings position pose geometry as the main discriminative source and contour information as auxiliary evidence.

The study is limited by three independent participants, correlated video-derived frames, and frame-level rather than temporal behavior modeling. Future work should expand participant diversity, conduct external-dataset validation, and incorporate temporal persistence or state-transition cues in a real-time

monitoring prototype. Such a system should support human review of suspicious segments rather than equating a single-frame prediction with academic misconduct.

Acknowledgements

This work was supported by the 2024 Scientific Research Project of Zhejiang Yuexiu University (Project No. N2024013), "Research on Examination Integrity Detection Based on Computer Vision Contour Extraction".

References

- [1] Nigam A, Pasricha R, Singh T, et al. A systematic review on AI-based proctoring systems: Past, present and future[J]. *Education and Information Technologies*, 2021, 26(5): 6421-6445.
- [2] Garg M, Goel A. A systematic literature review on online assessment security: Current challenges and integrity strategies[J]. *Computers & Security*, 2022, 113: 102544.
- [3] Han S, Nikou S, Ayele W Y. Digital proctoring in higher education: A systematic literature review[J]. *International Journal of Educational Management*, 2023, 38(1): 265-285.
- [4] Kaddoura S, Gumaei A. Towards effective and efficient online exam systems using deep learning-based cheating detection approach[J]. *Intelligent Systems with Applications*, 2022, 16: 200153.
- [5] Tejaswi P, Venkatramaphanikumar S, Kishore K V K. Proctor Net: An AI framework for suspicious activity detection in online proctored examinations[J]. *Measurement*, 2023, 206: 112266.
- [6] Verma P, Malhotra N, Suri R, et al. Automated smart artificial intelligence-based proctoring system using deep learning[J]. *Soft Computing*, 2024, 28(4): 3479-3489.
- [7] Hu Z, Jing Y, Wu G, et al. Multi-Perspective Adaptive Paperless Examination Cheating Detection System Based on Image Recognition[J]. *Applied Sciences*, 2024, 14(10): 4048.
- [8] Singh T, Nair R R, Babu T, et al. Enhancing Academic Integrity in Online Assessments: Introducing an Effective Online Exam Proctoring Model using YOLO[J]. *Procedia Computer Science*, 2024, 235: 1399-1408.
- [9] Malhotra M, Chhabra I. MANIT: A multilayer ANN integrated framework using biometrics and historical features for online examination proctoring[J]. *Scientific Reports*, 2025, 15: 29302.
- [10] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science[J]. *Patterns*, 2023, 4(9): 100804.
- [11] Wu W, Peng H, Yu S. YuNet: A tiny millisecond-level face detector[J]. *Machine Intelligence Research*, 2023, 20: 656-665.
- [12] Feng Y, Yu S, Peng H, et al. Detect Faces Efficiently: A Survey and Evaluations[J]. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2022, 4(1): 1-18.
- [13] Guo J, Deng J, Lattas A, et al. Sample and Computation Redistribution for Efficient Face Detection[C]//*International Conference on Learning Representations*. 2022.
- [14] Liu Y, Wang F, Deng J, et al. MogFace: Towards a Deeper Appreciation on Face Detection[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022: 4093-4102.
- [15] Li H, Guo Z, Rhee S M, et al. Towards Accurate Facial Landmark Detection via Cascaded Transformers[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022: 4176-4185.
- [16] Dhingra N. LwPosr: Lightweight Efficient Fine-Grained Head Pose Estimation[C]//*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022: 1204-1214.
- [17] Zhang C, Liu H, Deng Y, et al. TokenHPE: Learning Orientation Tokens for Efficient Head Pose Estimation via Transformers[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023: 8897-8906.
- [18] Canny J. A computational approach to edge detection[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1986, PAMI-8(6): 679-698.
- [19] Hu M K. Visual pattern recognition by moment invariants[J]. *IRE Transactions on Information Theory*, 1962, 8(2): 179-187.
- [20] Rabiha S G. Cheating Scenario Dataset in Online Exam[DS/OL]. *Mendeley Data*, 2024, Version 1. DOI: 10.17632/mjrfmvsh7d.1.