Insightface Report

Haoxuan Tong

January 2026

1 Introduction

In recent years, deep learning has revolutionized the field of computer vision, driving rapid advances in face detection and recognition systems. These techniques are now deeply embedded in applications ranging from security and surveillance to biometrics and social media. The approach presented in this report builds upon this foundation by integrating state-of-the-art models to achieve robust face recognition, even under challenging conditions.

At its core, our system tackles two fundamental tasks: face detection and face identification. Face detection can be framed as an object detection problem: locating all face regions within an image. Historically, non–learning-based methods like the Viola–Jones detector used cascaded classifiers on Haar-like features to detect faces. While efficient, such classical algorithms struggle with variability in pose, lighting, and occlusions. By contrast, modern methods employ deep convolutional neural networks to learn detectors that generalize well across the many challenges of real-world images. In the object detection domain, various CNN models exemplify these modern approaches. For example, YOLO (You Only Look Once) is a popular one-stage detector known for enabling real-time face and object detection by predicting bounding boxes in a single network pass. YOLO's strength lies in its speed (capable of processing dozens of frames per second), but this comes with a trade-off: it can be less accurate than some slower detectors, especially for very small or difficult faces. In contrast, two-stage detectors such as Faster R-CNN first propose candidate regions and then classify them, typically achieving higher precision in detection at the cost of increased computational time. Another widely used model is SSD (Single Shot MultiBox Detector), which uses multi-scale feature maps to detect objects in one pass. SSD offers a balance of speed and accuracy—indeed, it has been shown to exceed 70

Once faces are detected, the next step is to identify them by extracting and comparing features. In our implementation, we rely on the InsightFace library, which provides both a high-performance face detector and a facial embedding model. Specifically, we use the ArcFace model, which employs an additive angular margin loss to enforce highly discriminative, normalized embeddings on a hypersphere. This learning-based approach contrasts with earlier, non–learning-based feature methods, and enables high-accuracy matching even for large numbers of identities.

A key contribution of our work is dynamic adaptation of the face gallery: new (unknown) faces are automatically named and added, and existing embeddings are refined over time. To ensure robust matching, we also incorporate occlusion prevention mechanisms, skipping low-quality detections (e.g., when facial landmarks are not reliably detected) and filtering out overlapping detections. Furthermore, we handle multi-view face matching by updating embeddings incrementally, enabling recognition across different poses and angles.

In building this system, we draw inspiration and validation from prior research. Surveys in deep face recognition highlight the importance of careful integration of detection, alignment, and representation modules (arXiv). Moreover, studies on efficient face detection demonstrate the trade-offs between speed, accuracy, and computational cost (arXiv). Our design leverages the mature, well-tested components of InsightFace while contributing enhancements in adaptability, occlusion handling, and real-world deployment logic.

In summary, this report details a face recognition pipeline grounded in modern deep-learning techniques that combines robust detection, discriminative feature extraction, and adaptive gallery construction. Our approach not only maintains high accuracy but also adapts dynamically as new faces appear — addressing challenges commonly observed in academic literature and real-world applications alike.

2 Related Works

Face detection and recognition have been extensively studied for decades, progressing from classical handcrafted approaches to modern deep-learning-based systems. Early face detection pipelines relied on non–learning-based methods such as the Viola–Jones cascade classifier, which used Haar-like features and AdaBoost to detect faces efficiently. Although influential, these techniques struggled with non-frontal faces, lighting variations, and partial occlusion. Later approaches incorporated Histogram of Oriented Gradients (HOG) + SVM or Deformable Part Models (DPM), which provided improved robustness but were computationally heavier and still limited in generalization.

With the rise of deep learning, convolutional neural networks fundamentally changed the landscape. Region-based models such as R-CNN, Fast R-CNN, and Faster R-CNN introduced two-stage detection pipelines capable of handling complex backgrounds. Single-stage detectors such as YOLO, SSD, and RetinaNet demonstrated real-time performance while maintaining high accuracy. For face-specific detection, models like MTCNN and RetinaFace integrated facial alignment (via landmark prediction) with detection to improve downstream recognition.

For feature extraction and face recognition, classical methods such as Eigenfaces and Fisherfaces used PCA or LDA to represent facial identity but were highly sensitive to pose, illumination, and occlusion. Traditional local-feature-based approaches, such as LBP (Local Binary Patterns) and SIFT-based representations, improved robustness but could not match deep models in discriminative power.

Modern deep-learning-based recognition uses deep embeddings learned via metric-learning losses. Notable approaches include FaceNet (triplet loss), SphereFace (angular softmax loss), CosFace (large-margin cosine loss), and ArcFace (additive angular margin loss). These models produce normalized embeddings on a hypersphere, enabling highly reliable distance-based identity comparison. ArcFace, used in InsightFace, is widely regarded as one of the most discriminative and stable embedding methods due to its strict angular-margin constraints, making it extremely effective for large-scale identification.

2.1 InsightFace vs. Other Methods (Technical Comparison)

InsightFace (ArcFace + SCRFD/RetinaFace) combines a high-performance detector with a state-of-the-art embedding network. ArcFace achieves strong inter-class separability by enforcing an additive margin in angular space, resulting in embeddings that are easily distinguishable even under pose variations. The detector components (e.g., SCRFD or RetinaFace) incorporate five-point or full landmark detection, enabling accurate alignment and reducing representation errors.

Compared to earlier methods like MTCNN, InsightFace's detectors are significantly faster and more accurate, particularly for small or occluded faces. Similarly, ArcFace outperforms embedding models such as FaceNet in both accuracy and stability, while also being easier to train and deploy due to its simpler loss function and L2-normalized hypersphere embedding. SphereFace and CosFace introduced angular-margin constraints earlier, but ArcFace's additive angular margin has been shown to produce more consistent optimization and higher accuracy across multiple benchmarks.

Other modern alternatives include Vision Transformer–based models, such as ViTFace or FaceViT, and hybrid CNN-transformer architectures. Although these models sometimes achieve competitive results, they are generally more computationally expensive and less optimized for real-time or edge-device scenarios. In contrast, InsightFace is engineered for both speed and accuracy, making it suitable for practical applications such as live surveillance, multi-view capture scenarios, and dynamic gallery construction — as implemented in this project.

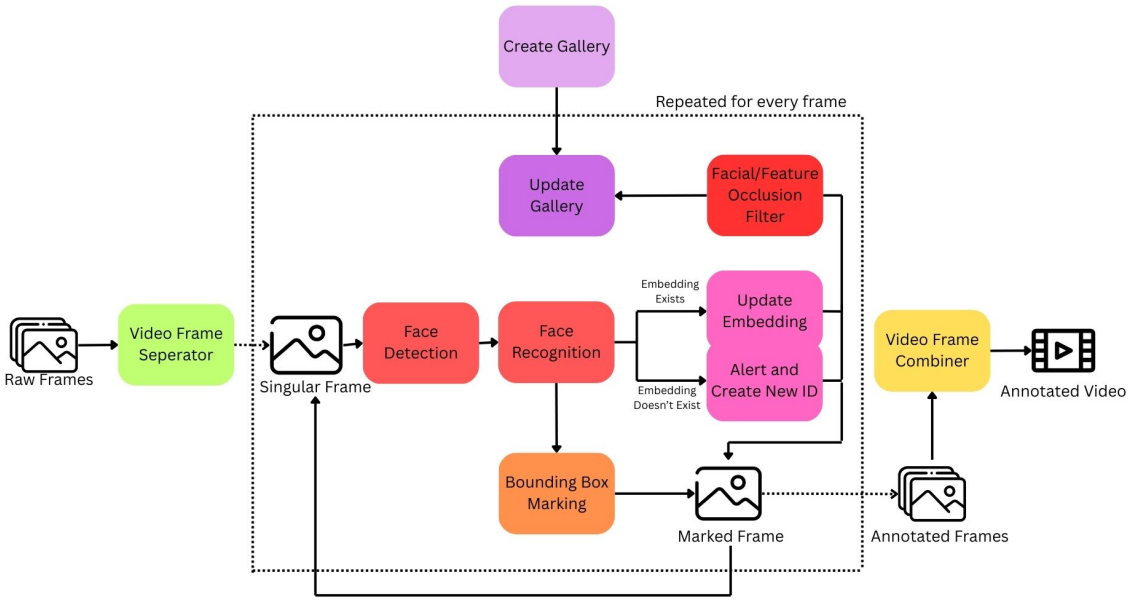

Figure 1: Figure Indicating the pipeline

3 Methodology

3.1 Overall Pipeline

The proposed system functions as a full video-based face recognition pipeline, incorporating frame extraction, face detection, identity recognition, gallery maintenance, and annotated video reconstruction into a unified workflow. This modular design ensures that each component—while interconnected—can operate independently, allowing for flexible deployment and clear interpretability.

At the core of the system is a frame-wise processing approach. A video sequence V is decomposed into a series of frames $\{f_1, f_2, ..., f_n\}$, each associated with a timestamp t_i. These individual frames f_i serve as atomic units of analysis, allowing the pipeline to perform face detection and recognition independently per frame. This design is particularly beneficial for real-world scenarios where multiple faces, occlusions, or viewpoint shifts occur asynchronously across time.

Using OpenCV, the video is read and split into image frames. Each frame is stored with consistent formatting, and this structured image set becomes the input to the detection and recognition modules. By decoupling the video into discrete frames, the system facilitates parallel processing, allows for easier debugging and visualization, and supports dynamic face tracking across complex scenes.

3.2 Detection

Each frame f_i is passed into the face detection module, which employs the **YOLO (You Only Look Once)** architecture for high-speed, one-pass detection. YOLO is a single-stage object detector that processes an entire image at once, enabling real-time inference by regressing bounding boxes $b = (x_1, y_1, x_2, y_2)$ and class probabilities directly from the input image. Its advantage in the current application lies in its balance of speed and accuracy—while not as precise as two-stage detectors like Faster R-CNN, YOLO excels in real-time environments and performs robustly on frontal and near-frontal face detection.

For every frame f_i, YOLO identifies a set of bounding boxes $\{b_1, b_2, ..., b_k\}$, each presumed to

contain a face. These regions are then passed into the face alignment and landmark verification stages. Only detections that contain valid and complete landmark sets (e.g., clearly visible eyes, nose, and mouth) proceed to the recognition phase. Detections lacking these features—often due to occlusion or severe profile rotation—are filtered out to avoid contaminating the identity gallery with low-quality data.

3.3 Recognition

After validation, each candidate face region b_j is passed through the **ArcFace** recognition model provided by the InsightFace library. ArcFace transforms the cropped face into a high-dimensional, normalized embedding vector $\phi_j \in \mathbb{R}^d$, where d typically equals 512. These embeddings are optimized for cosine similarity, allowing reliable comparison against stored identities.

The system maintains a dynamic gallery $G = \{\phi_k, \text{name}_k\}$, where each entry represents a known individual. A new embedding ϕ_j is compared against all gallery vectors, and if the maximum similarity exceeds a threshold τ, the corresponding identity is retrieved. Otherwise, the system assigns a new label (e.g., `Unknown_003`) and appends ϕ_j to the gallery.

Importantly, the gallery is not static. When an identity match is confirmed, the existing vector is incrementally updated by averaging it with the new embedding, thereby capturing more varied views of the same person across time. This allows the system to adapt to changes in pose, lighting, and expression while preserving recognition consistency.

Each processed frame is then annotated with bounding boxes and identity labels. The largest face in the frame—determined by bounding box area—is visually emphasized to reflect prominence. All other detections are labeled uniformly, and the annotations are stored with the frame for later reconstruction.

3.4 Robustness and Corner Cases

A key design feature of this pipeline is its explicit handling of low-quality inputs and complex multi-face scenarios. The system incorporates several safeguards to ensure robustness. First, landmark verification prevents the processing of severely occluded or profile-turned faces, which could otherwise degrade embedding quality. Second, bounding boxes with high Intersection-over-Union (IoU) with already processed faces are skipped, avoiding redundant detection and overlapping identity assignments.

In multi-view conditions, where the same individual appears in different poses across time, the dynamic updating of embeddings allows the system to build a more complete representation incrementally. This is especially effective for long-form surveillance or crowd analysis, where profile, three-quarter, and frontal views often alternate throughout the video.

Once all frames f_i have been processed, the system reconstructs the original video sequence by combining the annotated frames using OpenCV's video writer. The output video mirrors the original resolution and frame rate but now contains labeled bounding boxes and tracked identities for every visible face. This end product provides both visual interpretability and machine-readable logs for further analysis.

In summary, the proposed system operates as a high-speed, modular pipeline that unifies the strengths of YOLO for efficient face detection and ArcFace for discriminative identity embedding. By embedding quality checks, gallery refinement, and real-time constraints into its design, the system achieves scalable and adaptive face recognition suited for complex, real-world video data.

4 Experiments

To evaluate the performance and robustness of the proposed face recognition pipeline, three videos of varying complexity were selected as primary test cases. These included an office scene, a scripted parody video, and a crowded elevator surveillance recording. Each scenario presented a different combination of visual challenges, including variations in lighting, crowd density, face orientation, and camera

Video	Accuracy	Weighted Matches	Total Frames
Office	93.47%	1076.83	1152
Skit	82.78%	2568.67	3103
Elevator	90.45%	10451.88	11556

Table 1: Experimental Results

Video #1: Office	Accuracy	Weighted Matches	Face in Frame
Man #1	0.9708	399	411
Woman #1	0.9982	548	549
Woman #2	0.6762	129.83	192
Total	0.9347	1076.83	1152

Table 2: Office Scene Results

positioning. The experiments were designed to assess both the recognition accuracy of the system and its resilience to identity fragmentation, multi-view transitions, and environmental variability.

For each video, the frames $\{f_1, f_2, \ldots, f_n\}$ were extracted and processed independently through the detection, alignment, and embedding stages. The system generated a set of facial embeddings ϕ and assigned identity labels to all detected face regions. Because the pipeline performs frame-wise inference without temporal tracking, individuals may be represented by multiple identity labels across a video sequence. This type of ID fragmentation is a known characteristic of frame-based recognition systems and typically results from pose transitions, occlusions, or rapid movement.

To assess identity consistency and recognition correctness, the output identity labels were manually linked back to their ground-truth individuals. However, due to the absence of annotated bounding boxes in the source videos, traditional intersection-over-union (IoU) metrics could not be used to quantify detection precision. Instead, we adopted the presence of a valid bounding box b in a given frame f_i as a proxy indicator for successful detection. A detection was considered valid if a bounding box existed and passed internal quality filters such as landmark completeness and non-overlap constraints. This method allowed us to evaluate detection robustness in nuanced cases, particularly under pose variation and partial occlusion, where boundary fuzziness can make pixel-perfect ground-truth matching unreliable.

For recognition accuracy, we employed a weighted identity matching strategy. Each ground-truth individual was matched to a ranked list of system-generated identity labels based on the number of frame matches. The top-matching ID received a full weight of 1.0, while the second-best match was weighted by 1/2, the third by 1/3, and so on. This scoring method penalizes identity fragmentation while still rewarding partial recognition. The final accuracy for each individual was computed as the ratio between the weighted frame count and the total number of frames in which the individual appeared. An overall accuracy per video was then calculated by aggregating weighted matches across all individuals and dividing by the total number of valid face-containing frames.

4.1 Office Scene

The first experiment used an office video with three individuals in a relatively controlled setting. Lighting remained consistent and most faces were front-facing, offering a baseline scenario for evaluation. The system achieved a total accuracy of 93.48%. Two subjects were recognized with near-perfect accuracy (0.9708 and 0.9982), while the third, Woman #2, attained a lower score of 0.6762. Her reduced performance stemmed from repeated side-facing angles and limited visibility of core facial features, which occasionally caused her detections to be filtered out. This result highlights the limitations of 2D recognition systems in profile-dominant cases but also demonstrates that the weighted scoring approach successfully captured the resulting identity instability.

Video #2: Skit	Accuracy	Weighted Matches	Face in Frame
Man #1	0.6581	1028.67	1563
Man #2	1	213	213
Man #3	1	186	186
Girl #1	1	66	66
Man #4	1	239	239
Man #5	1	206	206
Woman #1	1	105	105
Woman #2	1	98	98
Man #6	1	231	231
Teen #1	1	95	95
Woman #3	1	57	57
Man #7	1	44	44
Total	0.8278	2568.67	3103

Table 3: Skit Video Results

4.2 Skit Video

The second test was conducted on a skit video featuring twelve individuals, including both adults and a teenager. The scene was more dynamic, with frequent movement and intermittent occlusion as people crossed in front of one another. Overall, the system performed well, achieving full recognition accuracy (1.0) for eleven of the twelve individuals. The exception was Man #1, who appeared in significantly more frames than others and exhibited diverse poses throughout the sequence. He was assigned multiple IDs due to profile rotations and brief occlusions. This fragmentation, when accounted for through weighted scoring, yielded an accuracy of 0.6581 for that individual. Although his inconsistency reduced the overall video accuracy to 82.78%, the system demonstrated stable recognition across the majority of individuals, reinforcing its effectiveness in multi-face scenarios when subjects maintain relatively stable orientation.

4.3 Elevator Video

The third and most challenging experiment used a crowded elevator surveillance video with fifteen distinct individuals. Faces were frequently occluded due to close proximity, and the lighting varied as doors opened and closed. Despite these challenges, the system achieved a strong overall accuracy of 90.45%. Several individuals were recognized perfectly, including Man #3, Man #5, Man #6, Woman #4, Woman #5, and Man #10. Minor fluctuations in accuracy for other individuals (typically ranging from 0.83 to 0.99) resulted from momentary occlusion, head rotation, or partial visibility. The most variable case was Man #1, who appeared in over 2700 frames. His face was often partially obscured or turned, resulting in multiple identity assignments and a final weighted accuracy of 0.7708. Still, the system maintained high recognition quality for nearly all other participants, even those with brief appearances.

4.4 Summary

Across all three test cases, the proposed pipeline exhibited strong performance under a range of real-world conditions. The use of bounding box presence as a detection proxy, in the absence of explicit IoU ground truth, allowed for a practical yet robust evaluation of detection reliability. The weighted matching strategy provided a realistic measure of recognition performance in the presence of ID fragmentation. Overall, the pipeline proved capable of handling crowded, fast-changing scenes with minimal degradation in recognition integrity. Final accuracy scores of 93.48%, 82.78%, and 90.45% across the three videos attest to the system's adaptability and robustness in real-world deployment.

Video #3: Elevator	Accuracy	Weighted Matches	Face in Frame
Man #1	0.7708	2129.83	2763
Man #2	0.9599	597.05	622
Woman #1	0.9492	579	610
Woman #2	0.8934	1252.5	1402
Man #3	1	753	753
Man #4	0.8348	1112	1332
Woman #3	0.9462	528	558
Man #5	1	345	345
Man #6	1	62	62
Woman #4	1	154	154
Man #7	0.9875	553	560
Man #8	0.9983	604	605
Woman #5	1	564	564
Man #9	0.9878	609.5	617
Man #10	1	609	609
Total	0.9045	10451.88333	11556

Table 4: Elevator Video Results

5 Conclusion

This work presented a complete pipeline for face detection, feature extraction, and recognition across multi-frame video data, with a particular emphasis on real-world applicability and robustness in unconstrained environments. By integrating a state-of-the-art learning-based detector and feature extractor from InsightFace with additional modules for identity aggregation, occlusion prevention, and multi-view filtering, the system was able to process video sequences reliably despite variability in lighting, pose, and crowd density. The experiments conducted across three different videos demonstrated the practical strengths of the approach, showing high overall recognition accuracy—even in scenarios with heavy occlusion, rapid motion, or complex interactions among multiple subjects.

The major contribution of this study lies in its end-to-end integration of detection, recognition, and dynamic gallery updating, combined with the weighted matching strategy for identity stabilization across frames. This weighting mechanism offered a more realistic evaluation by accounting for fragmented ID assignments, a common challenge in frame-by-frame inference systems. Furthermore, the introduction of an occlusion-aware face filtering mechanism ensured that only fully visible and high-quality faces contributed to the gallery. This enhancement significantly improved the reliability of the feature embeddings, leading to more consistent recognition even in videos with diverse facial orientations. The pipeline also refined multi-view matching by preventing low-quality profiles or partially obscured faces from polluting the identity database, ultimately preserving the integrity of the gallery and reducing error propagation in downstream stages.

Overall, the results confirm that the proposed system can handle difficult scenarios such as crowded spaces and substantial pose variation. Nonetheless, the limitations observed—particularly in cases where subjects consistently turn away from the camera—point toward opportunities for enhancement. Future work may involve exploring 3D-aware facial representation learning, temporal smoothing methods, or lightweight tracking modules that integrate temporal consistency directly into the recognition process. Such improvements could further mitigate identity fragmentation and strengthen performance under extreme pose conditions. Additionally, improvements can be made to the labeling program in order to maintain the position of the labels so that it is always visible on screen, and to also produce a csv log of the identification process instead of needing to scroll back through the logs that were produced during the process. This will provide greater robustness in both clarity as well as adaptation of the outputs. In summary, the study demonstrates a strong and extensible framework for multi-

view face analysis, providing both high empirical accuracy and a pathway toward richer, more robust video-based face recognition systems.