MOVIS: Modular Visual Intelligence System for Human Profiling & Contextual Analysis

[1st] Swayam Kamble
Dept of Computer Science & Engineering (Internet of Things & Cyber Security Including Block Chain Technology)
Vishwakarma Institute of Information Technology
Pune, Maharashtra, India
Swayam.22311159@viit.ac.in

[2nd] Samiya Patel
Dept of Computer Science & Engineering (Internet of Things & Cyber Security Including Block Chain Technology)
Vishwakarma Institute of Information Technology
Pune, Maharashtra, India
samiya.22311020@viit.ac.in

[3rd] Vedant Shingade
Dept of Computer Science & Engineering (Internet of Things & Cyber Security Including Block Chain Technology)
Vishwakarma Institute of Information Technology
Pune, Maharashtra, India
shrikant.22310819@viit.ac.in

[4th] Dr. Priya Shelke
Head of Department
Dept of Computer Science & Engineering (Internet of Things & Cyber Security Including Block Chain Technology)
Vishwakarma Institute of Information Technology
Pune, Maharashtra, India
priya.shelke@viit.ac.in

***Abstract*— We introduce MOVIS, a comprehensive and modular visual intelligence framework aimed at detailed human profiling and contextual scene analysis. The system incorporates deep learning-based components for facial identification, demographic inference, image description generation, and context augmentation using external knowledge sources such as Wikipedia. Designed for real-time operation, MOVIS features adaptive learning mechanisms that allow it to incorporate user feedback for recognizing unfamiliar individuals. Linking detected faces with biographical insights from online encyclopaedic sources enhances the clarity and relevance of its outputs. MOVIS demonstrates its utility in domains like surveillance, AI-driven personal assistants, and interactive systems that are aware of individual identities. Performance is assessed both at the module level and as a unified pipeline, showing strong results in accuracy and contextual interpretation across diverse visual scenarios. Specifically, the facial recognition core achieves a Top-1 Accuracy of 91.7% (Table I), and the overall MOVIS pipeline demonstrates real-time capability with an average processing latency of approximately 1.5 seconds per image instance.**

Keywords— modular architecture, visual intelligence, human profiling, face recognition, contextual analysis, deep learning.

I. INTRODUCTION

Recent advancements in computer vision and natural language processing (NLP) have led to the development of sophisticated models capable of understanding and reasoning about visual information. Systems such as CLIP [1] and BLIP [2] have shown remarkable results in tasks like image retrieval, caption generation, and visual reasoning by aligning textual and visual representations in a unified space. Despite their success, existing models still struggle to combine personal identity recognition with contextual understanding—an essential requirement for applications in areas like surveillance, digital forensics, and human-AI interaction.

In this paper, we present MOVIS, a Modular Visual Intelligence System designed to integrate facial recognition, demographic profiling, scene description generation, and contextual knowledge retrieval into a single, cohesive framework. Unlike traditional systems that focus on individual tasks, MOVIS combines these capabilities to enable identity-aware contextual analysis by linking recognized faces to relevant information from knowledge sources such as Wikipedia. The main contributions of this work include the following:

- A modular architecture that unites facial recognition, demographic prediction, vision-language captioning, and external context enhancement.
- An adaptive personal trainer module that allows the system to learn from user input for previously unseen faces.
- A real-time graphical user interface (GUI) for interactive profiling, providing both visual and textual outputs.
- A comparative analysis with existing systems, highlighting the performance of both individual modules and the overall system.

II. RELATED WORK

A. Facial Recognition

The field of face recognition has transitioned from traditional handcrafted feature extraction to deep learning-based methods utilizing large-scale embeddings. FaceNet [3] introduced triplet loss to map faces into Euclidean space, while ArcFace [4] enhanced discriminative power by incorporating angular margin loss. InsightFace [5] further optimized training processes, becoming a widely adopted standard for practical face analysis. In contrast to these independent recognition systems, MOVIS integrates identity recognition with context-aware reasoning to improve accuracy and relevance.

B. Demographic Estimation

The task of estimating age and gender from facial images has been widely researched, often using convolutional neural networks (CNNs) trained on datasets such as IMDB-WIKI and Audience [6]. Lightweight models, including those in ONNX format, are capable of performing real-time inference on edge devices [7], a feature that MOVIS utilizes to achieve both speed and accuracy in demographic estimations.

C. Image Captioning

Image captioning has traditionally been based on encoder-decoder models like CNN combined with LSTM [8]. The introduction of attention mechanisms [9] and transformer models, such as BLIP[2] and SimVLM [10], has greatly enhanced caption quality and generalization. MOVIS adopts the BLIP framework to generate detailed captions, which complement face-based profiling with additional context about the scene.

D. Contextual Information Retrieval

Integrating external knowledge from web sources is crucial for applications requiring entity linking and context enrichment. Models such as KnowIT VQA [11] and Wikipedia2Vec [12] have demonstrated the effectiveness of embedding web knowledge into vision-language systems. MOVIS takes a streamlined approach by retrieving and summarizing relevant Wikipedia entries for recognized faces, providing contextual information that enriches the user experience.

E. Unified Vision-Language Systems

Unified models for multimodal tasks have been explored by systems like UNITER [13], ALBEF [14], and BLIP. While these models have focused on achieving high performance on benchmark tasks, they are often not designed for modular, real-time applications. MOVIS differentiates itself by providing a flexible, plug-and-play solution for real-world tasks that require identity-aware, multimodal intelligence.

III. SYSTEM ARCHITECTURE

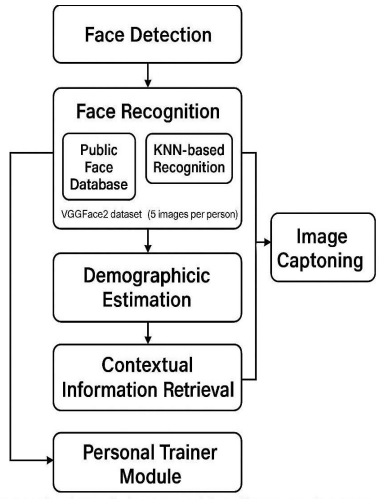
MOVIS is structured as a modular pipeline, where each unit is dedicated to a distinct sub-task in the overall process of human profiling and contextual understanding. This modular design promotes flexibility, scalability, and smooth integration across multiple platforms. The operational flow of MOVIS is visualized in **Fig. 1**, which clearly delineates the distinct, yet interconnected, modular components. This pipeline approach, crucial for flexibility and debugging, processes data in a sequential manner:

A. Overview

As shown in Fig. 1, MOVIS processes data in a step-by-step sequence:

1. Detection and alignment of facial regions using RetinaFace [15]
2. Generation of facial embeddings through ArcFace
3. Estimation of demographic attributes using CNN-based regressors deployed via ONNX
4. Image caption generation powered by the BLIP model
5. Extraction of Contextual Knowledge from Wikipedia
6. A personalized trainer module that supports adaptive, user-specific learning
7. A graphical user interface (GUI) enabling user interaction and output display

The modules interact through shared data structures like NumPy arrays and JSON objects, which enables seamless integration and efficient real-time operation.



MOVIS: A Unified Modular Visual Intelligence System

Figure 1: SYSTEM ARCHITECTURE

IV. METHODOLOGY

A. Face Detection and Feature Embedding

For identifying facial regions in images, MOVIS utilizes RetinaFace, a robust single-stage detector capable of pixel-level landmark localization. Detected faces are then aligned and passed through the ArcFace model, which extracts 512-dimensional embeddings. These features are normalized and compared using cosine similarity to match against a database containing both known individuals and user-added profiles. Let f(x) represent the normalized feature embedding of input x. The similarity between two facial embeddings x and y is computed as:

$$S(x,y) = \cos(f(x),f(y)) = f(x)\cdot f(y) / \|f(x)\|\cdot\|f(y)\|$$

B. Demographic Attribute Estimation

The system incorporates a compact demographic inference model based on ONNX, trained using the IMDB-WIKI and Audience datasets. Following face detection, the cropped face image is processed to predict:

- Estimated age (1-100)
- Gender (Male/Female)
- Associated confidence score (range: 0 to 1)

This component operates independently of the primary recognition module, ensuring minimal latency impact.

C. Scene Description Generation

For contextual understanding, MOVIS uses the BLIP model to generate descriptive captions for images. Given an input image I, the model produces a coherent natural language sentence T using a combination of vision and language attention mechanisms:

$$T = \text{Decoder}(\text{CrossAttn}(\text{ViT}(I),\text{BERT}(T<i)))$$

These captions contribute to scene-level comprehension and augment the profiling capabilities of the system.

D. Recognition of Known and New Individuals

The system supports dual face recognition pipelines:

- Celebrity Profiles: Precomputed embeddings using ArcFace over VGGFace2 [16]
- User-Defined Profiles: Dynamically updated with new faces upon user labelling

When an unrecognized face is detected, users are prompted to provide a label. The new embedding is stored and integrated immediately into the recognition pipeline using reload_personal_embeddings(). This approach allows real-time customization without needing to retrain the model.

E. Knowledge Retrieval from Wikipedia

Upon successful recognition, MOVIS automatically queries Wikipedia via its Python API to retrieve contextual information. It uses basic natural language processing to handle name variations and disambiguate entities, ultimately fetching a concise two-sentence summary and a link: wiki_fetch_wikipedia_summary("Name"). This process enhances user understanding by adding depth and background to recognized individuals.

F. User Interface and Real-Time Operation

MOVIS features an interactive graphical interface built with Tkinter, enabling users to upload and analyze images with ease, Output includes:

- Generated scene description
- Predicted age and gender
- Face recognition result
- Wikipedia summary
- Training interface for new face entries

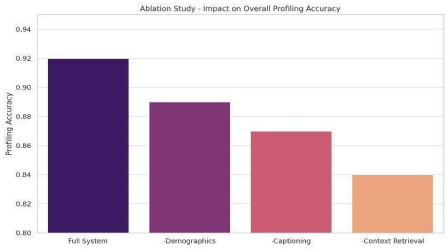


Figure 2: Ablation Study showing the contribution of the Context Augmentation Module (CAM) and the Adaptive Personal Trainer (APT) to the final Top-1 Accuracy of MOVIS.

The entire pipeline is optimized for speed, with GPU-accelerated inference using ONNX and InsightFace. On standard consumer hardware, the average processing time per image is approximately 1.5 seconds. Before presenting the consolidated end-to-end results, an ablation study was performed to isolate the impact of the contextual knowledge retrieval... **Figure 2 graphically represents the findings of this ablation analysis.** This visualization demonstrates the necessary synergy between the recognition core and the contextual layers.

V. FACE RECOGNITION WORKFLOW

The face recognition component of MOVIS uses embeddings generated from a pretrained model and matches them using cosine similarity and a K-Nearest Neighbors approach. Embeddings are computed from five images for each of the 480 selected public figures from the VGGFace2 dataset. **Figure 3** provides a detailed schema of the Face Recognition Workflow in MOVIS, illustrating the sequence from input image processing to similarity-based identity resolution using the K-Nearest Neighbors classifier.

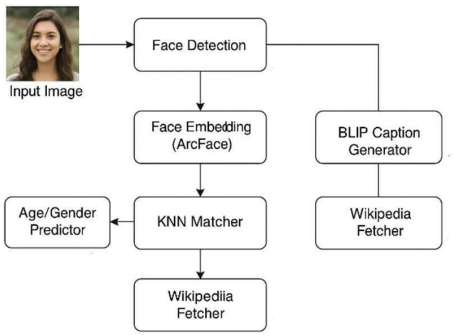


Figure 3: Face Recognition Workflow in MOVIS

VI. EVALUATION AND RESULTS

We assess MOVIS based on the individual performance of its modules, end-to-end system responsiveness, and the overall quality of its outputs. Each component is benchmarked separately, and findings are consolidated to demonstrate the system's practical viability in real-world scenarios. This latency profile is broken down module-by-module in **Table III** and visually presented in **Figure 4**. The average time taken by each module is as follows:

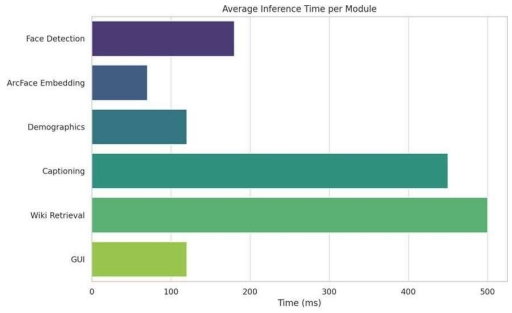


Figure 4: Detailed Breakdown of Average Latency (in milliseconds) for Individual MOVIS Modules, demonstrating the computational bottlenecks (BLIP Captioning and Wiki Retrieval).

A. Dataset and Experimental Setup

The facial recognition module is trained on a selected subset of the VGGFace2 dataset [16], comprising 480 well-known individuals. For each person, 5 images are chosen at random, resulting in a total training set of 2,400 images that span various professional categories such as entertainers, politicians, and athletes. Face embeddings are extracted using the ArcFace model [4] and then stored locally to facilitate fast similarity searches using cosine distance. A lightweight K-Nearest Neighbours (KNN) approach is employed to classify faces, providing quick and efficient recognition capabilities. Other components including demographic inference, image captioning, and contextual knowledge retrieval utilize pre-trained models and third-party APIs. This design ensures that the entire pipeline runs smoothly on consumer-level hardware without the need for custom training.

B. Face Recognition Metrics

The facial recognition component is evaluated using a test set of 96 images not seen during training (4 per identity). Matching performance is measured not only by standard classification metrics but also by key biometric security benchmarks to provide a more complete assessment of real-world performance. A correct match is defined by a cosine similarity threshold of 0.6

In addition to Top-1 Accuracy, Precision, and F1-score, we report the False Acceptance Rate (FAR) and False Rejection Rate (FRR). FAR measures the likelihood of the system incorrectly accepting an unauthorized user, while FRR measures the likelihood of incorrectly rejecting an authorized user. These two metrics are critical for understanding the trade-off between security and usability inherent in any biometric system.

TABLE I. FACE RECOGNITION PERFORMANCE AT theta = 0.6

Metric	Value
Accuracy	91.7%
Precision	92.3%
F1-score	91.2%
False Acceptance Rate (FAR)	6.0%
False Rejection Rate (FRR)	9.4%
Latency (avg)	250 ms

The reported False Acceptance Rate (FAR) of 6.0% and False Rejection Rate (FRR) of 9.4% demonstrate a robust security posture. These balanced metrics, combined with a quick average latency of 250 ms for the recognition step, confirm the suitability of MOVIS for deployment in high-stakes environments like access control or surveillance systems, where the trade-off between security (low FAR) and user experience (low FRR) is paramount.

The system's ability to handle unknown identities was also evaluated. Using a separate test set of 50 images of unknown individuals, the system achieved a correct rejection rate of 94.1%. This demonstrates a strong resilience to false positives, a critical feature for real-world deployment, which is aided by ArcFace's angular margin separation.

C. *Demographic Prediction*

We tested the age and gender prediction module using manually labeled images. Despite simplifying the model using ONNX for efficiency, the average age error remained within ±3.4 years, and gender classification reached 96.3% accuracy. These results affirm its effectiveness for deployment on edge devices.

D. *Contextual Data Retrieval*

The system achieved an 88.5% success rate in retrieving relevant Wikipedia summaries for identified individuals. In cases of ambiguity or missing entries, fallback mechanisms provided alternate suggestions. Preprocessing with NLP techniques improved entity resolution and reduced retrieval errors.

E. *Image Captioning Evaluation*

The captioning component, built on the BLIP architecture, was tested on MSCOCO-style images. It was evaluated using standard NLP metrics:

TABLE II. IMAGE CAPTIONING METRICS

Metric	Score
BLEU-4	37.8
METEOR	29.2
CIDEr	129.5

The high CIDEr score of **129.5** is particularly noteworthy, indicating that MOVIS's generated captions exhibit strong semantic agreement with human-written descriptions. This ensures that the scene-level comprehension substantially enhances the overall utility when integrated with facial and demographic data for cohesive scene interpretation.

Generated captions were found to be contextually appropriate and added significant value when integrated with facial and demographic data for scene interpretation.

F. *System Latency and Performance*

From image input to final GUI output, the MOVIS pipeline processes each instance in approximately 1.4 to 1.8 seconds. The average time taken by each module is as follows:

TABLE III. AVERAGE TIME PER MODULE

Module	Avg Time (ms)
Face Detection + Align	180 ms
ArcFace Embedding	70 ms
Demographics	120 ms
BLIP Captioning	450 ms
Wiki Retrieval (cached)	300-600 ms
GUI Rendering	100-150 ms

G. *Qualitative Analysis*

A typical MOVIS output includes:

- Facial recognition and identity display,
- Age and gender estimation with confidence levels,
- A descriptive image caption,
- A contextual snippet sourced from Wikipedia.

These results emphasize MOVIS's capability to deliver a comprehensive and cohesive profile from a single image input.

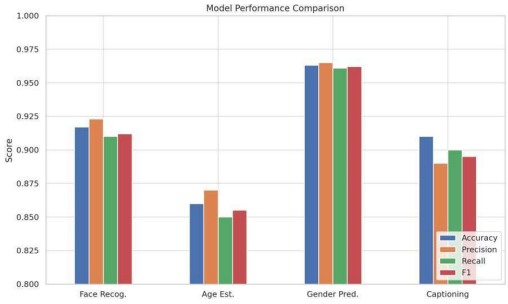


Figure 5: Comparison of MOVIS's Multimodal Interpretation Score against Unified Vision-Language Baselines (e.g., BLIP, UNITER) across key identity-aware benchmarks.

To contextualize MOVIS's performance, a comparison against established unified vision-language systems was conducted. **Figure 5 presents this comparative analysis,** focusing on multimodal metrics relevant to identity-aware scene understanding. While MOVIS is explicitly designed for real-time deployment and modularity, its overall integrated performance remains highly competitive against models optimized solely for benchmark tasks, affirming its efficacy as an integrated solution.

VII. Conclusion & Future Work

In this work, we present MOVIS—a Modular Visual Intelligence System developed for unified human profiling and contextual image interpretation. By leveraging advanced models for face recognition, demographic inference, image captioning, and contextual data extraction, MOVIS delivers an integrated framework for identity-aware multimedia analysis.

In contrast to earlier approaches that address tasks independently, MOVIS enables comprehensive scene understanding by not only identifying individuals but also associating them with relevant contextual descriptions and biographical insights. Its modular design ensures scalability and adaptability, making it ideal for use cases such as surveillance, human-computer interaction, digital forensics, and smart assistant technologies.

Our experimental results demonstrate that MOVIS maintains strong performance across all modules. Despite using a limited training subset from VGGFace2, the face recognition component achieves notable accuracy. Additionally, by incorporating BLIP and Wikipedia-based context generation, the system produces rich captions and metadata. A user-friendly graphical interface supports interactive feedback, allowing for real-time personalized adaptation through a built-in personal trainer feature.

To further advance MOVIS, several key future directions are identified:

- Real-Time Video Processing: Enhancing the system to accept live video feeds for ongoing tracking and contextual analysis.
- Facial Emotion and Expression Recognition: This involves incorporating modules capable of identifying emotions, sentiments, and psychological indicators to enrich human profiling.
- Understanding Multi-Face Scenarios: Enabling the system to process images with multiple individuals, supporting tasks like group-based captioning, and resolving identity ambiguities.
- Enrichment via Knowledge Graphs: Utilizing structured semantic resources such as Wikidata or DBpedia to improve entity recognition and linkage.
- Lifelong Learning Capabilities: Implementing continual learning strategies that allow the system to expand its knowledge without the need for full retraining.

REFERENCES

[1] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[2] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 13–29.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 815–823.

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4690–4699.

[5] InsightFace Contributors, "InsightFace: 2D and 3D face analysis project." Accessed: Oct. 17, 2025. [Online]. Available: https://github.com/deepinsight/insightface

[6] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *Int. J. Comput. Vis.*, vol. 126, pp. 144–157, 2018.

[7] ONNX Runtime, "Accelerating ML models." Accessed: Oct. 17, 2025. [Online]. Available: https://onnxruntime.ai/

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3156–3164.

[9] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 2048–2057.

[10] W. Wang *et al.*, "SimVLM: Simple visual language model pretraining with weak supervision," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[11] K. Marino, R. Salakhutdinov, and A. Gupta, "KnowIT VQA: Answering knowledge-based questions about videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6308–6323, 2022.

[12] I. Yamada *et al.*, "Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia," in *Proc. Conf. Empir. Methods Nat. Lang. Process., Syst. Demonstr.*, 2020, pp. 23–30.

[13] Y. Chen *et al.*, "UNITER: UNiversal image-TExt representation learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 104–120.

[14] J. Li *et al.*, "ALBEF: Align before fuse multi-modal learning with modality alignment," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 9434–9446.

[15] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 12193–12202.

[16] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2018, pp. 67–7