International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 09 | Oct 2026 www.irjet.net p-ISSN: 2395-0072

ThirdEye v2: A Hybrid Deep Metric Learning and Spatial Histogram of Oriented Gradients Fusion Framework for Cross-Modal Forensic Sketch-to-Photo Face Recognition

Mallikarjuna[1], K Daavan Kumar Reddy[2], Manu Kumar H N[3], Mrs. Rekha B.N[4]

[1,2,3]Student, Dept. of Computer Science & Engineering, Sir M. Visvesvaraya Institute of Technology, Bengaluru, India
[4]Associate Professor, Dept. of Computer Science & Engineering, Sir M. Visvesvaraya Institute of Technology, Bengaluru, India

---

Abstract - *When no photograph or closed-circuit television footage of a suspect exists, an eyewitness-derived sketch is often the only clue investigators have, yet face recognition models trained purely on photographs handle hand-drawn or composite sketches poorly. The reason is a substantial cross-modal gap: unlike photographic mugshots, sketches carry none of the photometric texture, skin tone, or lighting cues those models depend on, which drives a standard deep network such as FaceNet down to just 12.11 percent Rank-1 accuracy. ThirdEye v2 addresses this with a complete sketch-to-suspect pipeline: a JavaFX desktop application for assembling composite sketches interactively, paired with a Python and FastAPI machine learning service that performs the identification. Its recognition core blends two signals: a 512-dimensional FaceNet embedding passed through a projection head with 164,736 parameters trained using triplet margin loss, and a contrast-normalized, spatially weighted Histogram of Oriented Gradients descriptor, combined at the score level. Tested on the CUHK Face Sketch benchmark, this fused approach reaches 85.71 percent Rank-1 accuracy with an area under the curve of 0.9898 against a gallery of 189 candidates, and 92.00 percent on the standard 100-pair protocol, a gain of over seventy percentage points against plain FaceNet, with matching still completing in a median of 307.9 milliseconds.*

Key Words: Forensic face recognition, Sketch-to-photo matching, Cross-modal domain gap, Deep metric learning, FaceNet, Histogram of oriented gradients, Triplet loss, Score fusion

1. INTRODUCTION

1.1 Background

When a suspect leaves behind no photographic or video trace, law enforcement has to fall back on what a witness remembers: a composite sketch built from their description. That sketch is then checked, by eye or by software, against a database of mugshots. Doing this by hand is slow, depends heavily on the judgment of whoever is comparing the faces, and requires a trained forensic artist to even produce the sketch in the first place all of which point to the need for a tool that handles both the sketch construction and the matching step automatically. Several recent studies have addressed forensic sketch construction and recognition [1]–[6].

1.2 The Cross-Modal Domain Gap

Comparing a sketch to a photograph is inherently harder than comparing two photographs. A sketch is little more than sparse, colorless edge lines, while a mugshot carries full red-green-blue (RGB) texture, shading, and lighting variation. On top of that, human memory itself distorts the sketch eyes drawn too large, a chin shape that is slightly off adding further noise. Because networks like FaceNet learn from realistic color photographs, their earliest layers key on texture and color that simply do not exist in a sketch, so accuracy collapses when such a network is pointed at sketch-to-photo matching without adaptation.

1.3 Objectives and Scope

ThirdEye v2, as presented here, is a full forensic workstation built from two parts: an interactive desktop client for putting together a modular composite sketch, and a machine-learning service that closes the cross-modal gap by fusing a learned metric embedding with a hand-engineered spatial Histogram of Oriented Gradients (HOG) descriptor. Rank-N accuracy, Receiver Operating Characteristic Area Under the Curve (ROC-AUC), and inference latency are all measured on the CUHK Face Sketch (CUFS) dataset as the benchmark. Generating sketches automatically for instance with a diffusion model from a text description and handling faces turned more than roughly 15° off-axis are both left outside this project's scope.

2. LITERATURE SURVEY

Table -1 lists the prior work that shaped the design of ThirdEye v2, drawn from deep metric-learning embeddings, sketch-photo synthesis benchmarks, generative cross-modal approaches, and classical texture descriptors.

Table -1: Summary of related work

Author/Year	Method	Key Finding / Relevance
Schroff et al. [7]	FaceNet / triplet loss	99.63% on Labeled Faces in the Wild (LFW); backbone embedding, needs cross-modal adaptation
Wang & Tang [8]	CUFS dataset, Markov random field (MRF) synthesis	Established sketch-photo benchmark; source of splits
Dalal & Triggs [9]	HOG descriptor	Robust gradient shape descriptor; basis of HOG stream
Galea & Farrugia [10]	CycleGAN translation	Realistic but demands a graphics processing unit (GPU); motivates lightweight alternative
Ojala et al. [11]	Local Binary Patterns	Illumination-robust texture cue; evaluated in ablation

3. SYSTEM ARCHITECTURE

ThirdEye v2 is built as a two-tier client-service system, shown in Fig. 1. On the client side, a JavaFX desktop application gives the user an interactive canvas for stacking facial components hair, eyes, nose, mouth, chin, and facial hair into a single composite sketch; that sketch is exported as a Portable Network Graphics (PNG) image and sent as a Hypertext Transfer Protocol (HTTP) multipart POST to a Python FastAPI service. On the server side, the microservice aligns the incoming face, extracts its two-stream feature representation, scores it against every entry in the SQLite suspect database using late score fusion, and hands back a ranked list of the closest matches.

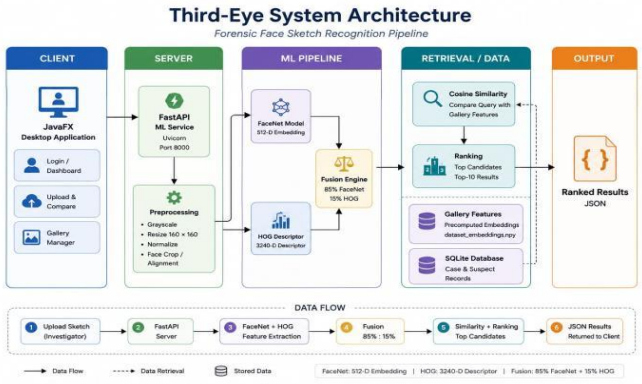


Fig -1: High-level architecture of the ThirdEye v2 client-service pipeline

3.1 Desktop Client

The desktop client is written in JavaFX around three main controllers: a Dashboard Controller that handles suspect-record management, an Upload_sketchController that submits a sketch for matching and renders the resulting gallery of candidates, and a DeepMatchClient that packages the composite canvas and makes the Representational State Transfer (REST) call to the machine learning (ML) microservice.

3.2 Machine Learning Microservice and Data Store

The microservice runs on FastAPI and warms up the model on startup, loading the FaceNet backbone and projection head into memory before the first request arrives so there is no cold-start delay. Suspect records image blobs and their precomputed HOG descriptors live in a local SQLite database, which keeps the whole system usable fully offline, a requirement in law-enforcement settings without reliable external network access.

4. METHODOLOGY

Two complementary feature streams feed the recognition engine and are combined at the score level, as described in the following subsections.

4.1 Stream 1: Metric-Learned Deep Embedding

Each aligned face, resized to 160×160×3, goes through a pretrained Inception-ResNet-v1 FaceNet backbone, producing a 512-dimensional, L2-normalized embedding z_{raw}. A projection head $\varphi(\cdot)$ Dense(256) → Batch Normalization (BN) → Rectified Linear Unit (ReLU) → Dropout(0.2) → Dense(128) → L2-normalize, 164,736 parameters in total then remaps z_{raw} into a shared, 128-dimensional cross-modal metric space:

$$z_{proj} = \text{L2_Normalize}(W_2 \cdot \text{Dropout}(\text{ReLU}(\text{BN}(W_1 z_{raw} + b_1))) + b_2) \quad (1)$$

This projection head is trained offline using a triplet margin loss (margin = 0.3) on anchor–positive–negative sketch/photo triplets; deep similarity S_{deep} is then simply the cosine similarity between the projected query and gallery embeddings.

4.2 Stream 2: Weighted Spatial Histogram of Oriented Gradients

Each grayscale input first goes through Contrast Limited Adaptive Histogram Equalization (CLAHE) contrast normalization (clip limit 2.0, tile grid 8×8) followed by a 3×3 Gaussian blur, which evens out the contrast gap between hand-drawn strokes and photographic detail. Sobel gradients are then computed on the 160×160 image,

split into 20×20 cells of 8×8 pixels each with nine unsigned orientation bins, giving a 3,600-dimensional descriptor in the deployed configuration. Before L2 normalization, each cell's histogram is weighted by an elliptical Gaussian face mask, $W_{face}(x,y) = 2.0 \cdot \exp(-2.0 \cdot d^2)$, centered on the key facial structures; HOG similarity S_{hog} is the dot product of the two resulting unit vectors.

4.3 Late Score Fusion

The two similarity scores are combined with a simple linear rule, $S_{hybrid}(q,g) = \alpha \cdot S_{deep}(q,g) + (1-\alpha) \cdot S_{hog}(q,g)$; grid search on the held-out validation split settled on $\alpha^* = 0.05$ as the production operating point.

5. EXPERIMENTAL SETUP

5.1 Dataset and Protocol

Testing draws on the CUFS database 189 gallery photographs and 190 query sketches split by identity into 60 training, 20 validation, and 21 held-out test pairs, with no identity appearing in more than one split, a division checked against the dataset's official split manifest.

5.2 Evaluation Metrics

Results are reported as Rank-1 and Rank-5 identification accuracy plus ROC-AUC, across three gallery sizes: the full 189-photo gallery, a 109-photo test-pool subset, and a 100-pair subset that matches the protocol commonly used in earlier sketch-recognition work. Latency is taken as the median of thirty repeated queries run warm on an ordinary central processing unit (CPU)-only machine. The hybrid rows in Table -2 are computed on the 21 held-out test queries, whereas raw FaceNet is computed on all 190 queries; the Rank-1 gain over raw FaceNet is therefore indicative, and a same-query comparison is left for future work. The 100-pair protocol is included for comparability with earlier work and may contain identities seen during training, so it is not a strict held-out result.

6. RESULTS AND DISCUSSION

Table -2 summarizes Rank-N accuracy across the evaluation protocols.

Table -2: Rank-N accuracy across evaluation protocols

Protocol	Gallery	Rank-1	Rank-5	AUC
Raw FaceNet (190 queries)	189	12.11%	25.79%	0.684
Baseline hybrid	189	71.43%	100.0%	0.981
Optimized hybrid (α=0.05)	**189**	**85.71 %**	**95.24 %**	**0.990**
Test-pool protocol	109	90.48%	95.24%	0.991
Standard 100-pair	100	92.00%	98.00%	0.993

These numbers make clear that a deep embedding on its own is not enough for cross-modal matching raw FaceNet manages only 12.11% Rank-1 accuracy because its features are tuned for photographic texture, not sketch outlines. Adding the trained projection head and fusing it with the spatially weighted HOG descriptor pushes Rank-1 accuracy up to 85.71% on the full 189-candidate gallery, a jump of more than seventy percentage points, while the baseline configuration still holds a 100.00% Rank-5 rate. Accuracy climbs further as the gallery shrinks 90.48% at 109 candidates, 92.00% on the 100-pair protocol which tracks with the expected trend that a smaller candidate pool is an easier identification problem. Median warm-inference latency comes in at 307.9 ms, fast enough for interactive investigative use.

That the optimal fusion weight settles so low ($\alpha^* = 0.05$) says most of the discriminative power here comes from the spatially weighted HOG stream, with the learned deep embedding acting more as a refinement on top of it. In other words, for a forensic gallery this small, a well-engineered spatial descriptor can hold its own against and even outperform a learned metric embedding.

7. CONCLUSION AND FUTURE SCOPE

ThirdEye v2, as described in this paper, is a complete forensic pipeline pairing a JavaFX composite-sketch client with a FastAPI microservice that fuses a deep metric embedding with a spatial HOG descriptor for cross-modal recognition. On the CUFS benchmark it reaches 85.71% Rank-1 accuracy against a 189-candidate gallery and 92.00% on a 100-pair protocol, a clear improvement over raw FaceNet, and stays fast enough sub-second per match for field use. Its current limitations are restricted access to additional sketch-photo datasets (CUFSF, IIIT-D) pending academic licensing, and weaker performance on faces turned more than roughly 15° off-axis. Planned next steps include 3-D landmark-based pose normalization, using diffusion-based sketch-to-photo synthesis as a pre-matching stage, and encrypting on-device suspect records for real operational deployment.

ACKNOWLEDGEMENT

The authors are grateful to their project guide and to the Department of Computer Science and Engineering for

their guidance and for the infrastructure that made this work possible.

REFERENCES

[1] N. Sádaba-Campo and H. Gómez-Moreno, "Exploration of Generative Neural Networks for Police Facial Sketches," Big Data and Cognitive Computing, 2025.

[2] S. Pushpalatha, Shashank V, Shreyank K. J., Sri Ranga B. K., and Pavan T. M., "Forensic Face Sketch Construction and Recognition," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 13, issue XII, Dec. 2025.

[3] K. Pasunuri, C. Rohan, G. Vaishnavi Reddy, and T. Sai Sree, "AI-Driven Forensic Face Sketch Construction and Recognition," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE 2025), Advances in Computer Science Research, vol. 124, 2025, doi: 10.2991/978-94-6463-858-5_231.

[4] D. A. Itole, M. P. Sardey, and M. P. Gajare, "Heterogeneous Sketch-Face Photo Recognition in Forensic Science Laboratories," International Journal of Electrical and Computer Engineering (IJECE), 2024.

[5] N. S. Shaikh, D. Wagh, S. Takawale, P. Singh, and A. Jadhav, "RecogniPro: Recognition and Construction of Forensic Facial Sketches," International Journal of Creative Research Thoughts (IJCRT), vol. 12, no. 3, Mar. 2024.

[6] A. Mohan, S. Matekar, and P. Itankar, "CHEHRA: An Application for Forensic Face Sketch Construction and Recognition," International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), vol. 2, no. 2, Apr. 2022, doi: 10.48175/IJARSCT-3179.

[7] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.

[8] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 11, pp. 1955–1967, 2009.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR), vol. 1, 2005, pp. 886–893.

[10] C. Galea and R. A. Farrugia, "Forensic face photo-sketch recognition using a deep learning-based architecture," IEEE Signal Process. Lett., vol. 24, no. 11, pp. 1586–1590, 2017.

[11] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, 2002.