International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Verbal-to-Visual AI-Powered Forensic Face Reconstruction

Suresh Kumar C[1]
Dept. of CSE
Don Bosco
Institute of Technology
Bengaluru, India Assistant
Professor
suresh.acgiri@gmail.com

Vikram Singh V[2]
Dept. of CSE
Don Bosco
Institute of Technology Bengaluru,
India
1DB23CS243
mr.vikram.singh170@gmail.com

Darshan M[3]
Dept. of CSE
Don Bosco
Institute of Technology
Bengaluru, India
1DB23CS405
darshanvikram3@gmail.com

Irfan Ali Baig D S[4]
Dept. of CSE
Don Bosco
Institute of Technology
Bengaluru, India
1DB23CS410
irfanalibaigds@gmail.com

Md Arshad[5]
Dept. of CSE
Don Bosco
Institute of Technology
Bengaluru, India 1DB23CS415
ma8147500767@gmail.com

Abstract

Forensic face reconstruction is crucial for conducting criminal investigations and identifying missing persons and unknown individuals. However, conventional reconstruction methods mainly rely on sketching by forensic artists, making them highly subjective and time-consuming. The paper discusses "Verbal to Visual: AI-Powered Forensic Face Reconstruction" which is a novel approach to automating the process of face reconstruction by means of Speech Recognition, Natural Language Processing, Deep Learning, Computer Vision and Facial Recognition. The system takes verbal descriptions of such facial traits as age, gender, skin color, facial shape, eyes, nose, hairstyle, and other unique features as inputs. Verbal description is transformed into textual format and analyzed via Natural Language Processing techniques to extract information related to facial attributes. Finally, attributes are provided to text-to-image generation algorithms implemented via AI technologies and the facial reconstruction process is performed automatically. The generated image is compared against facial images contained in a database with the use of facial similarity matching techniques and ranked matching results are produced to help in the investigation process.

I. INTRODUCTION

Forensic face reconstruction is a method in forensic science. Forensic face reconstruction helps estimate what a person looks like when only bones or incomplete evidence remain. Traditional reconstruction relies a lot on experts, knowledge of anatomy and manual or digital art. Even though these methods help they take a time can be subjective and depend on how experienced the person is. Because of these limits there is a need for technology-based tools to aid forensic work.

Recent progress in Artificial Intelligence, Natural Language Processing, Computer Vision and generative AI lets machines turn words into clear pictures. This ability is especially helpful in cases when witnesses or investigators give spoken descriptions of a person's face.

This paper introduces "Verbal to Visual: AI-Powered Forensic Face Reconstruction " a tool that turns spoken or written descriptions into AI-created faces. The tool looks at details such as age, gender, skin tone, facial shape, hairstyle and unique features to build a matching picture. By combining speech recognition, natural language processing, computer vision and generative AI the tool offers an automated method, for facial reconstruction. The tool can help police and forensic investigators quickly spot missing people, unknown victims and suspects while cutting the time and manual work needed by reconstruction.

II. RELATED WORK

Research in forensic face reconstruction and AI-based facial image synthesis has examined sketch-to-photo matching, text-to-image generation and image-to-image translation. Several researchers have explored applying computer vision, machine learning, and generative intelligence techniques to facial reconstruction, facial recognition, and image synthesis. These approaches have demonstrated that artificial intelligence can significantly improve the automation, visual quality and efficiency of facial analysis systems. However, most existing methods address stages of the reconstruction process and do not provide an integrated system capable of directly converting natural verbal descriptions into facial representations. The following studies provide the foundation, for the proposed verbal-to-visual forensic face reconstruction approach.

I find that Han and colleagues [1] described a component-based method that matches sketches to real facial photographs. Their method uses landmarks, multiscale local binary patterns and component-level similarity matching. The result was a rank-100 accuracy of 77.2 % and 89.4 % on two databases showing that forensic sketch matching can be very effective.

Xu and colleagues [2] presented AttnGAN a text-to-image generation framework that uses attention to link each word in a description with a part of an image. The model produced image details and achieved a 14.14 % increase on the CUB dataset and a 170.25 % increase on COCO, measured by Inception Score compared with earlier methods.

Isola and colleagues [3] offered pix2pix, a GAN that translates one image style into another. The method worked well for turning label maps into photos edge maps into objects and other visual changes.. It mainly handles image-to-image translation and does not turn natural-language descriptions into facial pictures.

Colleagues [4] introduced a Composition-Aided GAN (CA-GAN) for realistic face photo-sketch synthesis. Their method uses composition data, reconstruction loss and perceptual loss to keep the structure and identity of faces intact. It lowered the FID from 36.2 to 26.2 for sketch synthesis and from 60.9 to 30.5, for photo synthesis.

Ref.	Approach	Accuracy	Limitation
[1]	Compares forensic sketches with photographs by breaking faces into individual components and measuring feature-level similarity.	**Rank-100 accuracy: 77.2% and 89.4%** on two mugshot databases.	Requires an already-created sketch and access to a facial database; cannot generate a face directly from a verbal description.
[2]	**AttnGAN** uses attention mechanisms to associate words in text descriptions with corresponding image regions for fine-grained text-to-image generation.	**Inception Score improvement: 14.14% on CUB and 170.25% on COCO** compared with previous methods.	General-purpose text-to-image model; not specifically designed for forensic face reconstruction or accurate identity preservation.
[3]	**pix2pix / Conditional GAN** performs image-to-image translation by learning a mapping between paired visual representations.	Demonstrated effective results across multiple image-to-image translation tasks; **no single accuracy value reported**.	Requires an input image and therefore cannot directly convert text or speech descriptions into a facial image.
[4]	**CA-GAN** incorporates facial composition information along with reconstruction and perceptual losses for realistic photo-sketch synthesis.	**FID: 26.2 for sketch synthesis and 30.5 for photo synthesis**, improving over previous methods.	Limited to photo-sketch synthesis and requires visual input; cannot directly process natural-language or audio descriptions.

Table I: Summary of Related Work

III. METHODOLOGY

This system uses a sequential approach to turn the description of a suspect by a user into a face image of the suspect, which is then compared against other images present in the forensic database. The entire process comprises of speech recognition, speech to text conversion, face image generation through AI technology, face image matching with other images in the database, selection of matches and face comparison in detail. Every step uses the output of the preceding step.

A. Speech Input

The whole process starts with the user giving a verbal description of the individual in question through the speech input module. This includes descriptions of some of the features of his face, including age, gender, skin color, facial structure, eyes, nose, hair style, and any other distinct features. The speech recognition component gets the user's voice and processes it as the first step in feeding the input into the system. In our implementation of the speech recognition module, the audio input is obtained through the microphone, and Speech Recognition library, with Google Speech Recognition, is used to convert the speech into text form.

B. Text Processing and Prompt Formation

Once the textual transcription process is complete, the recognized textual transcription is fed to the face generation unit. This part of the process does not use any pretrained NLP model either. What the text is fed to is a prompt generation unit that generates a forensic image generation prompt from the provided facial textual description. The generated prompt contains information on factors like facial appearance, expression, viewing direction, and even the style, with a realistic forensic police sketch representation done using graphite pencil. The generated prompt is then sent to the selected text-to-image generation model using the AI generation pipeline.



Fig 1: Text Processing and Prompt Formation

C. AI-Based Face Image Generation

After text processing and prompt formation, the generated descriptive prompt is passed to an AI-based image generation module to reconstruct a facial image corresponding to the verbal description. The proposed system employs a multi-model generation strategy to improve reliability and maintain image generation even when a particular service is unavailable. The system initially attempts image generation using **DALL-E 3**, followed by **Google Imagen**, **Craiyon**, **Pollinations**, **Segmind Stable Diffusion 1.5**, **DeepAI**, **Prodia SDXL**, and **Hugging Face Stable Diffusion models** as fallback alternatives. Each model receives the processed textual prompt and generates a facial image based on the described attributes such as facial structure, hairstyle, skin tone, age, gender, and other distinguishing characteristics. The generators are accessed through their respective APIs or inference interfaces, depending on service availability. If the primary model fails due to API errors, service unavailability, rate limitations, or generation failure, the system automatically proceeds to the next available model in the predefined fallback sequence. This sequential fallback mechanism increases the robustness and availability of the proposed system by reducing dependency on a single generative AI service. The successfully generated image is then forwarded to the subsequent face-matching module, where it can be compared against the available facial image database for identification.



Fig 2:AI -Based Face Image Generation

D. AI-Generated Face Matching and Identification

The generated facial image by the AI-based image generation module is further processed by the face matching and identification module, which aims to find the most similar person in the existing set of facial images. In particular, the generated image and all the images in the database are first preprocessed by resizing them, converting into grayscale color scheme and enhancing the image quality to ensure consistency when comparing them. Then, the system assesses the similarities between the generated image and other images by means of several image-comparison approaches, such as Structural Similarity Index Measure (SSIM), Normalized Cross-Correlation (NCC), histogram correlation, histogram-based distance measures, histogram intersection and ORB-based feature matching. The individual similarity scores are calculated using these approaches and are later aggregated by means of a weighted score aggregation approach. The resulting overall similarity scores of each of the images are used to rank them and select the one that surpasses the matching threshold as the most similar image. Using this multi-metric approach allows for achieving more reliable identification outcomes as it considers the global and local features of the images simultaneously instead of using only one comparison metric.



Fig 3: Database Matching sketch

IV. RESULTS AND DISCUSSION

I evaluated the proposed Verbal to Visual AI-Powered Forensic Face Reconstruction system by taking spoken descriptions and turning them into pictures of faces. The speech-to-text module successfully converted the input provided by the user into textual information, allowing the proposed Verbal to Visual AI-Powered Forensic Face Reconstruction system to capture important facial attributes from the description. The extracted information included characteristics such as age, gender, skin tone, facial structure, eyes, nose and hairstyle, which were subsequently used for reconstruction.

The processed textual description was then fed to the AI-based text-to-image generation module to produce a forensic facial reconstruction. The generated facial image closely matched the characteristics specified in the description, showing that the proposed Verbal to Visual AI-Powered Forensic Face Reconstruction system can transform natural-language descriptions into visual facial representations. This provides an automated alternative to manually creating a facial representation from witness descriptions.

The reconstructed facial image was further examined using the face matching module. The generated image was compared with images stored in the database using facial similarity matching techniques. The proposed Verbal to Visual AI-Powered Forensic Face Reconstruction system successfully performed comparisons with the database records and produced matching results based on the similarity between the reconstructed face and stored images.

The matching results were presented in a ranked form so the investigator could consider the similar database images first. This ranking helps reduce the time needed to examine a large number of candidate images. The use of Computer Vision and Deep Learning techniques also improved the clarity of the generated representation and the overall reconstruction quality.

The overall results demonstrate the feasibility of combining speech recognition, NLP, AI-based image generation and face matching into a forensic face reconstruction system. The developed system can transform witness descriptions into facial representations and then use those representations to search for potential matches within a facial database. The proposed Verbal, to Visual AI-Powered Forensic Face Reconstruction system is intended to support investigators during analysis and candidate identification rather than serve as an independent final identification mechanism.

V. FUTURE ENHANCEMENT

The proposed system can be extended by adding 3D face reconstruction and real-time face generation which will help to generate more detailed and interactive forensic face representations. We may also add multilingual speech support to enhance accessibility for witnesses and investigators from different locations.

Future work can focus on further integrating larger criminal databases and training the AI models with larger and diverse datasets. Such improvements may lead to better accuracy of reconstruction, reduce possible bias and improve reliability of face matching for practical forensic applications.

VI. CONCLUSION

The proposed Verbal to Visual AI-Powered Forensic Face Reconstruction system demonstrates the feasibility of utilising Artificial Intelligence for translating verbal descriptions into visual face representations. The system combines Speech Recognition, Natural Language Processing, AI-based image generation, Computer Vision, and face matching to provide an automated approach for forensic face reconstruction and candidate searching from the database.

The results show the ability of the system to generate facial representations from verbal descriptions and match them with stored database images to provide ranked matching results. Thus, the proposed

system can help forensic investigators in suspect visualization and identification. The system is a decision support tool, not an identification system.

VII. ACKNOWLEDGMENT

The authors extend their sincere gratitude to Suresh Kumar C, Assistant Professor, Department of Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, for his continuous guidance, valuable suggestions, encouragement, and constant support throughout the development of this project. His technical expertise, constructive feedback, and insightful recommendations greatly contributed to the successful completion of this work.

The authors sincerely thank the Department of Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, for providing the necessary infrastructure, resources, and technical support for this project. Special thanks to the faculty members and peers for their valuable guidance, cooperation, and assistance throughout the project.

Finally, the authors sincerely thank their families for their constant support and encouragement throughout this work.

VIII. REFERENCES

[1] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain, "Matching Composite Sketches to Face Photos: A Component-Based Approach," IEEE Transactions on Information Forensics and Security, vol. 8, no. 1, pp. 191-204, 2013, DOI: 10.1109/TIFS.2012.2228856.

[2] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2018, pp. 1316–1324, 2018, DOI: 10.1109/CVPR.2018.00143.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 5967-5976, Jul. 2017. DOI: 10.1109/CVPR.2017.632.

[4] J. Yu, X. Xu, F. Gao, S. Shi, M. Wang, D. Tao, and Q. Huang, "Towards Realistic Face Photo-Sketch Synthesis via Composition-Aided GANs," IEEE Trans. Cybern., vol. 51, no. 9, pp. 4350–4362, 2021, DOI: 10.1109/TCYB.2020.2972944.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," Advances in Neural Information Processing Systems (NeurIPS), vol. 27, 2014.

[6] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 11, pp. 1955–1967, 2009.

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks," in 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2223–2232, 2017.

[8] T. Karras, S. Laine, and T. Aila, "A Style Based Generator Architecture for Generative Adversarial Networks", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4401–4410, 2019.

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695, 2022.

[10] A. Radford, J. W. Kim, *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in Proceedings of the 38th International Conference on Machine Learning *(ICML)*, vol. 139, pp. 8748–8763, 2021.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823, 2015. DOI: 10.1109/CVPR.2015.7298682.

[12] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1701–1708, 2014.

[13] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition. (CVPR), pp. 4690–4699, 2019.

[14] O. M. Parkhi, A. Vedaldi and A. Zisserman, "Deep Face Recognition", Proceedings of the British Machine Vision Conference (BMVC), pp. 41.1-41.12, 2015.

[15] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An Efficient Alternative to SIFT or SURF," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2564–2571, 2011, DOI: 10.1109/ICCV.2011.6126544.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.

[17] D. Amodei et al., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," in Proceedings of the 33rd International Conference on Machine Learning (ICML), vol. 48, pp. 173–182, 2016.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT, pp. 4171–4186, 2019, DOI: 10.18653/v1/N19-1423.