International Journal of Computing and Artificial Intelligence

E-ISSN: 2707-658X
P-ISSN: 2707-6571
Impact Factor (RJIF): 5.57
www.computersciencejournals.com/ijcai
IJCAI 2025; 6(2): 139-143
Received: 23-06-2025
Accepted: 26-07-2025

Ashik Kumar
SRU, Bagar Rajput,
Rajasthan, India

Corresponding Author:
Ashik Kumar
SRU, Bagar Rajput,
Rajasthan, India

Multi modal based realistic synthetic media generation using AFMT-ALKM and 3GPAN

Ashik Kumar

DOI: https://www.doi.org/10.33545/27076571.2025.v6.i2b.190

Abstract

The generative adversarial network (GAN) framework has emerged as a powerful tool for various image, video, and audio synthesis tasks, allowing the synthesis of visual content in an unconditional or input-conditional manner. It has enabled the generation of high-resolution photorealistic images and videos, which is a more challenging or impossible task than prior methods. It has also led to the creation of many new applications in content creation. In this article, an overview of GANs is provided with a special focus on algorithms and applications for visual synthesis. Also, several important techniques are covered to stabilize GAN training, which has a reputation for being notoriously difficult. Here, GAN's applications, such as image translation, image processing, video synthesis, and neural rendering are also discussed. Therefore, the proposed model will be developed for Realistic Synthetic Media Generation for Multi-Model, including Images, Videos, and Music using 3GPAN and AFMT-ALKM. Initially, the Audio, Video, and Image datasets will be taken, and pre-processing steps, such as frame conversion, noise removal, contrast enhancement, and background subtraction will be done for three datasets. Next, the object detection will be done by using YOLO, and Motion will be estimated using LET-BMA. Next, the Audio Features, Video Features, Image Features, Appearance Features, Action Features, Emotion Features, and 3D model will be given to the Realistic Media Generation phase for training using 3GPAN. Finally, AFMT-ALKM-based rendering will be done for the retrieved 3D model, and the proposed model will be compared with various models using the result parameters like MSE, RMSE, MAPE, etc.

Keywords: Computer vision, generative adversarial networks (GANs), image, video, and audio synthesis; image processing, rendering, realistic media generation, synthetic human creation, voice synchronization, dynamic environments

1. Introduction

The advancement in Artificial Intelligence (AI) and the growing interest in media as well as entertainment leads to generative video technology. This synthetic generation of video is comprised of animation videos, character videos, special effects videos, and so on. By creating synthetic human-like entities using Machine Learning (ML) and computer graphics technology, a realistic media is generated, which mimics the specific human behavior and appearance. Synthetic video plays a significant role in various social fields, including education, entertainment, and customer services. The virtual human models and the associated voice interactions formulate the virtual teacher models in the smart education system. Further, in the entertainment domain, diverse virtual characters for the gaming and film industry can be created with the efficient generation of synthetic humans. Also, virtual customer service through the realistic media creation rapidly responds to the customers, resulting in improved communication services. In recent days, numerous research has emerged for human video generation mainly focusing on the quality, efficiency, and interactivity of the video. For developing realistic human videos, Deep Learning (DL) based techniques like Generative Adversarial Networks (GAN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Autoencoders, and Deep Reinforcement learning models are highly utilized. Among them, CNN effectively recognizes the synthetic human actions and GAN provides excellent performance in realistic video synthesis. However, the existing research did not concentrate on the complex dynamic scenes and the gender-related features, causing ineffective and mismatched generation of synthetic human data. Therefore, this paper proposes an efficient realistic data generation by focusing on the aforementioned issues by using 3GPAN and AFMT-ALKM techniques.

2. Literature Survey

Authors	Objective	Method	Dataset	Result	Advantage	Limitation
(Wendland *et al*., 2022) [5]	Realistic synthetic data generation	Multimodal Neural Ordinary Differential Equations	Synthetic patient data	Achieved higher performance regarding progression score and sample size.	Highly realistic synthetic data trajectories were created on a continuous time scale.	However, the differential equations and multiple modalities increased the computation complexity.
(Li *et al*., 2024) [2]	Constrained keyframe-based Virtual Human video generation	Multimodal information fusion strategy and interaction framework	Voxceleb2 dataset and Flickr Faces High Quality (HQ) dataset	Achieved higher Structural Similarity Index (SSIM) and Beat align score.	The stability of the cross-mixed frames in synthetic videos was ensured by the approach.	Yet, the intricate dynamic background of the data was not analyzed, thus lowering the quality of realistic video.
(Varol *et al*., 2021) [3]	Human action recognition of synthetic humans	Spatial-temporal CNN model	Synthetic hUmans foR REal ACTions (SUR-REACT) dataset	Achieved enhanced performance in terms of video accuracy.	The presented approach had improved generalizability to unseen viewpoints.	However, a large amount of data samples were required and the video events were not properly captured, thus affecting the performance.
(Hwang *et al*., 2023) [1]	Synthetic data generation for human action recognition	ElderSim model	Real world data based KIST SynADL dataset.	Attained higher accuracy.	The human actions were exactly recognized by the augmentation of synthetic data.	Nevertheless, the gender-based biases were not concentrated, thereby degrading the system's effectiveness.
(Wang *et al*., 2024) [4]	Generative Latent image animator-based Human video synthesis	Latent Motion Diffusion model	TaichiHD, FaceForensics, and CelebV-HQ datasets	Obtained improved performance regarding Kernalized Video Distance (KVD).	The global and local deformations of the video were analyzed by expressing motion as a flow map sequence.	But, the presented model was suscepted to overfitting issues, which influenced the data training.

3. Problem Statement

Existing research methodologies have some drawbacks, which are described as follows,

- None of the prevailing models concentrated on verification and improvement of video-audio-scaling-position variation during synchronization in Synthetic human generation.
- [3], Most of the existing works were ineffective on synthetic human generation by considering background. Thus, it was challenging to analyze the actual object from the media.
- The work [1] performed the realistic video generation for multimedia data. But, it failed to select the appropriate model for realistic media generation.
- The work [2] mainly concentrated on generating a realistic video. But, it failed to concentrate on biases (appearance, emotion, activity, etc.), which could be reflected in the generated media. This could lead to misrepresentations or reinforcement of stereotypes.
- Some of the traditional works were less significant owing to the lack of effective audio, image, and video pre-processing schemes. The raw media had poor recording quality and environmental disruptions.

4. Significance of the study

Image, video and audio synthesis are closely related areas aiming at generating content from noise. While rapid progress has been demonstrated in improving image based models to handle large resolutions, high-quality renderings, and wide variations in image content, achieving comparable video generation results remains problematic.

The central contribution of this research is as follows,

- To introduce novel pre-processing steps to transform the image data into a format that is more easily and effectively processed in image processing.
- To introduce efficient features to differentiate the correlation between the data for better classification.
- To introduce an object detection technique to segregate the object from the image. It is a computer vision task that involves identifying and tracking objects of interest within a video sequence. Video object detection tasks incorporate temporal information to keep track of the objects across different video sequence frames.
- To introduce efficient motion estimation model to analyze and predict the movement of objects in a sequence of video frames.
- To introduce modified Generative Adversarial Networks to perform efficient Realistic Synthetic Media Generation for Multi Model.
- To introduce efficient rendering to generate a realistic video image from 3D model or other input data.
- To compare the proposed technique with the existing technique using the result parameters like Error, Bias, Mean Error (ME), Mean absolute error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE), Root Mean Squared Scaled Error (RMSSE), Relative Absolute Error (RAE), Prediction Accuracy, R-squared, Adjusted R squared for demand forecasting, fitness vs. iteration, clustering time, etc.

5. Objective of the study

The main aim of this paper is to architect a Realistic Synthetic Media Generation of a Multi-Model, including Images, Videos, and Music using 3GPAN and AFMT-ALKM. The objectives of this research work are enlisted as follows,

- To perform Background Subtraction, the Gaussian Root Wendland Kernel Mixture Model (GRWK2M) technique will be used.
- To estimate motion on video, the Log-Easom

Transform-Block Matching Algorithm (LET-BMA) will be used.

- To generate Realistic Synthetic Media, a Gradient Gish Generative Penalty Adversarial Network (3GPAN) will be used.
- To perform an efficient Rendering process, an Affine Fourier-Mellin Transform-based Adaptive Lucas-Kanade Method (AFMT-ALKM) will be introduced.
- To compare the proposed technique with the existing technique using the result parameters like MAE, MASE, RMSE, MAPE, MSE, RAE WAPE, etc.

6. Hypothesis/research question

Following is the hypothesis or research question that is familiarity with the subject that helps define an appropriate research question for a subject area or field of study.

- What are the ethical implications of using AI-generated synthetic media in various domains?
- Can AI models generate synthetic video content indistinguishable from real-world footage across appearance, action and emotion?
- What are the most effective techniques for detecting AI-generated synthetic media?
- How can advances in synthetic media generation be used to enhance virtual reality (VR) experiences?
- Exposure to realistic synthetic media will increase the difficulty of distinguishing between real and fake news, leading to higher susceptibility to misinformation.
- AI-generated deepfake videos that accurately mimic human expressions and voices will evoke similar emotional responses to real videos in viewers.
- Advancements in synthetic media detection algorithms will significantly reduce the spread of deepfakes on different domains.

7. Methodology

Human video generation is a dynamic and rapidly evolving task that aims to synthesize 2D human body video sequences with generative models and control conditions, such as video, audio, and image. With the potential for wide-ranging applications in film, gaming, and virtual communication, the ability to generate natural and realistic human video is critical. Recent advancements in generative models have laid a solid foundation for the growing interest in this area. Despite the significant progress, the task of human video generation remains challenging due to the consistency of synchronization in audio with video, the complexity of human motion, and difficulties in their relationship with the environment.

To create high-resolution accurate and realistic 3D models, the proposed framework will be designed. The proposed system will start from the multi datasets, such as video dataset, audio dataset, and image dataset. Next, the Audio pre-processing that is noise removal will be done by using a median filter. After that, audio features, such as Spectrogram, Mel-Frequency Cepstral Coefficients (MFCC), Mel-filter bank Energy (MFE), Spectral Centroid, Spectral Contrast, Rhythm, Beats, and Tempo will be extracted. Next, the video dataset will be taken; from that dataset, the audio will be separately extracted. Next, the Audio pre-processing step will be done, and audio features will be extracted. Meantime, the Video pre-processing step that is frame conversion will be done. Next, the sampling rate and frame rate will be extracted and checked for synchronization correlation using the Pearson correlation technique. The image dataset will be taken and Image pre-processing steps, such as Noise Removal using a Median Filter, Contrast Enhancement using Contrast Limited Adaptive Histogram Equalization (CLAHE), and Background Subtraction using the Gaussian Root Wendland Kernel Mixture Model (GRWK2M) will be carried out. Here, the Gaussian Mixture Model (GMM) eliminated the background properties of the image. The GMM is a flexible probabilistic model that can capture complex distributions of pixel intensities in an image of diverse backgrounds. Also, it can represent both foreground and background regions with multiple Gaussian distributions, allowing for more accurate modeling. Yet, it had suboptimal solutions due to the random initialization of the initial parameters of the Gaussian distributions for each component. Hence, the Root Wendland kernel function is incorporated to initialize the Gaussian distribution parameters. The root Wendland kernel function offers a balanced approach between linearity and non-linearity, making it a potentially useful choice for parameter initialization. Thus, the proposed work is named as Gaussian Root Wendland Kernel Mixture Model (GRWK2M).

After that, Objects will be detected from the pre-processed video frames and images using the You Only Look Once (YOLO) algorithm. After that, the motion will be estimated only from the video frames using the Log-Easom Transform-Block Matching Algorithm (LET-BMA). Here, the Block Matching Algorithm (BMA) is used to estimate the motion of the moving objects. The BMA allows for the use of different block sizes, enabling a balance between accuracy and computational efficiency, which is crucial for varying vehicle sizes and speeds. However, it had high estimation errors because it randomly selected the blocks when trying to match with subsequent frames. Therefore, the proposed method introduces the Log-Easom Transformation to select the blocks. Thus, the proposed work is named as Log-Easom Transform-Block Matching Algorithm (LET-BMA). Next, the Body Skeleton will be constructed on pre-processed video frames and images. Next, the Video Features, Image Features, Appearance Features, Action Features, and Emotion Features will be extracted from pre-processed video frames and images. Finally, the 3D model will be constructed for the detected objects. At the end, Audio Features, Video Features, Image Features, Appearance Features, Action Features, Emotion Features, and 3D models will be given to the Realistic Media Generation phase for training using Gradient Gish Generative Penalty Adversarial Network (3GPAN). Here, GANs are capable of generating high-resolution and realistic data, making them ideal for applications like image synthesis and video generation. Also, GANs can overfit the training data, leading to poor generalization of new, unseen data. In order to solve this issue, the Gradient Penalty technique will be included. To solve low learning efficiency, the Gish Activation function will be used.

At the end of the training phase, an Image, Video, or Image will be given to the Realistic Media Generation model. Before that, the input media attributes will be extracted and the type of media will be identified using the If then condition. Next, the Realistic Media Generation model gives a highly accurate 3D model, and it will be rendered using the Affine Fourier-Mellin Transform-based Adaptive Lucas-

Kanade Method (AFMT-ALKM). One of the most significant advantages of the Fourier-Mellin Transform (FMT) is its invariance to scale and rotation. This means that when an image is resized or rotated, the FMT can still recognize and match the image features, making it useful for applications like image recognition and classification. FMT is primarily effective for linear transformations like scaling and rotation. It may not perform well with non-linear transformations (e.g., perspective distortions), which can limit its applicability in some scenarios. In order to solve this issue, the Affine transform technique will be included. Lucas-Kanade Method uses local averaging to compute flow estimates, which helps to reduce the impact of noise in the input images. This robustness makes it effective in real-world scenarios, where images may be noisy. The performance of the Lucas-Kanade method can be sensitive to the choice of the window size used for local motion estimation. If the window is too small, then it may not capture enough information; if it's too large, then it may include irrelevant data. In order to solve this issue, adaptive window size will be used based on local image properties. The block diagram of the proposed framework is depicted below in Figure 1.

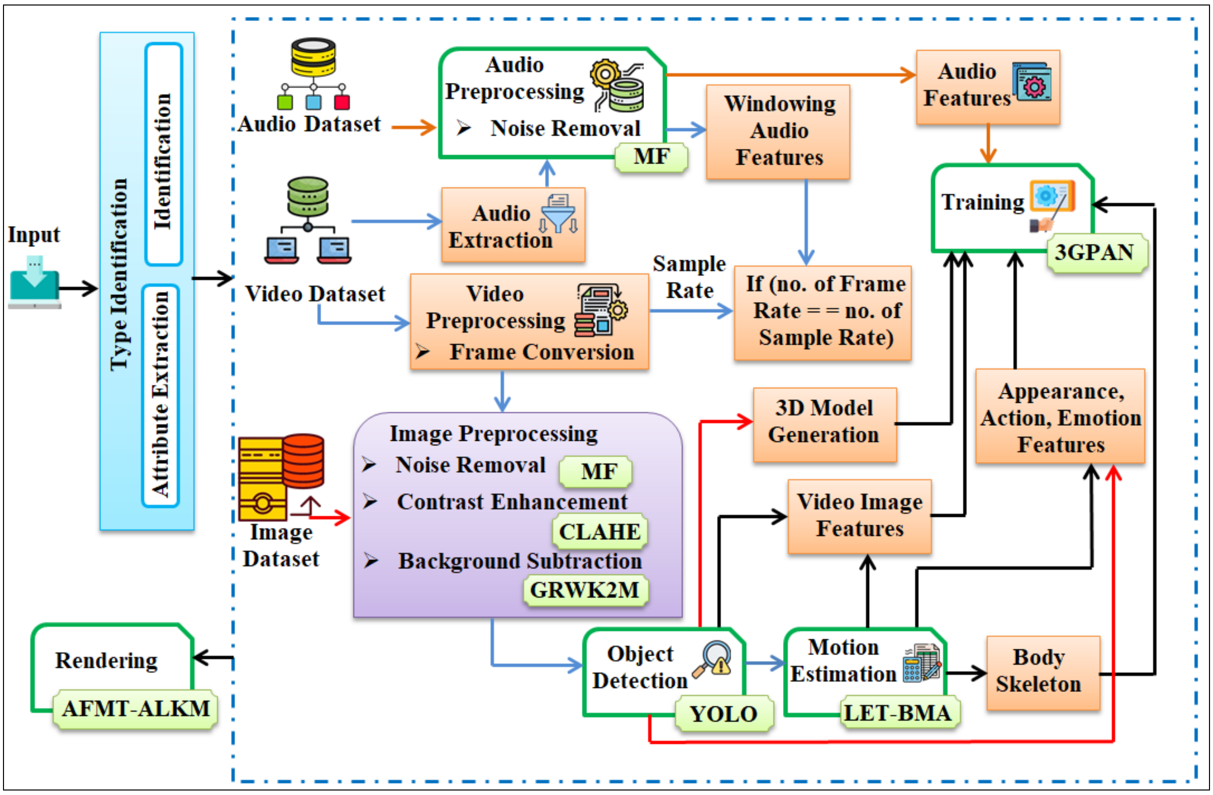

Fig 1: Block diagram of the proposed model

8. Data collection

8.1 Dataset used

The dataset used in the proposed framework is “Celeb-DF”, “FF++”, “NVlabs” and “voxceleb” from publically available Resources. This dataset is collected from publically available sources using the below link.

https://github.com/NVlabs/ffhq-dataset

https://github.com/YZY-stack/DF40/tree/main/Annotations_for_Facial_Attributes

https://huggingface.co/datasets/ProgramComputer/voxceleb/tree/main/vox1

8.2 Software requirement

The proposed work is implemented in the working platform of PYTHON. Python is a widely used general-purpose high-level programming language. It was mainly developed to emphasize code readability, and its syntax allows programmers to express concepts in fewer lines of code. Python is a programming language that lets you work quickly and integrate systems more efficiently. The virtual environment tool creates an isolated Python environment (in the form of a directory) that is completely separated from the system-wide Python environment.

8.3 Hardware requirements

- **Processor:** Intel i5/core i7
- **CPU Speed:** 3.20 GHz
- **OS:** Windows 10
- **System Type:** 64 bit
- **RAM:** 8GB

9. Proposed Analysis

The proposed technique is evaluated based on the performance metrics, such as Error, Bias, Mean Error (ME), Mean absolute error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE), Root Mean Squared Scaled Error (RMSSE), Relative Absolute Error (RAE), Prediction Accuracy, R-squared, Adjusted R squared for demand forecasting. These metrics evaluate the Realistic Synthetic Media Generation performance by comparing the proposed model with some other existing techniques like, Generative Adversarial Networks (GANs), Autoencoders and Variational Autoencoders (VAEs), Neural Style Transfer, Deepfakes and Neural Rendering. Finally, the related literature comparison will also be carried out to prove the accuracy of the proposed model.

References

1. Hwang H, Jang C, Park G, Cho J, Kim IJ. ElderSim: A synthetic data generation platform for human action recognition in eldercare applications. IEEE Access. 2023;11:9279–9294. doi:10.1109/ACCESS.2021.3051842
2. Li P, Li X, Yin L. Research and implementation of constrained keyframe-based virtual human video generation. IEEE Access. 2024;12:137089–137104. doi:10.1109/ACCESS.2024.3465469
3. Varol G, Laptev I, Schmid C, Zisserman A. Synthetic humans for action recognition from unseen viewpoints. International Journal of Computer Vision. 2021;129(7):2264–2287. doi:10.1007/s11263-021-01467-7
4. Wang Y, Ma X, Chen X, Chen C, Dantcheva A, Dai B, Qiao Y. Leo: Generative latent image animator for human video synthesis. International Journal of Computer Vision. 2024;1–16. doi:10.1007/s11263-024-01890-7
5. Wendland P, Birkenbihl C, Gomez-Freixa M, Sood M, Kschischo M, Fröhlich H. Generation of realistic synthetic data using multimodal neural ordinary differential equations. NPJ Digital Medicine. 2022;5(1):1–10. doi:10.1038/s41746-022-00666-x
6. Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations (ICLR); 2019. Available from: https://openreview.net/forum?id=B1xsqj09Fm