Demo Abstract: AD-CLIP: Privacy-Preserving, Low-Cost Synthetic Human Action Dataset for Alzheimer's Patients via CLIP-based Models

Heming Fu
hemingfu@link.cuhk.edu.hk
The Chinese University of Hong Kong, Hong Kong SAR

Hongkai Chen
hkchen@ie.cuhk.edu.hk
The Chinese University of Hong Kong, Hong Kong SAR

Guoliang Xing
glxing@cuhk.edu.hk
The Chinese University of Hong Kong, Hong Kong SAR

Abstract

With the increasing demand for smart health applications that emphasize privacy and efficiency, we introduce AD-CLIP, a synthetic data generation framework using CLIP-based models for Alzheimer's patients. Leveraging the public dataset and data we collected from Alzheimer's patients, AD-CLIP synthesizes human action videos featuring Alzheimer's disease. To address privacy concerns, labeling cost, and imbalanced data distribution, AD-CLIP generates a comprehensive labeled human action skeleton dataset from depth cameras with balanced data distribution. Our preliminary experiments confirm the effectiveness of the synthesized dataset by improving the accuracy of human activity recognition up to 76.56%, which demonstrates AD-CLIP's potential to enhance smart health applications.

Keywords

Synthesis dataset, Alzheimer's disease (AD), Human Action dataset.

1 Introduction

Smart health technologies with AI and sensors have redefined early diagnosis and intervention strategies by monitoring patients' daily activities for diseases like Alzheimer's Disease (AD) [6]. However, the performance of AI models strongly relies on high-quality training datasets, which are costly and rare. In particular, as the prevalence of AD increases, with an estimated 35.6 million affected individuals in 2010 and a doubling rate every 20 years [3], the need for AD patient datasets is more critical than ever.

Our work is motivated by the urgent need to address three main challenges: **(1) the scarcity of labeled human action data** which requires large costs in data collection and process work, **(2) the imbalance of existing datasets** that causes problems like long-tail distribution, and **(3) the limitation of sensor coverage** such as field of view (FoV) occlusion and fixed-view constraints.

Some previous synthesized dataset approaches using Vision-language models like BERT [1] and MotionDiffuse [9], have shown remarkable capabilities in understanding and generating image content [4]. However, most existing models have not been tailored to simulate the complex behaviors and actions characteristic of AD patients in daily living environments due to the lack of domain-specific knowledge and the gaps between medical images and human activity videos [7].

To tackle these challenges, we propose *AD-CLIP*, a framework that employs a CLIP-based model to generate a synthetic human action dataset of AD patients. AD-CLIP not only solves the scarcity of AD patients' action data by transferring the knowledge from the public human action recognition dataset but also generates balanced physiological and behavioral data from a large dataset we collected from **99** AD patients' homes. Moreover, the synthesized 3D skeleton human action videos give a comprehensive observation of human action. AD-CLIP demonstrates a privacy-preserving and low-cost way to provide AD patients' action datasets.

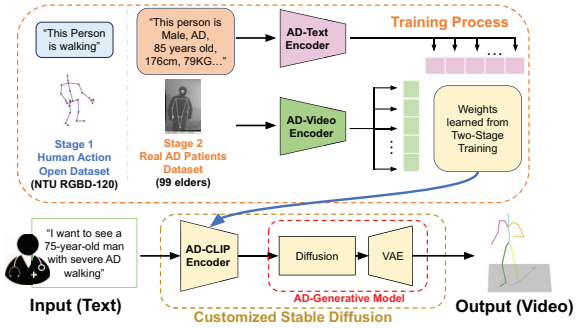

Figure 1: Design of AD-CLIP. The training process involves AD-Text Encoder and AD-Video Encoder. The deployment phase consists of a customized stable diffusion with AD-CLIP Encoder and AD-Generative Model.

2 System Design and Implementation

The AD-CLIP model, shown in Fig. 1, consists of three components: AD-Text Encoder, AD-Video Encoder and AD-Generative Model.

AD-Text Encoder: Leveraging the pre-trained OpenCLIP [2] encoder, the AD-Text Encoder receives descriptive textual prompts and encodes them into a feature space that captures the semantics necessary for generating corresponding action in the video domain.

AD-Video Encoder: The AD-Video Encoder is trained through a two-stage training process. Initially, training is conducted using the NTU RGBD-120 dataset [5], which contains human skeleton videos representing a variety of human actions. This allows the encoder to efficiently learn a wide spectrum of human actions.

In the second stage, we fine-tune the model using a dataset we collected with depth cameras in **99** elderly individuals' homes for one month, resulting in approximately **72,000** hours of videos and a total data volume of 100 TB. In addition, we have obtained access to their information under strict ethical standards and patient consent through a collaboration with a medical institution. The dataset showcases diverse physiological characteristics such as age, cognitive health status (including AD, Mild Cognitive Impairment (MCI), or normal cognition), and Montreal Cognitive Assessment (MoCA) scores. The raw video data is processed into human skeleton action videos.

AD-Generative Model: After completing the training of the encoders, we employ a customized version of the Stable Diffusion

979-8-3503-6201-5/24/$31.00 ©2024 IEEE
DOI 10.1109/IPSN61024.2024.00029

Table 1: HAR accuracies using vanilla and enhanced training data. The two HARs use the same model and testing data.

Method	Sitting	Standing	Walking	Turning	Lying
Vanilla	100%	0%	3.13%	0%	0%
Enhanced	76.56%	73.43%	78.12%	79.68%	75%

model [8] to generate videos from textual prompts. By substituting the original CLIP weights with our AD-CLIP trained weights, the AD-Generative Model is capable of synthesizing videos that reflect both the learned action features from the human action dataset and the physiological characteristics from our AD-behavioral dataset.

3 Preliminary Results and Demonstration

The synthetic dataset generated by AD-CLIP has a variety of uses, such as data augmentation. To evaluate the practicality of the AD-CLIP synthesized dataset, we trained a lightweight Human Action Recognition (HAR) model with both real-world AD data and synthetic data generated by AD-CLIP, and compared the results.

HAR Model: We used a simple HAR model with a composition of 5 CNN + 2 RNN layers for its light computational demand suitable for embedded systems in smart health applications.

Vanilla Training Data: Real-world labeled dataset of **21,910** videos from 79 elderly individuals, biased to the action of sitting, as it's the most routine activity for elders in the living room. The numbers of labeled activities in our dataset are: sitting 16,596, standing 3,121, walking 1,735, turning 272, and lying 186.

Enhanced Training Data: Synthetic dataset of **20,000** videos with balanced action distribution, evenly health condition representation (AD, MCI and NC), and 360-degree viewpoints around subjects.

Testing Data: 3,200 action videos from another 20 elders. To test the real-world performances of the two approaches more precisely, we select this cross-subject dataset with a balanced action distribution to mitigate the problems of high variance and high bias.

Result: Training with synthetic data (Enhanced HAR) led to a mean accuracy of 70-80% across actions, significantly higher than the 20.63% of the Vanilla HAR. Detailed accuracies for each action are shown in Table 1. The HAR model exhibited a significant performance improvement when trained on the AD-CLIP synthetic dataset. The synthetic dataset was meticulously crafted to address the limitations present in the real-world dataset:

Action Class Balance: The real-world dataset has a large proportion of sitting actions (over 75%), which led to a poor understanding of other actions. In contrast, the synthetic dataset was constructed with a uniform distribution of actions, ensuring that the model learned from an equal representation of each action class.

Physiological Variation: Patients with different ages and disease conditions exhibit distinct movement patterns and speeds. The AD-CLIP synthetic data generation process evenly distributed these physiological characteristics across all actions, leading to a dataset that covers the full spectrum of patient movement patterns.

Viewpoint Diversity: Camera placement in real-world settings resulted in varied perspectives, which can limit model exposure to certain viewpoints. The synthetic dataset, based on 3D skeletons, was

Figure 2: Demonstration of AD-CLIP. AD-CLIP generates a short video based on the input prompts.

evenly sampled from multiple viewpoints, providing a comprehensive set of angles, which led to more efficient observation of activities regardless of how patients are positioned in the clinical environment.

These insights demonstrate that the synthetic data generated by AD-CLIP provides a robust improvement for training more effective human action recognition models, particularly for embedded models that may struggle with the limitations of real-world datasets.

At the demo booth, we will show that AD-CLIP can interactively generate human action video according to the prompt. Users can explore its capabilities by requesting prompts for specific AD patient actions, and AD-CLIP will generate videos accordingly. A demo video is available at https://youtu.be/umRLpr-kxs0.

4 Conclusion

We present AD-CLIP as a privacy-preserving and low-cost tool for synthesizing human action datasets for AD patients. AD-CLIP demonstrates a significant improvement in the accuracy of human action recognition when enhancing embedded model using a synthesis dataset, which shows a promising potential in smart health.

5 Acknowledgment

This paper is supported by Hong Kong RGC C4072-21G.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:cs.CL/1810.04805

[2] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. OpenCLIP. https://doi.org/10.5281/zenodo.5143773

[3] Lampros C Kourtis, Oliver B Regele, Justin M Wright, and Graham B Jones. 2019. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *NPJ digital medicine* 2, 1 (2019), 1–9.

[4] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. arXiv:cs.CL/2310.07849

[5] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. 2020. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42. IEEE, 2684–2701.

[6] Xiaomin Ouyang, Xian Shuai, Yang Li, Li Pan, Xifan Zhang, Heming Fu, Xinyan Wang, Shihua Cao, Jiang Xin, Hazel Mok, Zhenyu Yan, Doris Sau Fung Yu, Timothy Kwok, and Guoliang Xing. 2024. ADMarker: A Multi-Modal Federated Learning System for Monitoring Digital Biomarkers of Alzheimer's Disease. arXiv:cs.LG/2310.15301

[7] Xiaomin Ouyang, Zhiyuan Xie, Heming Fu, Sitong Cheng, Li Pan, Neiwen Ling, Guoliang Xing, Jiayu Zhou, and Jianwei Huang. 2023. Harmony: Heterogeneous Multi-Modal Federated Learning through Disentangled Model Training. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services (MobiSys '23)*. Association for Computing Machinery, New York, NY, USA, 530–543. https://doi.org/10.1145/3581791.3596844

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:cs.CV/2112.10752

[9] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *arXiv preprint arXiv:2208.15001* (2022).