Privacy Preserving Representation Learning for Multimodal Behavior Understanding

AABHA RANADE, RAVI VIVEK AGRAWAL, SHIVANGI KOCHREKAR, and VAIDEHI VATSARAJ*, University of Southern California

1 PROBLEM DEFINITION

Behavior understanding, particularly through the analysis of biometric features such as voice patterns and facial expressions, holds significant promise for applications like emotion recognition and speech recognition. However, the utilization of such personal information raises substantial privacy concerns, particularly regarding the potential for unauthorized identification or surveillance. In this research paper, we propose an approach focused on learning representations of behavior that are informative for downstream tasks while safeguarding individuals' privacy. Our goal is to develop techniques that extract meaningful insights from behavior data without compromising individuals' anonymity. Through this project, we aim to explore various methods to preserve representation learning while ensuring privacy. With increasing data collection and privacy concerns, we identified this as an area that requires active attention and research.

2 LITERATURE SURVEY

To understand more about privacy issues in multimodal tasks, we reviewed research papers focusing on issues, causes and current methodologies available to mask the identity.

To understand the work already done in the domain of voice anonymization techniques, we reviewed various techniques addressing privacy preserving techniques in speech understanding and emotion recognition. The first paper [13] by Tran and Soleymani introduces the Privacy Transformer model, trained on a Voice cloning dataset, which utilizes the HuBERT pre-trained encoder to transform raw waveform data into high-level representations. Results demonstrate comparable accuracy in speaker identification tasks and superior performance in paralinguistic tasks compared to baseline models. Future work includes enhancing the model's performance with Distill-HuBERT. The second paper [3] by Feng and Narayana focuses on privacy and utility preservation through data transformation techniques such as Gradient Reversal Layer, Risk Model, and Replacement AutoEncoder. While they observed varied accuracy in predicting emotions, the use of RAE-transformed data diminishes accuracy in predicting gender. Moreover, [8] Luong et al. proposes the RDAL-M method, which employs source separation and adversarial learning to obscure speech presence in audio recordings while maintaining utility task performance. Jaiswal et. al. [5] found out was that data leakage is more prominent in audio input stream than in lexical stream but this effect compounds in multimodal. The paper experimented with adversarial strengths and found out that increasing the strength does not really affect the emotion recognition performance metric or the privacy metric. The main network consisted of three components: Embedding sub-network, emotion classifier and the gender classifier. The main network uses a GRU to unlearn gender while the attacker network having access to the held-out database tries to maximize emotion classification performance while minimizing gender classification performance. In the paper [1], the authors introduce a novel approach leveraging vector quantization to mitigate privacy concerns in voice data. By employing vector quantization, they effectively limit the representation space, thereby facilitating data anonymization. Notably, the authors highlight a crucial trade-off between preserving speech recognition accuracy and masking speaker identity, which can be modulated by adjusting the quantization size. This adaptive approach enables fine-tuning the level of anonymity while maintaining acceptable levels of speech recognition. Moreover, [10] proposes a novel method for speech anonymization. The authors extract an x-vector, which is an utterance-level embedding and then transform it by an autoencoder where speaker, gender, and accent information are suppressed through adversarial training. They generate an anonymized speech through a neural speech synthesizer. These papers collectively contribute innovative approaches to privacy preservation in speech understanding and gave detailed idea of how to approach this problem.

We explored numerous methods for facial features masking. Leong et.al. [7], presented LBP-TP for privacy-preserving facial recognition. Unlike traditional methods, which rely solely on spatial data, LBP-TP extracts information from the XT or YT planes of a video sequence. This approach aims to maintain privacy while ensuring high recognition accuracy. The study demonstrated promising results, achieving high accuracy rates across various databases. By striking a balance between privacy concerns and the utility of facial recognition technology, the authors offer a significant contribution to the field, reducing the amount of data required for recognition tasks. Chen et.al.[2], unveil PPRL-VGAN, a pioneering method devised to uphold visual privacy in facial recognition. Through a unique combination of Variational Auto-Encoders and Generative Adversarial Networks, this approach synthesizes faces while preserving expressions. Demonstrations on public datasets validate the efficacy of the technique, showcasing its potential for privacy-preserving applications and driving future advancements in secure facial recognition technology.

Moreover, Sun et. al. [11] use an auto encoder for the landmark converter and Target-specific landmark-to-face generator with a PixelShuffle layer. The observed that the Face Synthesis method is flexible to larger changes to the landmarks but has a limit. Too many changes to the landmarks lead to bad results.

Tang H. et. al. [12] uses a single image with a two stage strategy for their LandmarkGAN. The task is split into 2 subtasks: Category guided landmarks generation and Landmark guided expression to expression translation. In the first task they have employed a Convolutional GAN and use U-Net for their second task. They have

Authors' address: Aabha Ranade, aabharan@usc.edu; Ravi Vivek Agrawal, ravivive@usc.edu; Shivangi Kochrekar, kockreka@usc.edu; Vaidehi Vatsaraj, vatsaraj@usc.edu, University of Southern California.

compared their method to a lot of state of the art GANs and give an insight as to why their method works better.

Microsoft researchers [14] have performed 3D Face Reconstruction with Dense Landmarks, using over 700 dense landmarks to perform 3d face reconstruction by fitting a morphable face model to them. First, they predict probabilistic dense 2D landmarks using a traditional convolutional neural network (CNN). Then, they fit a 3D face model to the 2D landmarks, enabling intelligent aggregation of face shape information across multiple images. The results showcase robust outcomes in both multi-view and monocular facial performance capture scenarios.

Nousi [9] introduces a novel approach in face de-identification by using Deep Autoencoders and then fine tuning the encoder to perform face de-identification. They implemented various methods to finetune the encoder in both a supervised and unsupervised fashion to preserve facial attributes, while generating new faces which are both visually and quantitatively different from the original ones.

Feng et. al. [4] introduce the Position map Regression Network (PRN), an end-to-end method for simultaneous dense alignment and 3D face shape reconstruction. To achieve this, they designed a 2D representation called UV position map which records the 3D shape of a complete face in UV space, then trained a simple encoder decoder network to regress it from a single 2D image.

Looking at various literature in the domain we decided to go explore mask rendering and auto encoders for the video privacy preservation. For audio we explored variety of ways for anonymizing . We shall speak about it in detail in the other sections of the paper.

3 DATA DESCRIPTION

For this project we have selected the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA D) which consists of 7,442 original video clips from 91 actors (48 male and 43 female) having ages between 20 - 74 years, belonging to variety of races and ethnicities (African American, Asian, Caucasian, Hispanic and Unknown). The actors spoke from a selection of 12 sentences having average length of 2.63 ± 0.53 seconds. The videos express six different types of emotions (Anger, Disgust, Fear, Happy, Neutral and Sad) and four different emotion levels (Low, Medium, High and Unspecified). We have also used the Flickr-Faces-HQ (FFHQ) dataset[6] for training our auto encoder model. This dataset consists of 70,000 high-quality PNG images, having 1024×1024 resolution with considerable variation in terms of age, ethnicity, image background, presence of accessories such as eyeglasses, sunglasses, hats, etc.

4 METHODOLOGY

Our proposed method operates on videos extracted from the dataset by independently processing the audio and video components. After modifying both elements and obtaining emotion recognition results for each, we plan to perform late fusion to integrate the outcomes for multimodal emotion recognition. Additionally, we intend to evaluate privacy preservation measures at this stage.

4.1 Face Anonymization

We had initially proposed to achieve face anonymization via facial landmark detection and masking. However, after conducting experiments and reviewing more recent work done in this domain we concluded that this might not be the best approach to solve this problem. We shifted our focus to other solutions. Using auto encoders to recreate certain portions of the face after augmenting it via "masking it" by deleting certain pixel showed some promise. We can breakdown are final process into the following steps:

- The first step is to extract the frames from the video file. conducting.
- Then we use a custom generator function for image preposessing in the form of resizing and adding noise and load the images in batches.
- We trained an auto encoder model on 54,000 images with the following hyperparameter tuning (learning rate = [0.01, 0.001], optimizers = [Adam, SGD], Loss = [mse, rmse], Batch size= [32, 64], Epochs = [50, 100, 150] Latent space size = 300).
- We identified the best hyperparameters as learning rate = 0.001, optimizer = Adam, Loss = mse, Batch size = 64, Epochs = 100, Latent space size = 300.
- Finally, we reconstruct the video from the augemented frames generated by the autoencoder.

For audio anonymization and testing in our study, we employed the following method:

4.2 Audio Modification:

We utilized voice anonymization techniques aimed at altering audio properties to obscure speaker identity. Specifically, we employed the following technique:

Vector Quantization: We applied vector quantization techniques to cluster Mel-frequency cepstral coefficients (MFCCs) using K Means Clustering. This process helped in grouping similar MFCCs together, thereby anonymizing the audio data.

Steps involved in Vector quantization:

1) We utilized the k-means algorithm available in the scikit-learn library. This algorithm clusters the input data into a specified number of clusters. We experimented with different number of clusters to see its effectiveness and finally decided to consider 16 clusters which gives 16 centroids.

2) For quantizing the data, we used the trained k-means model. This involves predicting the cluster assignments for each data point within the input dataset. Here, the we use MFCCs as the input daata points. Subsequently, each data point is assigned to its nearest centroid.

3) These centroids act as the new features for the reconstructed audio and we create a new audio signal by inverting these MFCCs back to an audio signal.

We then trained a classifier to classify the emotions on the processed audio signal.

5 EXPERIMENTS

We tested out anonymization model on the downstream task of emotion recognition. For testing we used a subset of CREMA D dataset having 180 data points spread across 6 emotions uniformly.

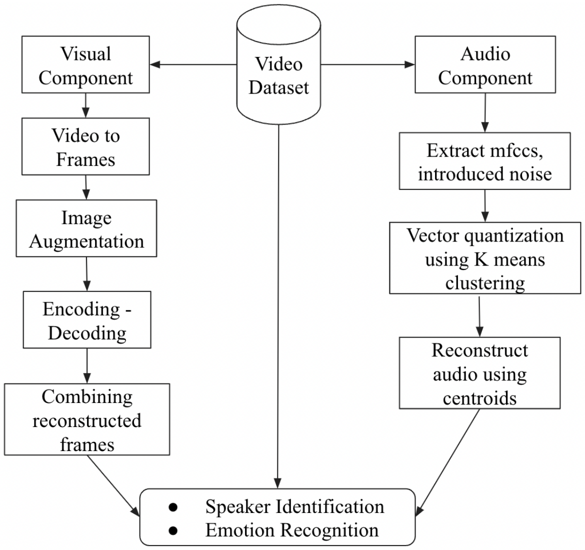
Visual Component
Video Dataset
Audio Component
Video to Frames
Extract mfccs, introduced noise
Image Augmentation
Vector quantization using K means clustering
Encoding - Decoding
Reconstruct audio using centroids
Combining reconstructed frames
Speaker Identification
Emotion Recognition

Fig. 1. General methodology

We implemented late fusion to test how our augmented privacy preserved data performs on emotion recognition.

5.1 Emotion Recognition Using Pretrained Model:

The main aim of the problem statement is to ensure that the performance of tasks such as emotion recognition is not impacted significantly by application of anonymization technniques.

- We used Facial Emotion Recognition(FER) model for the video component of the data. It exhibited 39% accuracy on 180 points of the CREMA D dataset and observed a 8% decline after applying facial anonymization technique.
- We used a pretrained 'emotion-recognition-wav2vec2' model trained on IEMOCAP dataset. This model displayed an accuracy of 33% on 180 points of the CREMA-D dataset. After applying the vector quantization techniques, the accuracy dropped to 17%.
- Since a pretrained model did not give a good accuracy, we trained our own Random Forest model on the original CREAMA-D dataset. The model displayed an accuracy of 79% on the test data. Training our own model directly on the CREMA-D dataset yielded significantly better results compared to using a pretrained model. This indicates that the pretrained model, although trained on a similar dataset (IEMOCAP), might not have generalized well to the CREMA-D dataset.

We combined the results of both the modalities using late fusion with a 60-40 weighted voting system and obtained an accuracy of 48% on the task of emotion recognition.

5.2 Identifying the speaker from Anonymized Face:

For testing privacy preservation, we first trained a VGG16 model on original frames taken from CREMA-D dataset which consists of 91 individuals. The model was trained on every fifth frame from each video for 20 epochs with 100 percent training accuracy.

After training a VGG16 model to identify individuals in original video frames, we achieved perfect accuracy on both the training and validation sets. However, upon testing the model on reconstructed video files generated by the autoencoder, we noticed a 0.06 accuracy.

5.3 Identifying the Speaker from Anonymized Audio:

To assess the effectiveness of the anonymization techniques in preserving speaker anonymity, we employed following techniques:

1) Initially, we trained a CNN model on the original CREMA-D dataset, focusing on speaker recognition through the analysis of MFCCs extracted from audio signals. While this approach yielded promising results, we acknowledged the potential biases inherent in training our own model.

2) After discussion and analyis of our results, we realized that training our own model can have biases. Hence we used a pretrained model 'Speaker Verification with ECAPA-TDNN embeddings on Voxceleb' named 'spkrec-ecapa-voxceleb'. This model gives a similarity score between the audio files based on their speaker. It gives a probability of the two files having the same speaker. To test our model's privacy, we tested the similarity scores between the raw audio files of the same speaker, having different emotions/ same emotions and the similarity scores of the reconstructed files with that of the original file. Table 1 shows the summary of the results obtained from the speaker identity recognition on the reconstructed audio.

R1 results indicate the scores between two audio files of the speaker having different emotions and R2 results indicate the scores between the original file and reconstructed file.

6 RESULTS

6.1 Results on reconstructed audio

In our evaluation of voice anonymization techniques, we found that the Vector Quantization (VQ) method yielded superior results in

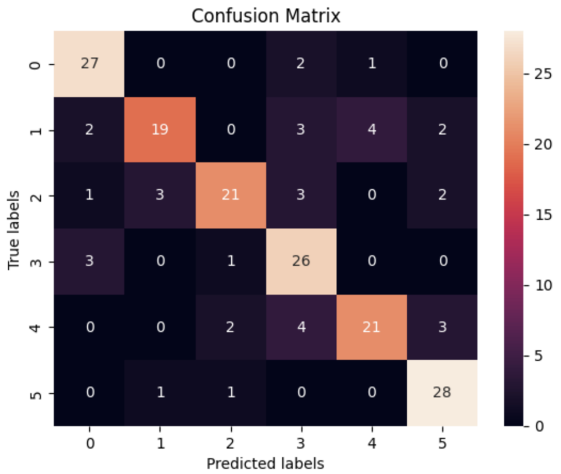

Fig. 2. Emotion Recognition on reconstructed audio: Confusion Matrix

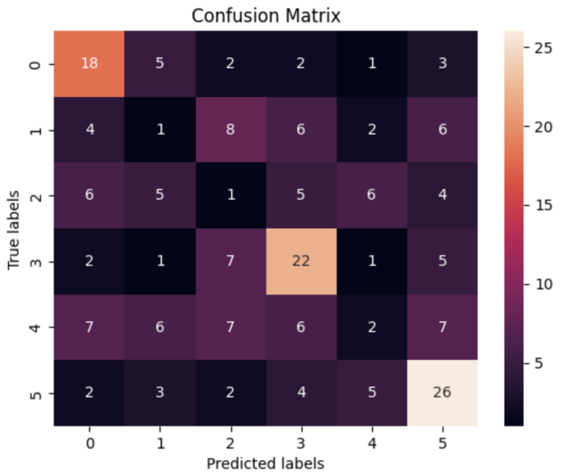

Fig. 3. Emotion Recognition on reconstructed face: Confusion Matrix

Table 1. Results for Speaker Identity Recognition

Results	R1	R2
Mean	0.483	0.052
Median	0.496	0.056
Minimum Score	0.042	0
Maximum Score	1	0.266

both voice emotion recognition and speaker anonymization tasks. We obtained the confusion matrix as displayed in figure 2.

6.2 Results of extracted facial expressions

Confusion matrix for the best model trained using auto-encoder was tested on the FER pretrained model and we obtained the confusion matrix as displayed in figure 3.

7 CONCLUSION AND LESSONS LEARNED

We ran into a lot of challenges trying to achieve privacy while simultaneously trying to preserve the useful features required for further downstream tasks. We initially tried to implement a face mask generator for the video anonymization task but eventually shifted to a more defacing and reconstructing approach using autoencoders. We observed that this task is quite complex and difficult because of the fact that the features required for tasks such as emotion recognition etc. overlap with the key features that are used to identify a person. While trying to maintain balance and achieve an acceptable trade off between privacy and preservation seemed to be quite a challenge. The use of auto encoders though might not be the perfect solution right now, seems to show promising results. Maybe with the increasing development in the auto encoders and/or new techniques this task might seem trivial in the recent future. For the audio task, we observed that altering the properties of speech are reversible tasks and are not robust enough. We conducted a series of experiments to enhance the anonymization of audio data through the introduction of noise. Employing various levels of sensitivity and epsilons, our analysis revealed that while there was not a significant decrease in speaker identification accuracy, there was a small decline in emotion recognition performance. Consequently, we decided not to integrate noise into our data anonymization process. With the increasing collection of data in the form or audio/video etc there also a rise in concern regarding the privacy and this seems area seems worthy of further research and focus.

8 SUMMARY AND CONTRIBUTION

Video privacy while maintaining emotional features is a more convoluted task than we initially estimated. Audio privacy seems to be a relatively less intricate task as evident from the results. 3d mesh generation is a highly computationally heavy task, which does not justify the trade off for it be used in the pipeline. Privacy is a major concern with the increasing use of AI models in the present scenario and this domain requires further research. We used auto encoders for video privacy preservation and vector quantization methods for audio anonymization. Each team member has contributed significantly to these efforts:

(1) Aabha Ranade - Conducted literature review for face reconstruction techniques like autoencoders and 3d mesh construction, Code for parsing videos to frames and corrupting the frames, autoencoder architecture, video emotion recognition, trained VGG16 model for person identification from images.
(2) Ravi Vivek Agrawal - Conducted literature review for multimodal privacy preservation techniques, emotion recognition for videos model, combining the results for Emotion Recognition Task and late fusion of the audio and video results
(3) Shivangi Kochrekar - Conducted literature review for video anonymization techniques, extracted the facial expressions, autoencoder architecture, making the video privacy,emotion and re-identification pipeline
(4) Vaidehi Vatsaraj - Conducted literature review for audio and video speaker anonymization techniques and implemented the vector quantization, worked on audio anonymization technniques, training a random forest model for audio emotion recognition, and tested the speaker identitification model on reconstructed audio

REFERENCES

[1] Pierre Champion, Denis Jouvet, and Anthony Larcher. 2022. Privacy-preserving speech representation learning using vector quantization. *arXiv preprint arXiv:2203.09518* (2022).

[2] Jiawei Chen, Janusz Konrad, and Prakash Ishwar. 2018. Vgan-based image representation learning for privacy-preserving facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops.* 1570–1579.

[3] Tiantian Feng and Shrikanth Narayanan. 2021. Privacy and utility preserving data transformation for speech emotion recognition. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII).* IEEE, 1–7.

[4] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. 2018. Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network. *CoRR* abs/1803.07835 (2018). arXiv:1803.07835 http://arxiv.org/abs/1803.07835

[5] Mimansa Jaiswal and Emily Mower Provost. 2020. Privacy enhanced multimodal neural representations for emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7985–7993.

[6] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 4401–4410.

[7] Shu Min Leong, Raphaël C.W. Phan, Vishnu Monn Baskaran, and Chee Pun Ooi. 2020. Privacy-preserving facial recognition based on temporal features. *Applied Soft Computing* 96 (Nov. 2020). https://doi.org/10.1016/j.asoc.2020.106662 Publisher Copyright: © 2020 Copyright: Copyright 2020 Elsevier B.V., All rights reserved..

[8] Diep Luong, Minh Tran, Shayan Gharib, Konstantinos Drossos, and Tuomas Virtanen. 2023. Representation learning for audio privacy preservation using source separation and robust adversarial learning. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).* IEEE, 1–5.

[9] Paraskevi Nousi, Sotirios Papadopoulos, Anastasios Tefas, and Ioannis Pitas. 2020. Deep autoencoders for attribute preserving face de-identification. *Signal Processing: Image Communication* 81 (2020), 115699.

[10] Juan M Perero-Codosero, Fernando M Espinoza-Cuadros, and Luis A Hernández-Gómez. 2022. X-vector anonymization using autoencoders and adversarial training for preserving speech privacy. *Computer Speech & Language* 74 (2022), 101351.

[11] Pu Sun, Yuezun Li, Honggang Qi, and Siwei Lyu. 2022. Landmarkgan: Synthesizing faces from landmarks. *Pattern Recognition Letters* 161 (2022), 90–98.

[12] Hao Tang and Nicu Sebe. 2022. Facial expression translation using landmark guided gans. *IEEE Transactions on Affective Computing* 13, 4 (2022), 1986–1997.

[13] Minh Tran and Mohammad Soleymani. 2023. Privacy-preserving Representation Learning for Speech Understanding. *arXiv preprint arXiv:2310.17194* (2023).

[14] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljevic, Daniel Wilde, Stephan Garbin, Chirag Raman, Jamie Shotton, Toby Sharp, Ivan Stojiljkovic, Tom Cashman, and Julien Valentin. 2022. 3D Face Reconstruction with Dense Landmarks. In *2022 European Conference on Computer Vision.* https://www.microsoft.com/en-us/research/publication/3d-face-reconstruction-with-dense-landmarks/