A COMPUTATIONAL FRAMEWORK FOR EXPLORING AND MITIGATING PRIVACY RISKS IN IMAGE-BASED EMOTION RECOGNITION

A Thesis

by

VANSH NARULA

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Theodora Chaspari
Committee Members,	Zhangyang (Atlas) Wang
	Tie Liu
Head of Department,	Scott Schaefer

May 2020

Major Subject: Computer Engineering

ABSTRACT

Ambulatory devices and Image-based IoT devices have permeated our every-day life. Such technologies allow the continuous monitoring of individuals' behavioral signals and expressions in every-day life, affording us new insights into their emotional states and transitions, thus paving the way to novel well-being and healthcare applications. Yet, due to the strong privacy concerns, the use of such technologies is met with strong skepticism as they deal with highly sensitive behavioral data, which regularly involve speech signals and facial images and current image-based emotion recognition systems relying on deep learning techniques tend to preserve substantial information related to the identity of the user which can be extracted or leaked to be used against the user itself. In this thesis, we examine the interplay between emotion-specific and user identity-specific information in image-based emotion recognition systems. We further propose a user anonymization approach that preserves emotion-specific information but eliminates user-dependent information from the convolutional kernel of convolutional neural networks (CNN), therefore reducing user re-identification risks. We formulate an iterative adversarial learning problem implemented with a multitask CNN, that minimizes emotion classification and maximizes user identification loss. The proposed system is evaluated on two datasets achieving moderate to high emotion recognition accuracy and poor user identity recognition accuracy, outperforming existing baseline approaches. Implications from this study can inform the design of privacy-aware behavioral recognition systems that preserve facets of human behavior, while concealing the identity of the user, and can be used in various IoT-empowered applications related to health, well-being, and education.

DEDICATION

To my mother, my father, my sister, and my brother.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my advisor Dr. Theodora Chaspari for her constant support, patience, enthusiasm and optimism throughout this research journey. Her constant motivation enabled me to explore the field of research at my own pace allowing me to come up with solutions and learning from my own mistakes. I am forever grateful for giving me this opportunity to fulfill my ambition of being a researcher. I would like to thank Dr. Zhangyang (Atlas) Wang for the inspiration as his work in this field has been one of the major driving force behind this research work. I would also like to thank Dr. Tie Liu for serving on my committee and constant support.

I would like to thank my fellow researchers at HUBBS lab: Prakhar Mohan, Ehsanul Haque Nirjhar, Projna Promita and Kexin Feng for their constant support and encouragement. I also want to thank my friends, Nikhil Bhat, Manish Patel, Prakhar Mohan, Abhimanyu Chadha and Puneet Bansal for stimulating discussions, motivation and for all the fun times in the last 2 years.

Finally, I am thankful to my family and my friend Dr. Simran who have always supported me in every phase of my student life. It is due to their sacrifices, teachings and undying love that I have succeeded in achieving this goal.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

The thesis committee for this work include Professor Theodora Chaspari (Chair) and Professor Zhangyang Wang (Member) of the Department of Computer Science at Texas AM University and Professor Tie Liu (Member) of the Department of Electrical and Computer Engineering.

The ideation and implementation of this research was conducted with the help and guidance of Professor Theodora Chaspari. Professor Zhangyang Wang's prior work served as the inspiration behind this work. All other work conducted as part of this thesis was completed by the author independently.

Funding Sources

This research was supported by the Texas A&M University Program to Enhance Scholarly and Creative Activities (PESCA).

NOMENCLATURE

CNN	Convolutional Neural Network
HIPAA	Health Insurance Portability and Accountability Act of 1996
JAFFE	Japanese Female Face Expressions database
YALE	Yale face dataset
AI	Artificial Intelligence
IoT	Internet of Things
HoG	Histogram of Oriented Gradients
EHR	Electronic Health Records
RBF	Radial Basis Function
SIFT	Scale Invariant Feature Transformation
SVM	Support Vector Machine

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Privacy risks in IoT devices

Image and video-capturing devices have become increasingly ubiquitous and pervasive. From the millions of surveillance cameras installed all over the world to the newly introduced smart home devices, such ambulatory recording technologies allow the continuous monitoring of individuals over long periods of time rendering ecologically valid data of human emotional, mental, and psychological states [1, 2, 3]. Such “high-volume" and “high-velocity" data can be integrated with Artificial Intelligence (AI) methodologies resulting in smart sensing technologies to promote various healthcare applications beyond well-established applications related to security monitoring and community safety. For example, the monitoring of facial expressions and body gestures in a continuous manner can capture momentary and longitudinal patterns of human emotion, which can be reflective of users’ stress, depression, or even suicidal risk, therefore rendering such information a valuable biomarker for predicting and potentially intervening upon individuals’ mental and emotional health [2].

Recent advances in ambulatory sensing and Internet-of-Things (IoT) technologies allow the continuous monitoring of behavioral signals in everyday life, applications [4, 5, 6, 7]. Speech signals captured by voice-enabled smart-home devices can track prosodic, spectral, and temporal characteristics of human speech in high time-granularity, yielding valuable biomarkers of serious mental conditions [2, 8].

Despite the premises, the barrier of confidentiality and anonymity inherent in these smart-monitoring applications is an issue with various social and cultural implications preventing their wide adoption. Users are often skeptical of such technologies, since they are afraid that facial information relevant to their identity will be permanently stored in third-party servers or will be abused by hacker attacks [9]. User authentication and authorization is a significant challenge in IoT devices with well-established user authentication protocols to identify potential privacy breaches [10, 11]. Data anonymization has also been one of the basic mechanisms of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule [12], attesting to the significance and timeliness of this topic.

These do not come as a surprise: behavior recognition systems rely on rich speech spectrotermporal patterns and facial features. The Mel-Frequency Spectral Coefficients extracted from speech capture subtle spectrotermporal characteristics of the human voice which can be directly associated with both user state and user identity [13, 14]. The Histogram of Oriented Gradients (HoG) incorporates rich textural information of the facial image [15], eigen-faces [16] rely on the most significant eigen-vectors that preserve much of the total energy of an image, while other approaches leverage the frequency characteristics of an image through Gabor filters and Wavelets [17] which respond to change in illumination and texture. In addition, state-of-the-art representation learning models, such as convolutional neural networks (CNN) [18], and recent advancements, such as the Resnet, MobileNet and Inception network [19], whose input is the 2D speech spectrogram or 3D image tensor, are often trained on massive datasets containing sensitive data [18], therefore tend to preserve significant amount of facial information related to the user identity, social content, and emotional expression [20, 21]. For example, CNNs are known to capture general and highly reusable information in their convolutional basis, which may be useful for another target task. This results in both the desired utility-based information (e.g., emotional, mental, and psychological state) as well as the undesired privacy-sensitive information (e.g., user identity) being preserved in the convolutional base and the subsequent fully-connected layers. For this reason, even when CNNs are trained on a specific task of interest, the information required for a new similar task might be embedded in the pre-trained convolutional base, therefore the CNN can easily be fine-tuned on the new task [20]. This renders data privacy a major barrier for collecting and sharing human behavioral signals, stalling the research progress and preventing the wide adoption of smart health and well-being systems. This privacy compromising landscape renders essential the design of novel machine learning systems that conceal one's identity, while at the same time preserve useful information for emotion recognition.

1.2 User anonymization and differential privacy

As human behavioral signals become more and more ubiquitous and prevalent, there is a need to appropriately define privacy and design computationally rigorous algorithms that satisfy the considered privacy constraints. The current paper will focus on user anonymization, which refers to the problem of removing identifiable information that may lead to user identification, with an ultimate goal for the user of the device to remain as much anonymous as possible. Data anonymization is typically viewed as protecting the privacy of a user, concealing user-dependent information that might be preserved in the internal structure of the IoT mechanism, and maintaining anonymized traffic on data packets transferred between the devices [22, 23]. There has been an extensive prior work on cryptographic techniques that prevent an unauthorized attacker from gaining access into a set of data [24]. Various computational solutions have been further proposed that build machine learning models without sharing patient-level data and support privacy-preserving dissemination of data, including homomorphic encryption mechanisms [25, 26], secure multiparty computation [27, 28], and federated learning [29, 30]. Another line of work follows the field of differential privacy, which refers to describing patterns related to a utility-based information in a dataset, while withholding patterns relevant to privacy sensitive aspects that put users at-risk of re-identification, such as characteristics revealing the unique identifiability of the user.

Differential privacy methods have proposed to add controlled noise to the data to render user identification hard and have integrated a privacy-policy criterion to the loss function of well-known classifiers (e.g., logistic regression, support vector machine) [31, 32, 33, 34]. Despite the encouraging progress, the majority of approaches on privacy-preserving machine learning are highly focused on con-signal-based data, such as Electronic Health Records (EHR) and genomic data. In contrast, signal-based data have been sparsely explored with most of the work focusing on general human activity recognition [35, 36], which involves visual frames of the full body from one or multiple individuals captured from a long distance. The problem of privacy-preserving human emotion recognition from signal-based data, such as images, presents an additional set of unique challenges, since it involves the learning of subtle highly personalized emotional expressions. Images depict high spatial dependency compared to discrete EHR or genomic data, rendering the privacy-preserving emotion recognition problem much more challenging.

This thesis will focus on user anonymization, which refers to the problem of removing identifiable information that may lead to user identification, with an ultimate goal for the user of the device to remain as much anonymous as possible.

1.3 Prior work

IoT devices are prone to privacy threats at every step of the data life cycle, from the collection of the data to its final disposal. This has fostered significant concerns among users, developers, and researchers, who acknowledge that information about individuals must be protected and should not be exposed without explicit consent under any circumstance [37]. Prior work has outlined ways to promote user privacy through mutual user authentication, encryption of data communication, and user anonymization [37, 38]. Mutual user authentication aims to grant access to a user at the registered IoT services considering inter-device authentication and session-key distribution systems [39, 40]. Encryption promotes security of data and sensitive information when transmitted through the network with numerous techniques, such as homomorphic encryption, which have been proposed over the years [41]. Finally, user anonymization is the process of removing information that may lead to the identification of the user [22]. With the advent of real-life multimodal behavioral data collected by IoT devices, user anonymization has become a prevalent challenge.

The most well-known user anonymization techniques include the k-anonymity model, and its extenstions of l-diversity and t-closedness, as well as differential privacy approaches. The k-anonymity model is one of the pioneer user anonymization methods [42]. K-anonymity aims to guarantee that a user's privacy information cannot be distinguished from at least $k - 1$ individuals. l-diversity and t-closedness principles have extended k-anonymization by reducing the representation granularity of the data through suppression and generalization, mostly employed in EHR and genomic sequence data. For example, a multidimensional suppression technique has been proposed that combines data attributes in EHR through feature selection [43]. Yoo et al. [44] proposed a generalization method based on conditional entropy for measuring the loss of information of sensitive features. Heatherly et al. [45] attempted to anonymize clinical profiles of patients in a hypothyroidism study by employing a k-anonymization algorithm at three medical centers. The study demonstrated that record generalization was lower when anonymization was performed on the entire EHR record, compared to anonymization focused on a specific cohort of patients. Martínez *et al.* [46] developed a generalized framework which leveraged semantic properties of non-numerical attributes in EHR records. Tamersoy *et al.* [47] used sequence aligning and clustering methods to support secure sharing of patient-specific longitudinal data by aggregating temporal and diagnostic information while preserving data utility. Kim *et al.* [48] designed privacy-preserving “data cubes" based on global and local generalization and bucketization. Loukides Gkoulalas-Divanis [49] proposed a method to anonymize diagnosis codes with generalization and suppression, taking into consideration that a patient’s identity could be linked with genome sequences using diagnosis codes. Hughes *et al.* [50] developed an online system using replacement and suppression to anonymize patient-level clinical trial data with an objective to maximize the utility for research. Finally, Poulis *et al.* [51] presented an alternative approach in which users had the ability to specify utility constraints on their data.

Differential privacy is a principled approach of user anonymization aiming to control the degree of user re-identification by embedding predefined noise in the data, thus creating a trade-off between data utility and user re-identification risk [52]. This framework has been well-explored in biomedical research, including EHR data [53, 31, 54], with the lack of an objective approach for determining the right amount of noise to achieve an acceptable balance between privacy protection and utility being a significant challenge. Ji *et al.* [32] proposed a distributed logistic regression model that synthesizes public and private data across different sites in a differentially private manner. Li *et al.* [33] presented a hybrid support vector machine based on a radial basis function (RBF) kernel to handle non-linearly separable cases. Simmons *et al.* [55, 34] tested a differential privacy framework on a rheumatoid arthritis dataset.

Data anonymization techniques have also been extensively used in the area of computer vision and IoT. Pre-defined image transformation approaches have attempted to increase the amount of uncertainty in an image by adding noise [56] or performing filtering operations [57], such as image blurring. Wang *et al.* [58] introduced a scalable privacy-aware IoT architecture that degraded facial characteristics. Other approaches have attempted to encode the change in successive images as the input to the system, rather than the image pixels themselves. For example, Steil *et al.* [59] detected changes in users' eye movements to approximate potentially privacy-sensitive image frames. The authors used the inner camera of eyewear devices, such as head-mounted displays or augmented reality glasses, to track eye movements. Information from the secondary camera was subsequently encoded in the privacy-preservation system and used as a signal to control for the amount of information to be preserved in the primary camera. Another set of methods have formulated data anonymization in signal-based data as an optimization problem, according to which data transformations are learned based on the antagonistic criteria of preserving utility-based information and suppressing user-related characteristics. The ideal transformation would increase a target utility metric and minimize a privacy-based metric [60, 61, 62, 63, 64]. Wu *et al.* [61], for example, proposed an adversarial learning framework for privacy-preserving human activity recognition with promising results. Prior work has also attempted to combine the two approaches. Bertran *et al.* [65] used adversarial learning to generate a domain-preserving transformation of the input image. Toward this, they used very deep convolutional neural networks, such as the Xception Net, to generate a transformed image that performs well on a utility function, but underperforms on the privacy-sensitive function. While this approach yielded promising results, the domain-preserving constraint becomes challenging when the information related to the utility and the sensitive tasks are highly inter-dependent. The results obtained in prior work are further highly dependent of the type of utility and sensitivity functions used, which makes the proposed system hard to generalize to unseen contexts and conditions. Moreover, it is not clear how the domain-preserving adversarial approach can be implemented with a shallow rather than a very deep network, which is necessary for the limited-memory IoT devices. Feutry *et al.* proposed the problem of privacy-preserving image transformation, in which an autoencoder architecture was trained in an adversarial manner to yield a new image that does not retain user-specific characteristics. The authors evaluated their approach on the task of handwritten digit recognition and emotion classification with promising results. Finally, user anonymization has been implemented in automatic speech recognition and speech-based emotion recognition [66, 67, 68].

1.4 Proposed thesis contributions

This thesis advances existing literature in the following ways: (1) In contrast to the majority of previous work on privacy preservation for tasks which are not highly dependent on user identity, such as general human activity recognition, this paper proposes a privacy-preservation system specifically for the task of emotion recognition. Image-based emotion recognition is highly dependent on subtle facial characteristics, for which it is much more difficult to learn appropriate degradation transformations of the facial images while preserving the emotion dependent information; (2) While previous work has focused on data obtained with surveillance cameras or distant cameras capturing the entire body from one or multiple users [61, 62, 63], this thesis relies on cameras placed in close proximity to a user's face, therefore preserving a high amount of identity-specific information; (3) Most of the work does not provide a clear method for evaluating the trade-off between the degradation of utility-based information and preservation of user identity [57, 56, 61, 64]. The proposed adversarial learning framework results in a convolutional transformation that attempts to degrade user-specific information for any of the subsequent fully-connected layers. The output of the convolution is fed into two classifiers, one for emotion recognition and the other for face identification and the corresponding accuracies are recorded. These accuracies can quantify the amount of identity- and emotion-specific information preserved in the convolutional layers of the CNN.

1.5 Research objectives and contributions of this research

1.5.1 Research aims

In the light of the above challenges, this thesis will focus on three main research questions:

1. Can we understand and quantify the interplay between user identification risks and emotion utility in images?

2. Can we develop a privacy-preserving machine learning model to limit user identification risks while retaining emotion inference capability?

1.5.2 Proposed approach

First, we will explore the extent to which user identity is preserved in image-based emotion recognition systems. We implement this through a CNN initially trained on the task of emotion classification. We then freeze its convolutional layers and fine-tune its fully-connected layers on the task of user identification in order to quantify the extent to which the convolutional basis of the emotion-specific CNN is able to retain user-specific information. We evaluate our results on both unblurred and blurred images in order to determine whether image blurring can be effective for user anonymization in emotion recognition. We also compare our results with classical feature extraction and image classification methods such as SIFT and SVM. Such comparison would allow us to demonstrate the importance of using CNN architectures over classical and simple ML classification models. In an attempt to design a privacy-aware emotion recognition model, we further propose a multitask CNN architecture composed of a convolutional basis, which is followed by two parts of fully-connected layers, one for emotion classification and another for face identity recognition. The convolutional layers are shared among the two fully-connected parts and the corresponding weights are learned in an adversarial way to preserve emotion-specific information and degrade information related to users' identity. We follow an alternate weight freezing approach, in which the convolutional and emotion-specific weights are re-trained conditioned on user-specific weights that are effective for the task of user identification. In this way, we aim to eliminate user-related information in the convolutional basis so that it cannot be re-trained and employed for user re-identification. The alternate weight freezing is compared against no weight freezing during the adversarial learning. The final evaluation of our system is performed by adding a new set of fully-connected layers on the already learned convolutional basis, and re-training the new weights for the user classification task. Our quantitative and qualitative results obtained in the Japanese Female Facial Expression (JAFFE) database [69] and the Yale Face Dataset (YALE [70]) will be discussed and examined to check if they demonstrates the feasibility of the proposed framework for promoting user privacy in image-based emotion recognition.

2. METHODOLOGY

In the following, we will describe our work on quantifying user anonymity and designing a privacy-aware machine learning system for emotion recognition. We first use classical feature extraction and image classification algorithms such as SIFT and SVM to generate results for both face and emotion recognition that will serve as baseline for our Convolutional neural network based feature extractor and classifier (Section 2.1). Then, we examine the interplay between emotion- and identity-specific information in the convolutional transformation learned by CNNs (Section 2.2). Following this, We describe our proposed adversarial learning approach, which learns a convolutional transformation in a CNN that preserves of emotional information and suppresses identity-specific information (Section 2.3). The convolutional transformation is learned by optimizing an adversarial loss function. Optimization is implemented through an alternate weight freezing approach: the user-specific weights are learned in order to achieve a good user classification model and are subsequently frozen so that the convolutional and emotion-specific weights are learned toward a successful privacy-preserving emotion classification system, which is being compared against many possible effective user identification schemes. We further outline quantitative ways to evaluate the effectiveness of our system and describe how to demonstrate convergence of the proposed algorithm by computing the utility and privacy-related loss over multiple training iterations, as well as through the visualization of the resulting image transformations (Section 2.4).

2.1 Quantifying user-related and emotion-related information retained by classical feature extractors

We will first examine the features captured by classical feature extraction methods such as the Scale Invariant Feature Transform (SIFT) [71]. These features are not task dependent and hence would be same for both emotion recognition and face recognition task. We use combination of Bag of Features (BOF) extracted using SIFT and Support Vector Machine (SVM) classifier which has been successfully implemented in various classification tasks such as hand gesture and vehicle

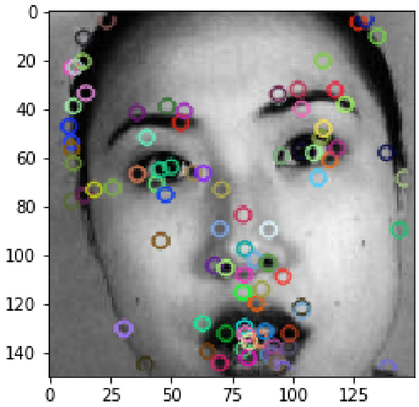


Figure 2.1: A visual representation of key points of an image as selected by SIFT descriptors

images.It is applied to both face and emotion classification task.

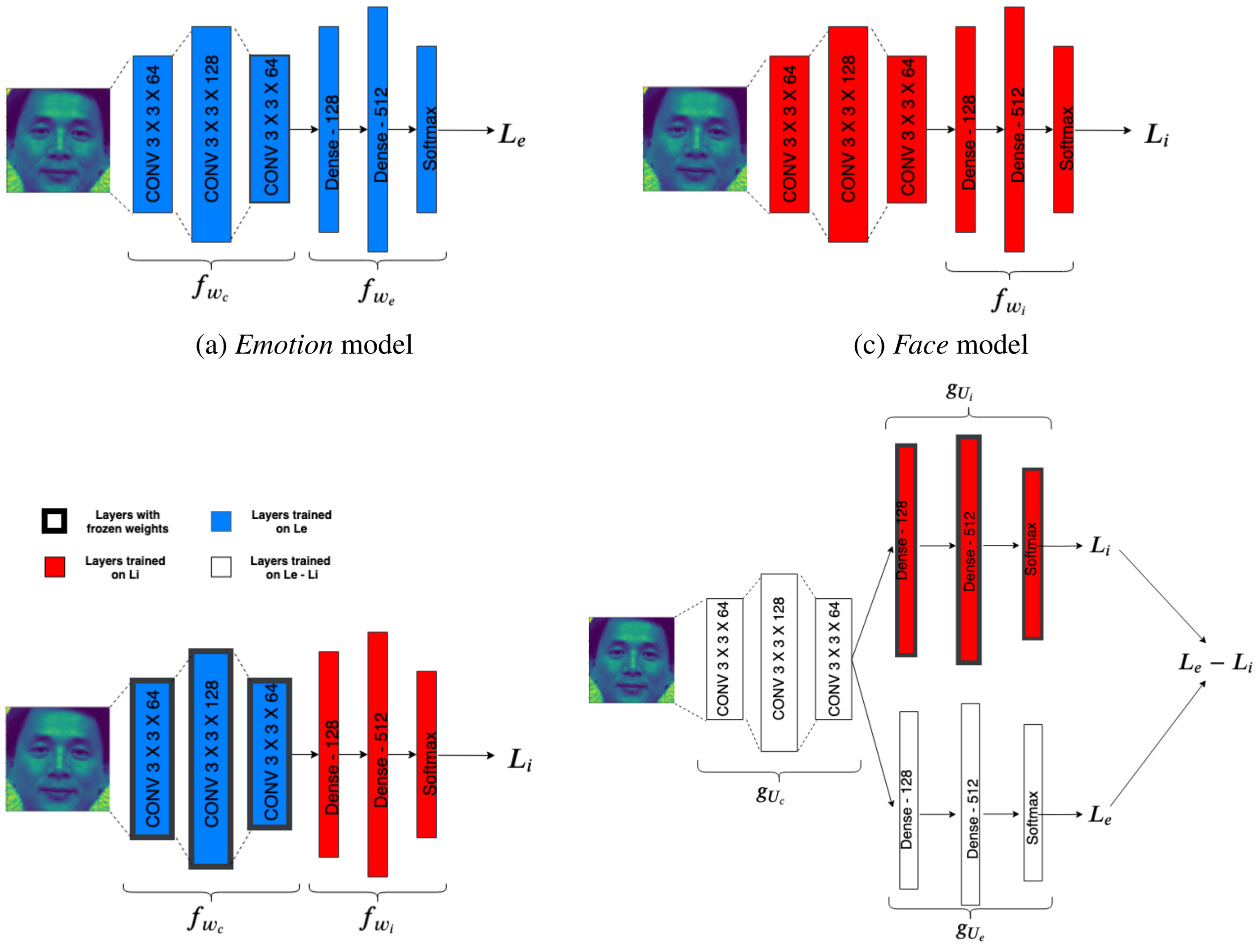
SIFT algorithm initially locates the position of key points in an image using extreme values in scale space. This step aims at getting rid of low contrast regions and edges which is followed by finding the most prominent regions of key points using direction of gradient of its neighbouring pixels. These steps make sure that SIFT features are both scale and rotation invariant as shown in Figure: 2.1. Each key point is then represented as a 128 dimensional SIFT vector. SIFT features can describe an image as a bundle of these 128 dimensional key points.

Since, the number of key points may vary from image to image, the dimensions of the image representation may not be consistent. We use Bag of Features or Visual Bag of Words to cluster together the descriptors which represents the same or similar feature of the image/object. This is performed using K-Means Clustering algorithm over all the descriptors of the dataset. Similar descriptor vectors are clustered together forming K clusters in total. Finally, each image is represented as a K dimensional vector where each dimension represents the number of descriptors present in the image belonging to a particular cluster.

These k dimensional representation of images are the trained for both emotion recognition task and face recognition task using SVM. We perform an extensive grid search on SVM hyper-parameters including ’c value’, ’gamma value’ and ’type of kernel’ and the model with best test accuracy is reported. The number of clusters is also considered as another hyper parameter and it is included in the final grid search where the best model for each K is selected and the corresponding accuracy is reported.

2.2 Quantifying user-related information on emotion-specific CNN models

In this section, We will examine the degree to which user-specific information is embedded in image-based emotion recognition models (CNN). Let $\mathbf{x} \in R^{D\times D}$ be an input image and $f_{\mathbf{W_c}} : R^{D\times D} \rightarrow R^{D'\times D'}$, $D' \leq D$ be a transformation function parameterized with $\mathbf{W_c}$, which translates the image $\mathbf{x}$ into another image $\mathbf{x}' = f_{\mathbf{W_c}}(\mathbf{x}) \in R^{D'\times D'}$ of same or lower dimensionality. Also let $f_{\mathbf{W_e}} : R^{D'\times D'} \rightarrow R$ be a function that translates the transformed image $\mathbf{x}'$ into an emotion decision $y_e = f_{\mathbf{W_e}}\left(f_{\mathbf{W_c}}(\mathbf{x})\right) \in R$. The latter can be implemented through a CNN, denoted as “*Emotion*" (Fig. 2.2a), whose convolutional layers approximate the transformation $f_{\mathbf{W_c}}$ and subsequent fully-connected layers approximate the transformation $f_{\mathbf{W_e}}$. Our goal is to examine the degree of user-specific information embedded in the network. Given that the convolutional layers are the ones that usually enable transfer of prior knowledge between tasks [72, 73, 74], we will explore the amount of user-dependent information that is embedded in the convolutional transformation $f_{\mathbf{W_c}}(\mathbf{x})$ of the *Emotion* model. We will do this by adding an additional function $f_{\mathbf{W_i}} : R^{D'\times D'} \rightarrow R$, that translates the transformed image $\mathbf{x}' = f_{\mathbf{W_c}}(\mathbf{x})$ learned for the task of emotion classification to the task of user recognition. This can be implemented by freezing the convolutional layers of the *Emotion* model and further adding a set of fully connected layers to be learned for the task of user recognition, yielding the corresponding decision $y_i = f_{\mathbf{W_i}}\left(f_{\mathbf{W_c}}(\mathbf{x})\right) \in R$. This model will be referred to as “*Emotion2Face*" (Fig. 2.2b), since its convolutional layers are learned based on emotion classification and subsequent fully-connected layers for user recognition. We will compute the user recognition accuracy of the *Emotion2Face* model. High accuracy would indicate that a large degree of user-specific information is embedded in the emotion-based convolutional transformation, while low accuracy would reflect the opposite. We will further compare the performance of *Emotion2Face* model with a CNN fully trained for face recognition, referred to as “*Face*" (Fig. 2.2c). The *Face* model will serve as a baseline to quantify the amount of information specific to the user identity captured through the convolutional layers of a CNN fully trained



(a) *Emotion* model

(c) *Face* model

(b) *Emotion2Face* model

(d) *Hybrid* model

Figure 2.2: Schematic representation of the: (a) *Emotion* model, trained on emotion recognition; (b) *Emotion2Face* model, trained on emotion and fine-tuned on face identification; (c) *Face* model, trained on face identification; and (d) *Hybrid* model, trained on an iterative adversarial framework for privacy-preserving emotion recognition.

for face identification, as measured by the corresponding face identification accuracy.

In accordance to prior work which has proposed image blurring as an attempt to conceal a person's identity [75], we will further train the same models with blurred images as the input. We will use 2-dimensional Gaussian blurring with kernel size of $(5, 5)$ and $(7, 7)$ to train CNNs on emotion classification and user identification. We will also fine-tune the CNNs pre-trained on emotion recognition to the task of user identification, so that we can examine the degree of user-related information embedded in emotion recognition models trained with blurred images.

2.3 User anonymity-preserving emotion classification

We will design an image-based emotion classification model, which can degrade the identity of the user, while performing well for the task of emotion classification. For this we will use a multitask CNN architecture trained in an adversarial manner. The goal of the proposed framework is to learn an image transformation which can reduce the identity-specific information relevant to the privacy-aware (or sensitive) task, while preserving the information required for the utility task of emotion recognition.

In the following, let $\mathbf{x} \in R^{D \times D}$ be an input image, $g_{\mathbf{U_c}} : R^{D \times D} \rightarrow R^{D' \times D'}$ the convolutional transformation of the original image, $g_{\mathbf{U_e}} : R^{D' \times D'} \rightarrow R$ the transformation that leads to the emotion decision $y_e \in R$, and $g_{\mathbf{U_i}} : R^{D' \times D'} \rightarrow R$ the transformation that provides the user decision $y_i \in R$. These are implemented with a multitask CNN architecture containing a convolutional basis that approximates transformation $f_{\mathbf{U_c}}$ and is common among the two tasks. The convolutional basis is followed by two distinct sets of fully-connected layers, one for the task of emotion classification and the other for the task of user identification, implementing transformations $g_{\mathbf{U_e}}$ and $g_{\mathbf{U_i}}$, respectively. The proposed architecture will be denoted as “*Hybrid*" and is schematically represented in Fig. 2.2d.

Given that the convolutional layers of the CNN are able to preserve a high degree of information reproducable across many tasks [72, 73, 74], the multitask CNN architecture of the *Hybrid* model will be learned so that the convolutional transformation $f_{\mathbf{U_c}}$ can withhold as less information as possible for the task of identity recognition and preserve as much information as possible for the task of emotion classification. In this way, the convolutional transformation will ultimately be useful as the input to the subsequent emotion-specific fully-connected layers $f_{\mathbf{U_e}}$, but will not be useful to the user-specific layers $f_{\mathbf{U_i}}$ for the task of user identification. Taking these into account, the weights $\mathbf{U_c}$, $\mathbf{U_e}$, and $\mathbf{U_i}$ of the CNN should be learned such that:

$$\min_{\{\mathbf{U_c},\mathbf{U_e},\mathbf{U_i}\}} \{L_e\left(g_{\mathbf{U_e}}(g_{\mathbf{U_c}}(\mathbf{x})), y_e\right) - \alpha L_i\left(g_{\mathbf{U_i}}(g_{\mathbf{U_c}}(\mathbf{x})), y_i\right)\} \tag{2.1}$$

where α is the hyper-parameter that balances the trade-off between minimizing the emotion loss $L_e(\cdot,\cdot)$ and maximizing the user identity loss $L_i(\cdot,\cdot)$.

The optimization of (2.1) involves an inherent limitation which lies in the fact that in order to maximize the user identity loss, or minimize the user classification accuracy, it might be enough for the system to just assign zero to the combination of weights $\mathbf{U_i}$ of the user-specific transformation $g_{\mathbf{U_i}}$, which will result in an "artificially" successful adversarial learning. However, we would like to learn a convolutional transformation $g_{\mathbf{U_c}}$ that can degrade the identity of the user no matter how good the user-specific fully-connected layers $g_{\mathbf{U_i}}$ are in the user recognition task. For this reason, we will employ an iterative adversarial procedure to learn a convolutional transformation against a number of good face recognition models and obtain a final transformation based on which there is no fully connected layer able to extract user identity-specific information. We will implement these by first freezing the user-specific weights $\mathbf{U_i}$ and jointly learning the weights of the convolutional transformation and the emotion-specific weights $\{\mathbf{U_c},\mathbf{U_e}\}$:

$$\min_{\{\mathbf{U_c},\mathbf{U_e}\}}\{L_e\left(g_{\mathbf{U_e}}(g_{\mathbf{U_c}}(\mathbf{x})),y_e\right)-\beta L_i\left(g_{\mathbf{U_i}}(g_{\mathbf{U_c}}(\mathbf{x})),y_i\right)\} \tag{2.2}$$

where β balances between positive emotion loss and negative face identity loss. We will then freeze $\{\mathbf{U_c},\mathbf{U_e}\}$ and learn the user-specific weights $\mathbf{U_i}$, such that:

$$\min_{\{U_i\}}\{L_i\left(g_{U_i}(g_{U_c}(\mathbf{x})),y_i\right)\} \tag{2.3}$$

This prevents $\mathbf{U_i}$ from becoming zero and allows us to obtain a competent user-specific transformation, which can then serve as a basis to re-learn a privacy-preserving convolutional transformation $\mathbf{U_c}$ based on (2.2). The process of alternating between the learning of weights $\{\mathbf{U_c},\mathbf{U_e}\}$ and $\mathbf{U_i}$ using (2.1) and (2.2) is repeated T times, until the emotion and user identity losses $L_e(\cdot,\cdot)$ and $L_i(\cdot,\cdot)$ converge, as outlined in Algorithm 1.

After learning a convolutional transformation $\mathbf{U_c^*}$, we will further evaluate its ability to eliminate user-specific information. We will do this by adding a set of new fully-connected layers that

```
Require: Image $\mathbf{x}$, emotion label $y_e$, user label $y_i$, hyperparameters $\alpha$, $\beta$, $T$
Initialize $\mathbf{U_c}$ (convolutional weights), $\mathbf{U_e}$ (emotion classification weights), $\mathbf{U_i}$ (face identification weights) with multitask learning:
  $\min_{\{\mathbf{U_c},\mathbf{U_e},\mathbf{U_i}\}}\{L_e\left(g_{\mathbf{U_e}}(g_{\mathbf{U_c}}(\mathbf{x})), y_e\right) + \alpha L_i\left(g_{\mathbf{U_i}}(g_{\mathbf{U_c}}(\mathbf{x})), y_i\right)\}$
for $t = 1, \ldots, T$ do
    Freeze $\mathbf{U_i}$
    Learn $\mathbf{U_c}$ and $\mathbf{U_e}$ using adversarial loss:
      $\min_{\{\mathbf{U_c},\mathbf{U_e}\}}\{L_e\left(g_{\mathbf{U_e}}(g_{\mathbf{U_c}}(\mathbf{x})), y_e\right) - \beta L_i\left(g_{\mathbf{U_i}}(g_{\mathbf{U_c}}(\mathbf{x})), y_i\right)\}$
    Freeze $\mathbf{U_c}$ and $\mathbf{U_e}$
    Learn $\mathbf{U_i}$ using user identity loss: $\min_{\{U_i\}}\{L_i\left(g_{U_i}(g_{U_c}(\mathbf{x})), y_i\right)\}$
end for
```

Algorithm 1: Adversarial learning for anonymity-preserving emotion recognition

implement transformation $h_{\mathbf{V_i}}$, which will yield a user-specific decision $h_{\mathbf{V_i}}(g_{\mathbf{U_c}}(\mathbf{x}))$. The new weights $\mathbf{V_i}$ will be learned such that they minimize the user classification loss $L_i\left(h_{\mathbf{V_i}}(g_{\mathbf{U_c}}(\mathbf{x})), y_i\right)$. Through this model, which will be referred to as "*Hybrid2Face*," we will be able to quantify the degree of user-specific information preserved in the learned convolutional transformation. A convolutional transformation which is successful in eliminating user-related information will yield low accuracy in the *Hybrid2Face* model.

We note that the proposed training approach with alternate weight freezing is slightly different compared to previous work on anonymized human activity recognition [61, 62, 63], in which all weights were simultaneously learned during training according to (2.1). This previously proposed approach is slightly more prone to local minima, can yield unstable solutions, and takes more time to propagate the error is propagated to the convolutional layers, therefore requiring a larger number of iterations. We will compare our proposed approach, that involves alternate weight freezing, with previously proposed training in which weight freezing was not included as part of the process [61, 62, 63], also denoted as "*HybridNoFreeze*." We will further concatenate a new set of user-dependent weights to the learned convolutional transformation of the *HybridNoFreeze* model, which will be learned based on the task of user identification. This will be called "*Hybrid2FaceNoFreeze*" and will give us an estimate of the amount of user-dependent information that is being preserved in the convolutional transformation of the *HybridNoFreeze* model.

2.4 Evaluation

We evaluate the proposed approach in both a quantitative and qualitative way. We will first compute the emotion classification and user identification accuracy for the proposed *Hybrid* model and compare this against the baseline models that use image blurring (i.e., *Emotion-Blur-5*, *Emotion-Blur-7*), as well as against the hybrid model that was trained without an alternate freezing of the weights (i.e., *HybridNoFreeze*). We will further evaluate the stability of the proposed training mechanism using alternate weight freezing by plotting the emotion and user identity loss functions, $L_e(\cdot, \cdot)$ and $L_i(\cdot, \cdot)$, across the number of iterations. The desired result would be that the emotion loss increases over time, while the user identity loss increases and remains consistently high as iterations progress. Our results will be also evaluated through visual inspection. We will visualize the transformed image resulting by the proposed training framework and compare it against the the hybrid model that was trained without an alternate freezing of the weights (i.e., *HybridNoFreeze*). Ideally, the transformed image that has completely lost the identity specific information should contain only the regions that are necessary for emotion recognition, which likely correspond to the eyes and mouth [76].

3. EXPERIMENTS

In this chapter, we first outline the data used in our experiments (Section 3.1). We then describe the experimental setting of our approach and the baselines that were used (Section 3.2). We present our findings on the user-specific information embedded in emotion recognition CNNs, as well as in comparison to the baseline approaches that employ image blurring (Section 4.2). Finally, we present the results of the proposed privacy-preserving multitask CNN trained with an adversarial loss (Section 4.3). We compare the proposed alternate weight freezing of the adversarial learning approach to an adversarial learning trained without any weight freezing.

3.1 Data description and pre-processing

Our experiments involve two datasets of facial images, the Japanese Female Facial Expression (JAFFE) database [69] and the Yale Face Dataset (YALE) [70]. We chose those two datasets, since they provide a constrained framework to evaluate our approach, since they include images taken by cameras in close proximity to the user's face preserving a high amount of identity-specific information and containing both emotion and identity labels. In JAFFE dataset, we have 10 female users and 7 emotions (neutral, sadness, surprise, happiness, fear, anger, and disgust), with a total of 213 static images. All images in the dataset included labels for both emotion and user identity, therefore they were all used. For the YALE dataset, we used only the images which included both the user and emotion labels, resulting in a total of 60 images of 15 male and female users and 4 emotion classes (happy, sad, normal, surprised). Since the number of images in both datasets was small for a CNN model to be adequately trained, we used data augmentation techniques related to random rotation, horizontal flip, and random noise addition. This resulted in 3038 and 3033 images for JAFFE and YALE datasets, respectively [77].

3.2 Experimental setting

We used a hybrid grid search for generating baseline results using SIFT and SVM keeping 'c value', 'gamma value', 'type of kernel' and 'number of clusters' as tunable hyper-parameters. For

CNN based models, we used 10-fold cross-validation for our experiments retaining the same set of train and test images for each fold across all systems. We also made sure that no samples generated from the same original image after data augmentation are concurrently present in the test set and train set. Based on the way our problem was formulated, we needed images from the same user in both the train and the test set, in order to perform the user identification task. For this reason, we were not able to perform a leave-one-subject-out cross-validation, which would involve separating users between the train and test sets. All architectures which were used to quantify identity-specific information in emotion recognition models (i.e., *Emotion*, *Face*, *Emotion2Face*, and their blurred counterparts) as well as the anonymity-preserving models of emotion recognition (i.e., *Hybrid*, *Hybrid2Face*, *HybridNoFreeze*, *Hybrid2FaceNoFreeze*) included 3 convolutional layers followed by 3 fully-connected layers. The ReLU activation function was used for all the hidden layers, while the output layer had a softmax activation. A 3×3 convolutional filter with a stride length of 3 was further employed. The number of nodes for each layer is depicted in Fig. 2.2. The hyperparameters balancing the ability of the *Hybrid* and *HybridNoFreeze* models to learn between the emotion classification and the user identity recognition tasks, as depicted in (2.1) and (2.2), were empirically set to $\alpha = 0.5$ and $\beta = 1$, respectively. The number of iterations for the adversarial learning optimization was $T = 70$ and $T = 40$ for the JAFFE and YALE datasets, respectively. Each iteration took 608 seconds using the NVIDIA GTX 1060 graphics card.

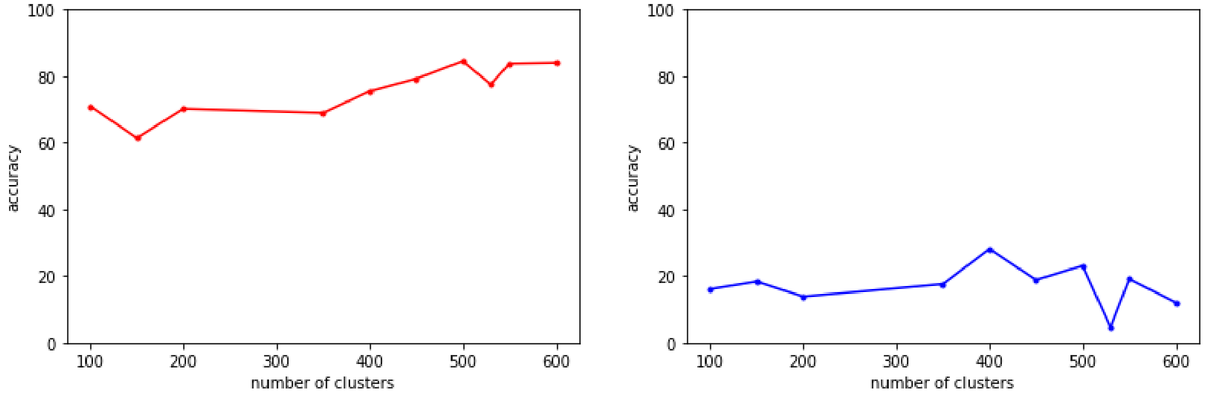
4. RESULTS

4.1 Quantifying user-related and emotion- related information captured by SIFT representation of image

We explore the quality of the features extracted by classical feature extraction methods such as Scale Invariant Feature Transformation (SIFT) for privacy preservation and emotion recognition (Section 2.1). The features extracted using such techniques are not task dependent and hence the transformation obtained using such extractors would be same for emotion and face recognition task. An ideal feature for privacy preservation would consistently achieve low face recognition accuracy and high emotion recognition accuracy. The results obtained are plotted for both YALE as shown in Figure 4.1 and JAFFE dataset as shown in Figure 4.2. We can see that we were able to achieve high face recognition accuracy for both YALE (Figure 4.1) and JAFFE dataset (Figure 4.2) while the emotion recognition accuracy was consistently low for both datasets. These accuracies would serve as baseline for deep learning models and our proposed algorithm. The best face recognition accuracy achieved using SIFT and SVM on YALE and JAFFE dataset was 84.46% and 99.50% respectively and the best emotion recognition accuracy achieved was 28.07% and 85.50% respectively.

4.2 Quantifying user-related information in emotion-specific models

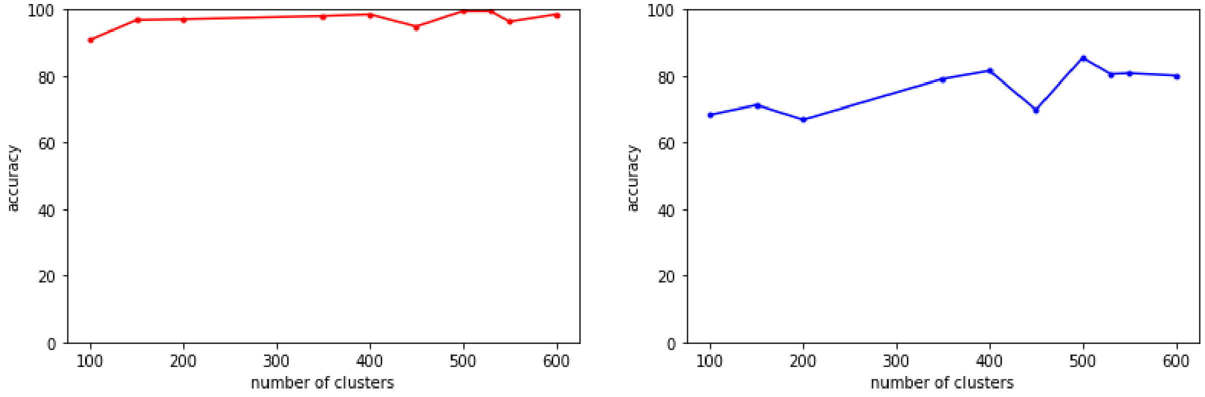
We explore the degree of user identity information which is embedded in the convolutional layers of an emotion-specific CNN (Section 2.2). Results obtained in the YALE and JAFFE datasets with unblurred and blurred images as the input are shown in Table 4.1. All results reflect simple classification accuracies, since the distribution of samples for the user and emotion categories was balanced for both datasets. High face recognition accuracy was achieved in both datasets (i.e., 96.25% and 99.26% for YALE and JAFFE, respectively) indicating that the corresponding task is relatively easy. Emotion classification on the other hand depicts higher accuracy for JAFFE compared to YALE, potentially due to the high variability of the latter. Still, the accuracies obtained



(a) Face recognition accuracy

(b) Emotion recognition accuracy

Figure 4.1: Classification accuracy on YALE dataset for (a) face recognition (b) emotion recognition using the Scale Invariant Feature Transformation (SIFT) features with Support Vector Machines (SVM). SIFT descriptors were extracted using K-Means with varying number of clusters.



(a) Face recognition accuracy

(b) Emotion recognition accuracy

Figure 4.2: Classification accuracy on JAFFE dataset for (a) face recognition (b) emotion recognition using the Scale Invariant Feature Transformation (SIFT) features with Support Vector Machines (SVM). SIFT descriptors were extracted using K-Means with varying number of clusters.

were consistently higher than the corresponding accuracies of model trained using SVM on features extracted using SIFT. This shows that the CNN models can help improve the classification accuracy and thus, the use of such models are indispensable. But, when the *Emotion* model is fine-tuned for face identity recognition, the corresponding accuracies of the *Emotion2Face* model remain high (i.e., 98.77% for JAFFE, 74.18% for YALE) suggesting that there is a substantial amount of identity-specific information in the convolutional layer of the *Emotion* model and sup-

(a) YALE Dataset

Model	**No Blur**	**Blur (5×5)**	**Blur (7×7)**
Emotion	47.34	37.09	37.34
Face	96.24	98.24	97.74
Emotion2Face	74.18	80.70	81.45

(b) JAFFE Dataset

Model	**No Blur**	**Blur (5×5)**	**Blur (7×7)**
Emotion	89.95	82.55	87.22
Face	99.26	99.754	99.75
Emotion2Face	98.77	99.75	99.26

Table 4.1: Emotion and user classification accuracies in YALE and JAFFE datasets, as obtained from a convolutional neural network (CNN) trained for the emotion-specific task (*Emotion*), a CNN trained for the user identity task (*Face*), as well as the CNN initially trained on emotion and fine-tuned on user identification (*Emotion2Face*). Experiments are performed without image blurring (No Blur), as well as using a 2-dimensional Gaussian kernel of 5×5 and 7×7 to blur the original images (Blur (5×5) and Blur (7×7)).

porting user re-identification concerns.

When we use blurred images as an input to the models, we observe a decrease on emotion recognition accuracy for the YALE dataset (i.e., 10% absolute decrease). Similar results were obtained for JAFFE with the corresponding drop in performance ranging between 2% to 7%. These indicate that the blurring of the original images degrades the task of emotion classification, since the emotional information can be sensitive to fine-grain fluctuations in the image, which tend to vanish with blurring. On the other hand, the face recognition accuracy does not appear to decrease (i.e., 1-2% and 0.5% absolute difference for YALE and JAFFE, respectively), which suggests that the identity-specific information might also depend on more coarse image characteristics, which are preserved even after blurring. These indicate that image blurring might not be an effective approach to the problem of anonymity-preserving emotion recognition, since blurring appears to degrade emotion recognition performance while allowing for user re-identification.

4.3 Evaluating user anonymity-preserving emotion classification

We further report the emotion and user classification performance that was achieved using the proposed multitask CNN architecture with adversarial learning and alternate weight freezing (i.e., *Hybrid*), as well as the baseline model that was trained without freezing any of its weights (i.e., *HybridNoFreeze*) (Table 4.2). The proposed approach (i.e., *Hybrid*) yields high emotion classification accuracy (i.e., 62.65% for YALE, 92.62% for JAFFE) and low user classification accuracy (i.e., 0% for YALE, 5.65% for JAFFE), which was the desired outcome. In contrast, the baseline model that did not use the alternate weight freezing training (i.e., *HybridNoFreeze*) was not able to eliminate user-related information yielding high user classification accuracies in both datasets (i.e., 67.41% in YALE, 98.77% in JAFFE). We further evaluate the ability of the learned convolutional layers of the *Hybrid* and *HybridNoFreeze* models to preserve user-specific information when new fully-connected layers are re-trained on the user classification task. The corresponding user classification accuracy of the *Hybrid2Face* architecture remained fairly low (i.e., 1.25% for YALE, 31.31% for JAFFE), while the same metric for the *Hybrid2FaceNoFreeze* was high (i.e., 84.46% for YALE, 99.26% for JAFFE). This indicates the ability of the proposed training with alternate weight freezing not only to eliminate user-specific information during training, but also to prevent re-identification of the user even after additional user-based learning. On the other side, the baseline model achieved an “artificially" successful adversarial learning with substantial face-dependent information still remaining in its convolutional base. We note that Table 4.2 does not report emotion classification accuracy for the *Hybrid2Face* and *Hybrid2FaceNoFreeze* models, since these are trained for user identification.

The ability of our proposed *Hybrid* model to converge is further depicted by plotting the emotion and user identity loss functions against the number of learning iterations T (Fig. 4.3). We can see that the *Hybrid* model was able to achieve high user identification loss after approximately 20 iterations. We further observe that the proposed model was able to keep emotion classification loss consistently low, attesting to its ability to successfully recognize emotions.

Once we have an indication that our model has converged, it is important to also demonstrate *(a) YALE Dataset*

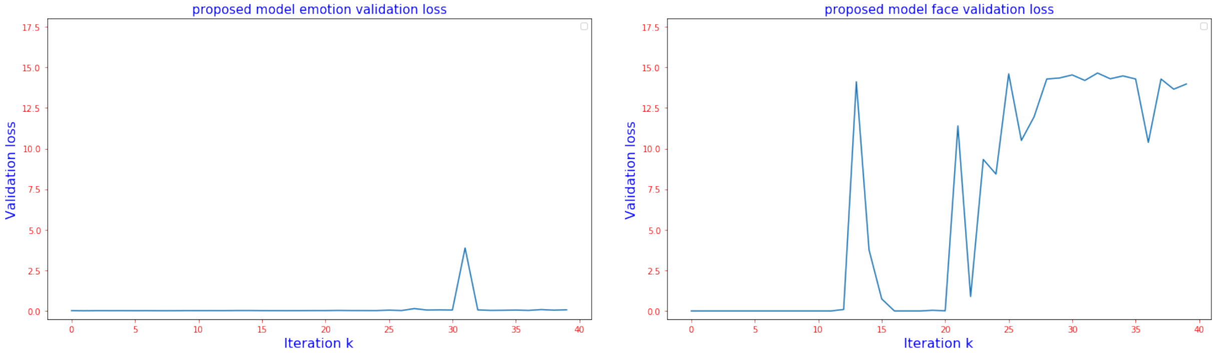
Model	**Emotion classification accuracy**	**User classification accuracy**
HybridNoFreeze (baseline)	25.06	67.41
Hybrid (proposed)	62.65	0
Hybrid2FaceNoFreeze (baseline)	N/A	84.46
Hybrid2Face (proposed)	N/A	1.25

(b) JAFFE Dataset

Model	**Emotion classification accuracy**	**User classification accuracy**
HybridNoFreeze (baseline)	86.27	98.77
Hybrid (proposed)	92.62	5.65
Hybrid2FaceNoFreeze (baseline)	N/A	99.26
Hybrid2Face (proposed)	N/A	31.13

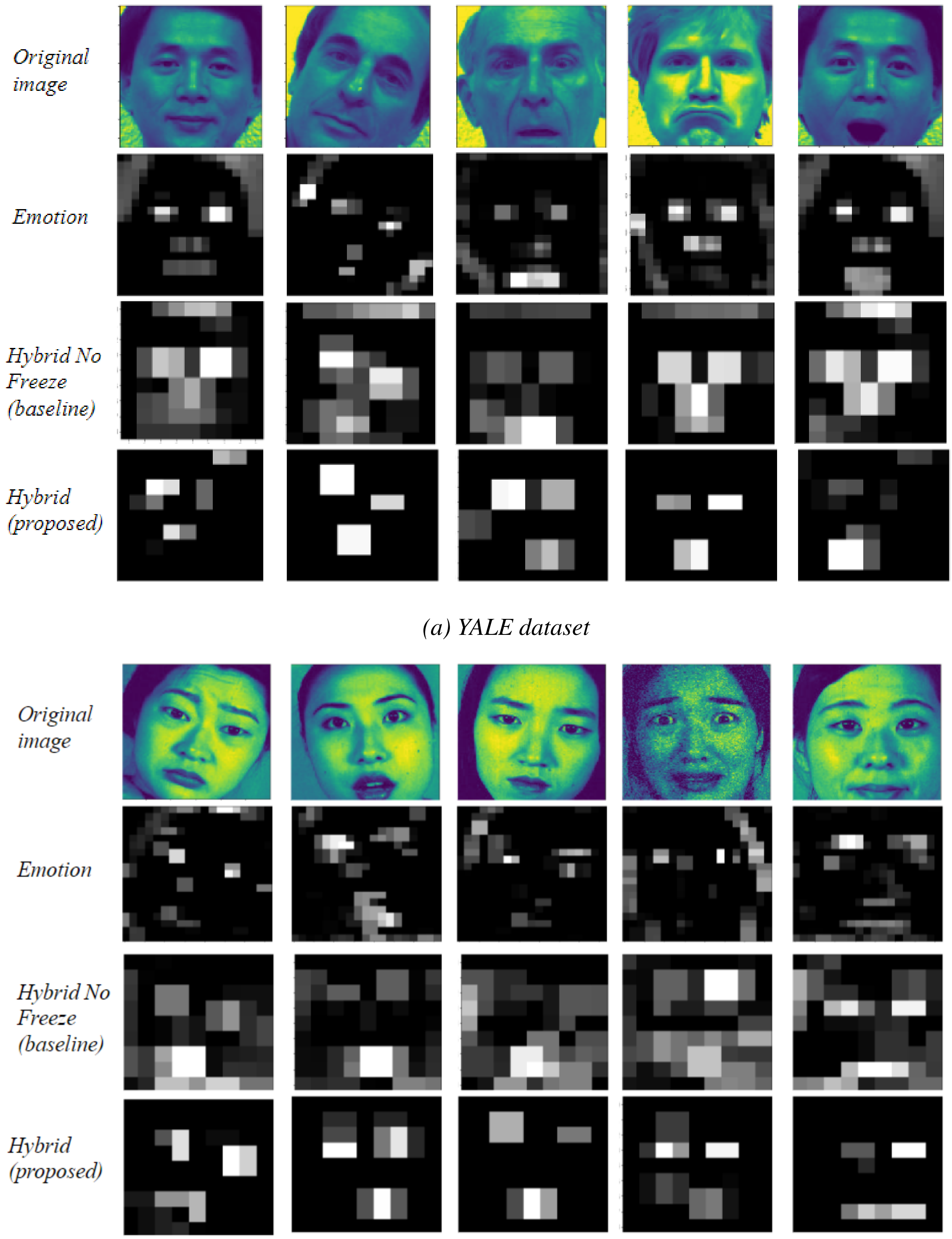
Table 4.2: Emotion and user classification accuracies in YALE and JAFFE datasets obtained by the proposed anonymity-preserving emotion recognition model with alternate weight freezing (*Hybrid*), as well as by the anonymity-preserving emotion recognition model without freezing of its weights during training (*HybridNoFreeze*). The ability of the proposed model to degrade user identity information is further evaluated by adding a set of new fully-connected layers on the learned convolutional transformation and fine-tuning for user identification. For the latter task, adversarial learning is performed with and without alternate weight freezing (*Hybrid2Face* and *Hybrid2FaceNoFreeze*, respectively).

that the model can converge to a good solution. For this reason, apart from the quantitative results (Table 4.2, Fig. 4.3), we also visualize the convolutional output. Results are depicted in Fig. 4.4. We can see that the output of the convolutional base trained using our proposed framework (*Hybrid*) is more concentrated towards the eyes and mouth region, whereas all other information seems to be completely lost. This aligns with previous findings which suggest that the emotion information tends to be depicted in a person's mouth and eye region of the face [76]. Moreover, it can be sometimes deceiving for the network to focus on other areas of the face, potentially further justifying the relatively higher emotion accuracy of our proposed model compared to the *HybridNoFreeze* baseline. In contrast to these, the output of the convolutional base of the *Emotion* model has retained almost all facial information, including the contour of the face, which was not retained by the *Hybrid* model. This is also the case for rotated images, suggesting that our model is not restricted to the value of pixels at a particular location of an image and has actually learned the shape of important facial landmarks irrespective of their relative position.



(a) Validation loss per iteration for emotion recognition(b) Validation loss per iteration for face recognition

Figure 4.3: Schematic representation of the loss function for (a) emotion recognition (b) face recognition of the proposed privacy-preserving emotion recognition models (*Hybrid*), which was trained on an adversarial manner and included alternate freezing between the convolutional and emotion-specific weights $\{\mathbf{U_c}, \mathbf{U_e}\}$ and the user-specific weights $\mathbf{U_i}$.



(a) YALE dataset

(b) JAFFE dataset

Figure 4.4: Examples of the original and transformed images obtained from the YALE and JAFFE datasets after applying a convolutional transformation learned by a convolutional neural network (CNN) solely trained for the emotion-specific task (*Emotion*), the proposed anonymity-preserving emotion recognition model with alternate weight freezing (*Hybrid*), and the baseline anonymity-preserving emotion recognition model without freezing of its weights during training (*Hybrid-NoFreeze*).

5. DISCUSSION

In this thesis, we studied the interplay between user identification risks and emotion utility in images using two datasets. We observed that user-dependent information is highly embedded in emotion recognition models learned using state-of-the-art representation models, such as CNNs. We further proposed an anonymity-preserving emotion recognition model that learns a convolutional basis able to perform well for emotion recognition, but poorly for user classification. Our results indicate that the proposed approach can achieve these goals and minimize user re-identification risks. Despite the promising results, the current study depicts the following limitations. We presented a first proof-of-concept that privacy-preservation through user anonymization is possible in emotion recognition models. For this reason, we evaluated our approach on two small datasets, the YALE [70] and JAFFE [69], which were collected in laboratory conditions and included acted emotions, therefore the considered data are clean and potentially not representing natural emotional expression. It would be beneficial to test the proposed approach in real-life applications with spontaneous emotional expressions obtained “in-the-wild" and using a larger number of more noisy samples. Examples of such datasets that could be used as part of future work include the CAS-PEAL Face Database [78] and the Indian Movie Face database [79], which contain approximately 33,000 and 99,000 images with a large number of users and conditions. Another limitation of this study lies in the fact that static images were taken into account. However, emotion is dynamically changing, therefore future work will concentrate on extending these techniques to video signals. The inherently unstable nature of adversarial learning can present various challenges related to finding the optimal number of iterations during the optimization to achieve a close-to-optimal solution. We used techniques like alternate weight freezing to tackle this, but additional experimentation would be useful in order to obtain more robust results. Finally, this work presented an experimental study on how emotion and user identity are inter-related to each other and embedded in image information. Providing a theoretical framework with privacy guarantees was outside the scope of this work. As part of our future work, we plan to provide a formal approach on user anonymity and identify potential privacy guarantees of our approach.

Although the thesis is focused on privacy-preserving emotion recognition system, our approach might not be necessarily limited to this particular application. Privacy and user anonymization are inherent issues in several behavioral studies involving psychological and cognitive outcomes. For example, leveraging publicly available data, we can develop privacy-aware systems for stress detection, cognitive demand recognition, and performance prediction [80, 81]. Beyond the image-based behavioral recognition, the proposed privacy-aware adversarial framework could be used for speech-based emotion recognition [66, 67, 68]. Instead of learning a convolutional transformation of an input facial image, we can learn a anonymity-preserving convolutional transformation of the 2-dimensional speech spectrogram. This might yield more reliable emotion recognition systems compared to current practices, which use 1-dimensional spectrotemporal acoustic features. The interplay between facial and acoustic information is also of high interest, especially in behavioral applications, where multimodality is an inherent part of the phenomenon that is being studied. Multimodal anonymity-preserving behavior recognition systems can be particularly helpful in real-life applications that involve IoT devices, such as hospitals, work offices, and classrooms [82, 83, 84].

Privacy-preservation is particularly relevant to IoT devices. The proposed framework was computationally and storage-wise quite expensive, since the *Hybrid* model involved the learning of approximately 1.6 million parameters. Despite the high computational capability and storage capacity of today's IoT devices, efficiency is a prominent issue. Designing compressed privacy-preserving behavior recognition systems remains an open problem, which will be explored in depth as part of our future work. Potential solutions toward this include the use of shallow neural networks and larger convolutional kernels, which can compress the input images without preserving high spatial granularity as presented in this work.

6. CONCLUSIONS

We have examined the interplay between emotion and user identity specific information in image-based CNNs trained for the task of emotion recognition. Our results indicate that CNNs, even when trained for a different task, tend to preserve a significant amount of user-related information, therefore presenting high user re-identification risks. We have also designed a user anonymity-preserving emotion recognition model using a multitask CNN architecture trained with an adversarial learning approach. Training of the proposed CNN was performed in an iterative way with alternate weight freezing so that the convolutional part of the network $g_{\mathbf{U_c}}$ learns to eliminate user identity related information from any potential user-related transformation $g_{\mathbf{U_i}}$. Results obtained on two publicly available datasets, YALE [70] and JAFFE [69], indicate the feasibility of our proposed approach in learning anonymity-preserving convolutional transformations which can perform well for the task of emotion recognition. Implications of our work can inform privacy-preserving machine learning across the span of various behavioral applications, including psychological and cognitive outcomes, where utility-based and sensitive information are highly inter-dependent, providing a foundation toward more secure and anonymous behavior recognition systems.

REFERENCES

[1] U. W. Ebner-Priemer, S. Koudela, G. Mutz, and M. K. Kanning, “Interactive multimodal ambulatory monitoring to investigate the association between physical activity and affect,” *Frontiers in Psychology*, vol. 3, p. 596, 2013.

[2] S. Narayanan and P. G. Georgiou, “Behavioral signal processing: Deriving human behavioral informatics from speech and language,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[3] V. A. Block, E. Pitsch, P. Tahir, B. A. Cree, D. D. Allen, and J. M. Gelfand, “Remote physical activity monitoring in neurological disease: a systematic review,” *PloS One*, vol. 11, no. 4, p. e0154335, 2016.

[4] R. Paradiso, A. Bianchi, K. Lau, and E. Scilingo, “Psyche: personalised monitoring systems for care in mental health,” in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 3602–3605, IEEE, 2010.

[5] E. Leon, M. Montejo, and I. Dorronsoro, “Prospect of smart home-based detection of subclinical depressive disorders,” in *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pp. 452–457, IEEE, 2011.

[6] B. Preschl, B. Wagner, S. Forstmeier, and A. Maercker, “E-health interventions for depression, anxiety disorder, dementia, and other disorders in old age: A review,” *Journal of CyberTherapy and Rehabilitation*, vol. 4, pp. 371–86, 2011.

[7] J. Gideon, E. M. Provost, and M. McInnis, “Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2359–2363, IEEE, 2016.

[8] S. Abdullah and T. Choudhury, “Sensing technologies for monitoring serious mental illnesses,” *IEEE MultiMedia*, vol. 25, no. 1, pp. 61–75, 2018.

[9] N. Kshetri and J. M. Voas, “Cyberthreats under the bed,” *IEEE Computer*, vol. 51, no. 5, pp. 92–95, 2018.

[10] J.-Y. Lee, W.-C. Lin, and Y.-H. Huang, “A lightweight authentication protocol for internet of things,” in *2014 International Symposium on Next-Generation Electronics (ISNE)*, pp. 1–2, IEEE, 2014.

[11] P. Porambage, C. Schmitt, P. Kumar, A. Gurtov, and M. Ylianttila, “Pauthkey: A pervasive authentication protocol and key establishment scheme for wireless sensor networks in distributed iot applications,” *International Journal of Distributed Sensor Networks*, vol. 10, no. 7, p. 357430, 2014.

[12] K. Benitez and B. Malin, “Evaluating re-identification risks with respect to the hipaa privacy rule,” *Journal of the American Medical Informatics Association*, vol. 17, no. 2, pp. 169–177, 2010.

[13] S. Lalitha, D. Geyasruti, R. Narayanan, and M. Shravani, “Emotion detection using mfcc and cepstrum features,” *Procedia Computer Science*, vol. 70, pp. 29–35, 2015.

[14] Y. Li, T. Zhao, and T. Kawahara, “Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning,” *Proc. Interspeech 2019*, pp. 2803–2807, 2019.

[15] M. Dahmane and J. Meunier, “Emotion recognition using dynamic grid-based hog features,” in *Face and Gesture 2011*, pp. 884–888, IEEE, 2011.

[16] M. A. Turk and A. P. Pentland, “Face recognition using eigenfaces,” in *Pattern Recognition*, 1991.

[17] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, “Dynamics of facial expression extracted automatically from video,” in *IEEE Conference on Computer Vision and Pattern Recognition: Workshop on Face Processing in Video*, 2004.

[18] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, and C. Fookes, “Deep spatio-temporal features for multimodal emotion recognition,” in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1215–1223, IEEE, 2017.

[19] S. Ghosh, A. Dhall, and N. Sebe, “Automatic group affect analysis in images via visual attribute and feature networks,” in *25th IEEE International Conference on Image Processing (ICIP)*, 2018.

[20] M. Huh, P. Agrawal, and A. A. Efros, “What makes imagenet good for transfer learning?,” *arXiv preprint arXiv:1608.08614*, 2016.

[21] M. Chamikara, P. Bertok, I. Khalil, D. Liu, and S. Camtepe, “Local differential privacy for deep learning,” *arXiv preprint arXiv:1908.02997*, 2019.

[22] B. Zhou, J. Pei, and W. Luk, “A brief survey on anonymization techniques for privacy preserving publishing of social network data,” *ACM Sigkdd Explorations Newsletter*, vol. 10, no. 2, pp. 12–22, 2008.

[23] D. Koukis, S. Antonatos, D. Antoniades, E. P. Markatos, and P. Trimintzios, “A generic anonymization framework for network traffic,” in *2006 IEEE International Conference on Communications*, vol. 5, pp. 2302–2309, IEEE, 2006.

[24] T. Heer, O. Garcia-Morchon, R. Hummen, S. L. Keoh, S. S. Kumar, and K. Wehrle, “Security challenges in the ip-based internet of things,” *Wireless Personal Communications*, vol. 61, no. 3, pp. 527–542, 2011.

[25] B. Brumen, M. Heričko, A. Sevčnikar, J. Završnik, and M. Hölbl, “Outsourcing medical data analyses: can technology overcome legal, privacy, and confidentiality issues?,” *Journal of medical Internet research*, vol. 15, no. 12, p. e283, 2013.

[26] X. Liu, R. Lu, J. Ma, L. Chen, and B. Qin, “Privacy-preserving patient-centric clinical decision support system on naive bayesian classification,” *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 2, pp. 655–668, 2015.

[27] Y. Lindell, “Secure multiparty computation for privacy preserving data mining,” in *Encyclopedia of Data Warehousing and Mining*, pp. 1005–1009, IGI Global, 2005.

[28] F. A. N. Pathak and S. B. S. Pandey, “An efficient method for privacy preserving data mining in secure multiparty computation,” in *2013 Nirma University International Conference on Engineering (NUiCONE)*, pp. 1–3, IEEE, 2013.

[29] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, “Practical secure aggregation for privacy-preserving machine learning,” in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, ACM, 2017.

[30] R. C. Geyer, T. Klein, and M. Nabi, “Differentially private federated learning: A client level perspective,” *arXiv preprint arXiv:1712.07557*, 2017.

[31] A. Gkoulalas-Divanis, G. Loukides, and J. Sun, “Publishing data from electronic health records while preserving privacy: A survey of algorithms,” *Journal of Biomedical Informatics*, vol. 50, pp. 4–19, 2014.

[32] Z. Ji, X. Jiang, S. Wang, L. Xiong, and L. Ohno-Machado, “Differentially private distributed logistic regression using private and public data,” *BMC medical genomics*, vol. 7, no. 1, p. S14, 2014.

[33] H. Li, L. Xiong, L. Ohno-Machado, and X. Jiang, “Privacy preserving rbf kernel support vector machine,” *BioMed Research International*, vol. 2014, 2014.

[34] S. Simmons and B. Berger, “Realizing privacy preserving genome-wide association studies,” *Bioinformatics*, vol. 32, no. 9, pp. 1293–1300, 2016.

[35] Z. Wu, Z. Wang, Z. Wang, and H. Jin, “Towards privacy-preserving visual recognition via adversarial training: A pilot study,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 606–624, 2018.

[36] J. Steil, M. Koelle, W. Heuten, S. Boll, and A. Bulling, “PrivacEye: Privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features,” in *Pro-*

ceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, pp. 1–10, 2019.

[37] T. Shinzaki, I. Morikawa, Y. Yamaoka, and Y. Sakemi, "Iot security for utilization of big data: Mutual authentication technology and anonymization technology for positional data," *Fujitsu scientific & technical journal*, vol. 52, no. 4, pp. 52–60, 2016.

[38] A. Otgonbayar, Z. Pervez, and K. Dahal, "Toward anonymizing iot data streams via partitioning," in *2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, pp. 331–336, Oct 2016.

[39] N. Park and N. Kang, "Mutual authentication scheme in secure internet of things technology for comfortable lifestyle," *Sensors*, vol. 16, no. 1, p. 20, 2016.

[40] D. J. Wu, A. Taly, A. Shankar, and D. Boneh, "Privacy, discovery, and authentication for the internet of things," in *European Symposium on Research in Computer Security*, pp. 301–319, Springer, 2016.

[41] C. Castelluccia, E. Mykletun, and G. Tsudik, "Efficient aggregation of encrypted data in wireless sensor networks," in *The second annual international conference on mobile and ubiquitous systems: networking and services*, pp. 109–117, IEEE, 2005.

[42] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[43] A. Aristodimou, A. Antoniades, and C. S. Pattichis, "Privacy preserving data publishing of categorical data throughk-anonymity and feature selection," *Healthcare technology letters*, vol. 3, no. 1, pp. 16–21, 2016.

[44] S. Yoo, M. Shin, and D. Lee, "An approach to reducing information loss and achieving diversity of sensitive attributes in k-anonymity methods," *Interactive Journal of Medical Research*, vol. 1, no. 2, p. e14, 2012.

[45] R. Heatherly, L. V. Rasmussen, P. L. Peissig, J. A. Pacheco, P. Harris, J. C. Denny, and B. A. Malin, “A multi-institution evaluation of clinical profile anonymization,” *Journal of the American Medical Informatics Association*, vol. 23, no. e1, pp. e131–e137, 2015.

[46] S. Martínez, D. Sánchez, and A. Valls, “A semantic framework to protect the privacy of electronic health records with non-numerical attributes,” *Journal of Biomedical Informatics*, vol. 46, no. 2, pp. 294–303, 2013.

[47] A. Tamersoy, G. Loukides, M. E. Nergiz, Y. Saygin, and B. Malin, “Anonymization of longitudinal electronic medical records,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 3, pp. 413–423, 2012.

[48] S. Kim, H. Lee, and Y. D. Chung, “Privacy-preserving data cube for electronic medical records: An experimental evaluation,” *International journal of medical informatics*, vol. 97, pp. 33–42, 2017.

[49] G. Loukides and A. Gkoulalas-Divanis, “Utility-aware anonymization of diagnosis codes,” *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 1, pp. 60–70, 2012.

[50] S. Hughes, K. Wells, P. McSorley, and A. Freeman, “Preparing individual patient data from clinical trials for sharing: the glaxosmithkline approach,” *Pharmaceutical Statistics*, vol. 13, no. 3, pp. 179–183, 2014.

[51] G. Poulis, G. Loukides, S. Skiadopoulos, and A. Gkoulalas-Divanis, “Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints,” *Journal of biomedical informatics*, vol. 65, pp. 76–96, 2017.

[52] M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings*, vol. 4051. Springer, 2006.

[53] S. A. Vinterbo, A. D. Sarwate, and A. A. Boxwala, “Protecting count queries in study design,” *Journal of the American Medical Informatics Association*, vol. 19, no. 5, pp. 750–757, 2012.

[54] F. K. Dankar and K. El Emam, "The application of differential privacy to health data," in *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pp. 158–166, ACM, 2012.

[55] S. Simmons, C. Sahinalp, and B. Berger, "Enabling privacy-preserving gwass in heterogeneous human populations," *Cell systems*, vol. 3, no. 1, pp. 54–61, 2016.

[56] Y. Rahulamathavan and M. Rajarajan, "Efficient privacy-preserving facial expression classification," in *Transactions on Dependable and Secure Computing*, IEEE, 2015.

[57] G. Letournel, A. Bugeau, V.-T. Ta, and J.-P. Domenger, "Face de-identification with expressions preservation," in *International Conference on Image Processing (ICIP)*, IEEE, 2015.

[58] J. Wang, B. Amos, A. Das, P. Pillai, N. Sadeh, and M. Satyanarayanan, "A scalable and privacy-aware iot service for live video analytics," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 38–49, ACM, 2017.

[59] J. Steil, M. Koelle, W. Heuten, S. Boll, and A. Bulling, "Privaceye," *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications - ETRA '19*, 2019.

[60] S. A. Osia, A. S. Shamsabadi, A. Taheri, K. Katevas, S. Sajadmanesh, H. R. Rabiee, N. D. Lane, and H. Haddadi, "A hybrid deep learning architecture for privacy-preserving mobile analytics," *arXiv preprint arXiv:1703.02952*, 2017.

[61] Z. Wu, Z. Wang, Z. Wang, and H. Jin, "Towards privacy-preserving visual recognition via adversarial training: A pilot study," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[62] J. Hamm, "Minimax filter: learning to preserve privacy from inference attacks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4704–4734, 2017.

[63] N. Raval, A. Machanavajjhala, and L. P. Cox, "Protecting visual secrets using adversarial nets," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1329–1332, IEEE, 2017.

[64] C. Feutry, P. Piantanida, Y. Bengio, and P. Duhamel, "Learning anonymized representations with adversarial neural networks," *International Workshop on Machine Learning and Artificial Intelligence (MLAI)*, 2019.

[65] M. Bertran, N. Martinez, A. Papadaki, Q. Qiu, M. Rodrigues, G. Reeves, and G. Sapiro, "Adversarially learned representations for information obfuscation and inference," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, (Long Beach, California, USA), pp. 614–623, PMLR, 09–15 Jun 2019.

[66] P. Arora and T. Chaspari, "Exploring siamese neural network architectures for preserving speaker identity in speech emotion classification," in *Proceedings of the 4th International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*, pp. 15–18, 2018.

[67] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-preserving adversarial representation learning in asr: Reality or illusion?," in *Interspeech*, ISCA.

[68] R. Aloufi, H. Haddadi, and D. Boyle, "Emotionless: Privacy-preserving speech analysis for voice assistants," *CoRR*, vol. abs/1908.03632, 2019.

[69] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek, "The Japanese female facial expression (JAFFE) database," in *Proceedings of third international conference on automatic face and gesture recognition*, pp. 14–16, 1998.

[70] "Yale face database." `http://vision.ucsd.edu/content/yale-face-database`, Last accessed on 2019-10-17.

[71] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[72] S. Li, Z.-Q. Liu, and A. B. Chan, "Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 482–489, 2014.

[73] T. Zeng and S. Ji, "Deep convolutional neural networks for multi-instance multi-task learning," in *2015 IEEE International Conference on Data Mining*, pp. 579–588, IEEE, 2015.

[74] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423, 2016.

[75] H. Chi and Y. H. Hu, "Face de-identification using facial identity preserving features," in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 586–590, Dec 2015.

[76] C. Padgett and G. W. Cottrell, "Representing face images for emotion classification," in *Advances in neural information processing systems*, pp. 894–900, 1997.

[77] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, p. 60, Jul 2019.

[78] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The cas-peal large-scale chinese face database and baseline evaluations," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 38, no. 1, pp. 149–161, 2007.

[79] S. Setty, M. Husain, P. Beham, J. Gudavalli, M. Kandasamy, R. Vaddi, V. Hemadri, J. Karure, R. Raju, B. Rajan, *et al.*, "Indian movie face database: A benchmark for face recognition under wide variations," in *2013 fourth national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG)*, pp. 1–5, IEEE, 2013.

[80] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.

[81] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2011.

[82] D. V. Dimitrov, “Medical internet of things and big data in healthcare,” *Healthcare informatics research*, vol. 22, no. 3, pp. 156–163, 2016.

[83] F. Moreira, M. J. Ferreira, and A. Cardoso, “Higher education disruption through iot and big data: A conceptual approach,” in *International Conference on Learning and Collaboration Technologies*, pp. 389–405, Springer, 2017.

[84] I. Lee and K. Lee, “The internet of things (iot): Applications, investments, and challenges for enterprises,” *Business Horizons*, vol. 58, no. 4, pp. 431–440, 2015.