ROYAUME DU MAROC
Université Mohammed V - RABAT -
جامعة محمد الخامس – الرباط –
Faculté des sciences
كلية العلوم
CENTRE D'ETUDES DOCTORALES - SCIENCES ET TECHNOLOGIES

N° d'ordre : 3400

THESE

En vue de l'obtention du: ***DOCTORAT***

Structure de Recherche : Laboratoire de Recherche en Informatique et Télécommunications
Discipline : Sciences de l'ingénieur
Spécialité : Informatique et télécommunications

Présentée et soutenue le 16/12/2020 par:

Safaa AZZAKHNINI

Approches basées sur la fusion multimodale pour les systèmes multi-capteurs : de l'amélioration de la performance vers les préoccupations éthiques.

Moulay Driss RAHMANI	PES	Université Mohammed V-Rabat, Faculté des Sciences	Président
Lahoucine BALLIHI	PH	Université Mohammed V-Rabat, Faculté des Sciences	Directeur de thèse
Mohammed El HASSOUNI	PES	Université Mohammed V-Rabat, Faculté des Lettres et des Sciences Humaines	Rapporteur/ Examinateur
Ahmed Drissi El MALIANI	PH	Université Mohammed V-Rabat, Faculté des Sciences	Rapporteur/ Examinateur
Mohamed Nabil SAIDI	PH	Institut national de statiques et d'économie appliquée, Rabat	Rapporteur/ Examinateur
Hicham LAANAYA	PH	Université Mohammed V-Rabat, Faculté des Sciences	Examinateur
Ralf C. STAUDEMEYER	PH	Schmalkalden University, Faculty of Applied Sciences, Germany	Examinateur

Année Universitaire : 2019-2020

"*Science is a way of thinking much more than it is a body of knowledge.*"

Carl Sagan

Acknowledgements

This thesis has been prepared within the laboratory of research in computer science and telecommunications (LRIT) at Mohammed V University under the direction and supervision of Pr. Lahoucine BALLIHI.
First, I would like to thank my thesis advisor, **Pr. Lahoucine BALLIHI**, qualified professor of computer science at Mohamed V Univerisity, for giving me the opportunity to do my PhD, forall his guidance and trust to follow my research interests.
I would like to express my gratitude to **Pr. Moulay Driss RAHMANI**, professor of higher education of computer science at Mohammed V University in Rabat for his feedback, assistance and for accepting to be the jury president for my defence.
My sincere appreciation and gratitude to **Pr. Mohammed ELHASSOUNI**, professor of of higher education of computer science at Mohammed V University in Rabat for accepting to report and examine my thesis and for his valuable comments and feedback.
Many thanks to **Pr. Ahmed Drissi El MALIANI**, qualified professor of computer science at Mohammed V University in Rabat for accepting to report and examine my thesis and for his valuable feedback and comments.
I would like to thank **Pr. Mohamed Nabil SAIDI**, qualified professor of computer science at the National Institute of Statistics and Applied Economics (INSEA) for accepting to report and examine my research and for his insightful comments.
Many thanks to **Pr. Hicham LAANAYA**, qualified rofessor of computer science at Mohammed V University in Rabat for accepting to examine my thesis and for his valuable feedback and comments.
I am also grateful to **Pr. Ralf C. STAUDEMEYER**, qualified professor at Schmalkaden University in Germany for his collaboration and for accepting to examine my thesis.

I was very lucky to cross paths with other special persons. I was lucky enough to have a wonderful set of friends and labmates at LRIT. I would like to thank the people that influenced my life choices. Thank you Mohamed for transmitting me your passion and energy for research, for inspiring me with more topics that we are able to follow and for being a great friend. Thank you Yassine, Chaimae, Lamiae, Brahim, Naima, Hicham, Mehdi, Bethaina, Asmae for being such awesome friends. Without your company, it would not have been so much fun.

Finally, and most importantly, I would like to thank my family for the inspiration and unconditional support that has allowed me to become the person that I am. More than anyone else, I am deeply indebted to my loving father and mother for being at my side throughout this challenging journey and for helping me to get to this stage in life as well as my brothers, Nisrine, Alae, Yasser, Youssra, who have always given me their unconditional support. I benefited very much from your support, and love. It is not easy to express my gratitude.

Abstract

In our real world, human interactions and learning are naturally characterised to be multimodal. We use multiple senses and modalities to explore our environment and to confirm with a unified view of our uncertain interpretations through perceived properties from each modality. Similarly to human experience, the increasing availability of multiple sensors gave rise to a diverse and large amount of data. Such growth gained remarkable attention from the machine learning community for finding suitable learning algorithms that allow exploiting those multimodal databases to form a unified picture, disambiguating and increasing the system robustness. Beyond this remarkable progress, society begins to realise that systems designed to assist people in various tasks can also harm individuals and society explicitly or implicitly through unwanted inferences. This thesis develops new approaches based on the fusion of multimodal data provided by Multisensor systems covering two main perspectives. The first is related to the learning problem point of view, where the goal is to study how to fuse multiple types of information to effectively exploit the potential they provide and therefore improve the recognition system performance. While the second relates to the ethical implications of the application by studying the potential risks related to privacy when unprotected data collection through zero permission sensors data is combined with unwanted inferences learned from multiple modalities. Two main applications were studied in this research. Namely, face classification related tasks including gender, ethnicity, and expressions classification using both RGB and depth data provided by the Kinect sensor, and the speech inference problem using motion data provided by zero permission sensors built-in mobile devices. Comprehensive experiments were conducted on the available benchmark datasets with various multi-feature and multimodal settings to show our models effectiveness and to discuss their concerns.

Key Words: Multimodal learning, multi-sensory systems, kinect, smartphones, privacy

Résumé

Dans notre monde réel, les interactions et l'apprentissage humain sont naturellement caractérisés comme étant multimodaux. Nous utilisons de multiples sens et modalités pour explorer notre environnement et pour confirmer avec une vue unifiée nos interprétations incertaines à travers les propriétés perçues de chaque modalité. De même, pour l'expérience humaine, la disponibilité croissante de capteurs multiples a donné lieu à une grande quantité de données diversifiées. Cette croissance a attiré une attention remarquable de la communauté de l'apprentissage automatique pour trouver des algorithmes d'apprentissage appropriés qui permettent d'exploiter ces bases de données multimodales pour former une image unifiée, lever l'ambiguïté et augmenter la robustesse du système. Au-delà de ces progrès remarquables, la société commence à se rendre compte que les systèmes conçus pour aider les gens dans diverses tâches peuvent également nuire aux individus et à la société de façon explicite ou implicite par des inférences non désirées. Cette thèse développe de nouvelles approches basées sur la fusion des données multimodales fournies par les systèmes Multicapteurs couvrant deux perspectives principales. La première est liée au point de vue du problème d'apprentissage, où l'objectif est d'étudier comment fusionner plusieurs types d'informations pour exploiter efficacement le potentiel qu'elles offrent et donc améliorer la performance du système de reconnaissance. Bien que la deuxième porte sur les implications éthiques de l'application en étudiant, les risques potentiels liés à la protection de la vie privée lorsque la collecte de données non protégées à travers des capteurs sans autorisation est combinée à des inférences non désirées tirées de multiples modalités. Deux applications principales ont été étudiées dans le cadre de cette recherche. Il s'agit notamment des tâches liées à la classification des visages, y compris la classification du genre, du groupe ethnique et des expressions à l'aide des données couleur et de la profondeur fournies par le capteur Kinect, et du problème d'inférence vocale à l'aide des données de mouvement fournies par des capteurs sans permission intégrés aux appareils mobiles. Des expériences exhaustives ont été menées sur des ensembles de données de référence disponibles avec divers paramètres multimodaux et multi caractéristiques pour montrer d'abord l'efficacité de nos modèles et ensuite pour discuter de leurs préoccupations.

Mots-clefs: Apprentissage multimodale, systemes multicapteurs, kinect, smartphones, confidentialité

Contents

List of Figures

List of Tables

Chapter 1

Résumé détaillé

1.1 Contexte generale

Cette thèse aborde l'apprentissage multimodal dans le cadre de deux applications. La première application concerne la classification des visages à l'aide de deux types d'informations visuelles fournies par le capteur Kinect, à savoir les images couleur et les données de profondeur. La deuxième aborde le problème de la reconnaissance vocale à l'aide de mesures de mouvement, à savoir les données de l'accéléromètre et les enregistrements du récupérateur d'énergie vibratoire. La présente thèse explore deux perspectives différentes. La première perspective examine les tâches du point de vue de l'apprentissage automatique, et la seconde se concentre sur la l'utilité des applications abordées dans cette recherche. Du point de vue de l'apprentissage automatique, la recherche présentée dans cette thèse se focalise sur l'apprentissage multimodal. Ce dernier peut être défini comme l'apprentissage des représentations pour des problèmes spécifiques en exploitant d'une manière simultanée une ou plusieurs types d'informations afin d'apprendre des informations complémentaires et améliorer la performance de l'apprentissage. Dans le contexte actuel du monde digital numériquement connecté, une grande quantité de données en perpétuelle croissance est intégrée et présentée sous différents formats (images, textes, vidéos...). Cette croissance offre des informations riches et diverses sur un phénomène donné. Par conséquent, elle soulève de plus en plus des questions et gagne l'attention de la communauté de recherche sur l'apprentissage automatique. L'apprentissage automatique multimodal présente plusieurs avantages par rapport à l'apprentissage automatique uni modal car il offre une image unifiée et une vue globale du phénomène étudié. De plus, il peut surmonter les limites des systèmes unimodaux, ce qui conduit à améliorer la prise de décision. Cependant, l'apprentissage des données multimodales est l'un des problèmes les plus compliqués dans le domaine de l'apprentissage automatique, étant donné l'hétérogénéité des données et la difficulté de combiner plusieurs informations sémantiques de haut niveau fournies par diverses sources. Selon Morency et al. [1], les principaux défis de l'apprentissage automatique multimodal sont classés en cinq catégories : représentation, traduction, alignement, fusion et co-apprentissage. Dans notre thèse, nous avons principalement abordé les aspects fusion et représentation.

Du point de vue pratique, les travaux de recherches présentés dans cette thèse exploitent des données multimodales dans deux applications différentes. Le point de départ portait sur le problème de classification de visages en utilisant deux types d'informations visuelles, à savoir les images couleur et données de profondeur. Avec le développement des technologies abordables de la détection de profondeur comme le Microsoft Kinect, l'acquisition d'images de haute qualité contenant des informations de couleur (texture) et de profondeur (forme) devient facile. Cela a attiré de nombreux chercheurs en vision par ordinateur de chercher à exploiter ces informations dans les problèmes de classification et la reconnaissance. Cette partie aborde le problème de la classification des visages dans le contexte des images couleur et de profondeur (données RGB-D). Le but était d'étudier combien les données de profondeur peuvent améliorer la qualité de reconnaissance des systèmes de vision standard, ainsi de comprendre comment nous pouvons exploiter le potentiel que ces données fournissent. Nous avons démontré à travers des problèmes de classification du genre, de la classification du groupe ethnique et des expressions que l'information de la profondeur est pertinente permettent d'améliorer la performance de ces systèmes de classification. Suite à la croissance rapide de la technologie des capteurs et de l'intelligence artificielle, la société a commencé à réaliser que les systèmes conçus pour aider les gens dans leur vie quotidienne peuvent également nuire aux individus et à la société. Un tel préjudice peut se produire dans plusieurs dimensions, allant de l'intrusion dans la vie privée causée par une utilisation abusive des données personnelles par des algorithmes d'apprentissage automatique avancés, à la discrimination causée par des algorithmes entraînés sur des données biaisées ou à des algorithmes non transparents causant un sérieux inconvénient. Ceci nous amène au point de départ de la deuxième partie de cette thèse. Nous avons étudié les problèmes de confidentialité liés à l'utilisation abusive de l'apprentissage automatique lorsque les données personnelles ne sont pas protégées. Nous nous sommes principalement concentrés sur les données collectées à partir de capteurs dites à zéro-autorisation intégrés dans les appareils mobiles, principalement l'accéléromètre et le récupérateur d'énergie vibratoire. Des études récentes révèlent que les capteurs de mouvement mobiles sont sensibles à la voix humaine et sont donc sujets aux attaques par canal latéral. Notre objectif ici est de sensibiliser aux risques potentiels liés à la fuite d'informations sensibles de ces capteurs sur la parole. Par conséquent, nous avons étudié à quel point la confidentialité des utilisateurs peut être compromise par l'extraction d'informations vocales via plusieurs capteurs de mouvement. Nous avons étudié la faisabilité et la complémentarité entre les modalités des capteurs de mouvement dans un scénario d'un appareil à multiples capteurs. Grâce à ce travail, nous avons démontré que l'apprentissage multimodal rend les attaques de confidentialité de la parole basées sur des capteurs de mouvement plus faciles à réaliser en extrayant les propriétés acoustiques utiles des structures de données complexes.

1.2 Objectifs et questions de recherche

Sur la base de la discussion ci-dessus, les deux principaux objectifs de cette recherche sont: Premièrement, améliorer la performance de reconnaissance dans les systèmes multi-capteurs en combinant les informations fournies par plusieurs sources. Deuxièmement, accroître la sensibilisation du public aux implications éthiques de l'utilisation abusive de l'apprentissage automatique multimodal lorsque les données ne sont pas protégées. Répondre à ces deux objectifs nous amène aux questions suivantes:

- Comment combiner efficacement les données acquises à partir de plusieurs sources de l'information afin d'exploiter le potentiel qu'elles offrent?
- Dans quelles tâches d'apprentissage, l'apprentissage multimodal présente-t-il des avantages par rapport à l'apprentissage monomodal?
- Quelles sont les préoccupations sociales / éthiques causées par l'utilisation abusive de données non protégées par des algorithmes d'apprentissage automatique?
- Dans quelle mesure la combinaison de données provenant de différentes sources peut augmenter le problème de violation de la vie privée grâce à de précises inférences non désirées?

1.3 Aspects étudiés

Trois aspects majeurs ont été explorés dans cette thèse. À savoir, les aspects liés au problème d'apprentissage, les défis liés à l'apprentissage multimodal et les aspects dérivés des implications éthiques de l'apprentissage automatique.

1.3.1 Défis liés au problème d'apprentissage

Parmi les différentes tâches d'apprentissage automatique, nous nous sommes concentrés dans cette thèse sur les tâches de classification. Dans un problème d'apprentissage, la qualité et la quantité des données de l'apprentissage sont importantes. Cependant, dans les applications de classification du monde réel, les ensembles de données souffrent de nombreux problèmes, notamment le bruit, une distribution déséquilibrée entre les exemples que contient chaque classe, la grande dimensionnalité et petite taille de la base de données. Les algorithmes d'apprentissage ont besoin d'une quantité suffisante et représentative de données pour pouvoir construire un algorithme capable de faire une généralisation. Sans un ensemble d'entraînement volumineux et informatif, un classificateur peut être induit en erreur ce qui peut amener à un problème de surapprentissage. Ainsi, il est important de bien explorer l'ensemble de données

et de comprendre les objectifs d'apprentissage afin de concevoir une approche d'apprentissage automatique appropriée. Cette thèse se concentre sur la classification et ses principaux défis, où de nombreuses techniques ont été étudiées pour le traitement de ces problèmes.

1.3.2 Défis liés à l'apprentissage multimodal

Les informations du monde réel peuvent être décrites par divers types de données fournies par différents capteurs. Dans cette thèse, nous développons différents algorithmes d'apprentissage pour différents types de données en exploitant leur complémentarité et leur potentiel. Deux problèmes où les données sont fournies par deux capteurs sont étudiés. Dans la première application, une approche de fusion a été proposée pour combiner les données RGB et de profondeur fournies par le capteur Kinect pour résoudre différentes tâches de classification des visages. Cette approche permet d'exploiter efficacement le potentiel des informations rgb-d et en même temps de faire face à la forte variance intra-classe des images de visage. Dans la deuxième application, la représentation multimodale a été apprise à partir des données fournies par deux appareils mobiles intégrés à capteurs de mouvement, à savoir le accéléromètre et le récupérateur d'énergie vibratoire. Le but est d'apprendre une représentation pertinente entre les deux types d'informations à travers les couches des réseaux de neurones profonds.

1.3.3 Implications éthiques de l'apprentissage automatique

La révolution des données, associée aux algorithmes d'apprentissage automatique, est rapidement adoptée dans le monde de l'économie et dans la société. Bien que cela apporte de nombreux avantages, des débats plus larges sont lancés pour sensibiliser le public aux préoccupations potentielles pour la société telles que le manque d'équité algorithmique (conduisant à des décisions discriminatoires), la manipulation potentielle des utilisateurs, la violation de la vie privée et des risques liés à la sécurité et à la cybersécurité. Dans cette thèse, nous avons étudié, en particulier, les risques pour la vie privée liés à l'utilisation abusive de l'apprentissage automatique par des inférences indésirables, principalement lors de l'utilisation de données provenant de plusieurs capteurs. L'objectif de ce travail est de garder les utilisateurs pleinement conscients de la manière de se protéger contre tout type de menace pour la vie privée et de les sensibiliser aux implications éthiques de l'utilisation abusive de l'apprentissage automatique pour encourager l'utilisation responsable des données et de l'Intelligence artificielle.

1.4 Contributions

Amélioration de la performances: analyse des visages à l'aide de données RVB-D

Etude préliminaire: Etude de l'utilité des données de profondeur pour la classification des visages Dans la première partie de cette recherche, l'objectif était d'améliorer les performances de classification en exploitant le potentiel fourni par les images de visage couleurs et de profondeur fournies par le capteur Kinect. Dans un premier temps, nous avons réalisé une étude expérimentale en comparant la performance des images de profondeur par rapport aux images couleur et leur combinaison pour comprendre l'utilité d'utiliser l'information de profondeur et d'évaluer sa pertinence pour trois problèmes de classification. À savoir, la classification du genre, du groupe ethnique et des expressions. Nous avons proposé un schéma de fusion des informations multimodales en sélectionnant le sous-ensemble constitué des caractéristiques pertinentes de chaque modalité a l'aide du classificateur Adaboost. L'avantage d'Adaboost est qu'il sélectionne les caractéristiques individuelles qui peuvent le mieux distinguer les classes. En effet, en faisant une analogie entre entre les classifieurs faibles d'Adaboost et les variables, chaque classificateur faible est associé à une caractéristique de l'ensemble complet. Ainsi, l'ensemble des meilleurs classifieurs faibles construits correspond au meilleur sous-ensemble de caractéristiques permettant une bonne séparabilité. Les résultats obtenus pour les trois tâches de classification montrent que l'utilisation des données de profondeur avait conduit à des résultats compétitifs par rapport à ceux des données RGB. Bien que nous ayons utilisé des descripteurs qui sont généralement utilisés pour les données RGB, elles peuvent toujours extraire des informations pertinentes dans les deux types de données. De plus, la fusion des deux informations basée sur le paradigme Adaboost a amélioré remarquablement les résultats, par rapport à l'utilisation d'une modalité, par rapport aussi la combinaison basée sur une simple concaténation et une autre sur la technique d'analyse en composantes principales.

Reconnaissance faciale multimodale à l'aide de données RGB-D Nous avons étudié le problème de classification de visage à partir des données RGB-D en l'attribuant à un système de classifieurs multiples. L'ensemble de composants de classifieurs correspond à l'ensemble de classifieurs formés sur des parties faciales séparées extraites des visages RGB et Profondeur. L'objectif était d'aborder l'apprentissage à partir des parties du visage locales pour d'abord traiter le problème de grande la variation intra-classe présente dans le visage et, ensuite, de comparer l'efficacité de la représentation locale de la fusion multimodale par rapport à la représentation globale de information. Nous supposons que chaque partie faciale est caractérisée par une forme et des

propriétés de texture spécifiques par rapport à l'autre partie. Ainsi, les modèles formés sur chaque partie du visage sont différents les uns des autres et diversifiés. Après avoir créé l'ensemble, la décision finale est obtenue en utilisant une approche de fusion comme suit. Pour chaque classifieur, un score correspondant est calculé sur la base de sa performance d'apprentissage mesurée par le rapport entre la précision d'apprentissage et le taux d'erreur. Nous calculons la somme des scores des modèles ayant le même label. Le label final attribué à un élément de l'ensemble de test est égale au label correspondant à la classe avec le score maximum. Nous examinons la pertinence de notre approche à travers trois expériences. Nous évaluons d'abord notre approche en comparant les résultats obtenus en utilisant l'ensemble proposé avec les résultats fournis par le visage et par une approche de combinaison à score égale. Deuxièmement, nous évaluons la pertinence du potentiel local par rapport au potentiel global de l'information unimodale et multimodale. Troisièmement, nous testons la diversité des différents composants de l'ensemble.

Problèmes de confidentialité: Extraction de la voix à partir des capteurs de mouvement Dans la deuxième partie de notre thèse, nous nous concentrons sur la mise en évidence des risques potentiels liés aux capteurs de mouvement intégrés dans les appareils mobiles et portables qui fuient des informations privées sur la parole, et les implications éthiques du mal utilisation de l'apprentissage automatique (profond) sur les données non protégées en étant une menace pour la vie privée. Nous présentons une attaque simple dans laquelle les données collectées à partir de l'accéléromètre et des capteurs VEH (Vibration Energy Harvester) peuvent être utilisées pour fuir des informations sur la parole. Nous proposons un modèle basé sur un réseau de neurones Auto-Encodeurs (AE) à plusieurs niveaux qui apprend la représentation temporelle et fréquentielle de chaque capteur. Après avoir extrait les caractéristiques séparément de chaque représentation, temporelle et fréquentielle, nous les combinons dans une représentation jointe temporelle-fréquentielle en utilisant une stratégie d'apprentissage par couche. Dans celui-ci, les caractéristiques apprises dans une couche cachée sont utilisées comme une entrée de l'AE suivant pour produire une nouvelle représentation des données. En représentant les données à travers couches, le modèle peut apprendre et découvrir des variations et corrélations dans les données. Nous démontrons l'efficacité en étudiant trois tâches de classification: l'identification du genre (i), la détection des mots clés (ii) et (iii) la reconnaissance de phrases simples sélectionnées à partir d'un ensemble de données préalablement bien étudié. Nos expériences démontrent l'efficacité de notre modèle et confirment que les capteurs mouvement sont une source riche de données personnelles, d'où émergent des informations hautement sensibles et privées sur utilisateurs des appareils mobiles.

Chapter 2

Introduction

This chapter introduces the research directions of this thesis and points out the investigated issues that will be addressed in subsequent chapters. Section 2.1 introduces the overall picture and context. Section 2.2 describes the research questions raised in this thesis. Section 2.3 presents the main aspects. Section 2.4 summarises the main contributions of this work and outlines the content of each subsequent chapter.

2.1 Context

This thesis addressed multimodal learning in the context of two applications. The first application concerned face classification using two types of visual information provided by the Kinect sensor, namely RGB images and depth data. The second one tackled the problem of speech recognition using motion measurements, namely the accelerometer data and the vibration energy harvester recordings. The current thesis explores two different perspectives. The first perspective looks at the tasks from a machine learning point of view, and the second centres around the practical relevance of the applications addressed in this research.

From a machine learning point of view, the research presented in this thesis delves into multimodal learning. This learning could be described as simultaneously learning task-specific representations using the experience acquired from two or more different types of information to learn complementary information and improve learning performance. With the digitally connected world of today, a rapidly growing amount of data is involved presented in many forms (images, text, videos, to name a few). This increasing growth offers rich and diverse information about a given phenomenon or an activity of interest. Consequently, it raises questions and gained remarkable attention from the machine learning research community. The multimodal machine learning research field has several advantages over unimodality since it offers a more unified picture and global view of the system at hand. Moreover, it can overcome the limitations of unimodal systems, which leads to improving decision making and exploratory research. However, learning from multimodal data has been one of the most

challenging problems in the machine learning field, given the data heterogeneity and difficulty of combining high-level semantic information delivered by various sources. According to Morency et al. [1], the core challenges surrounding multimodal machine learning consists of five categories: representation, translation, alignment, fusion and co-learning. In our thesis, we mainly tackled the fusion and representations aspects.

From an application point of view, the research presented in this thesis exploited multimodal data in two different contexts. The research starting point addressed the problem of face classification using two kinds of visual information, namely the RGB images and depth data. With the development of affordable depth-sensing technology such as the Microsoft Kinect, acquiring high-quality images containing colour (texture) and depth (shape) information becomes easy. This attracted many computer vision researchers seeking to exploit this information in classification and recognition tasks. This section addressed the problem of face classification in the context of RGB images and depth information (RGB-D data). The purpose was to study how much depth data can improve the recognition quality of the standard vision systems and to understand how we can exploit the potential that this data provided. We demonstrated through gender, ethnicity and expressions classification tasks that depth data is relevant information which leads to improved classification performance. Following the growing popularity of sensing technology and artificial intelligence, society began to realise that systems designed to assist people in their daily lives can also harm individuals and society. Such harm may occur across several dimensions, ranging from privacy intrusion caused by personal data misuse by advanced machine learning algorithms, discrimination caused by algorithms trained on biased data, or untransparent algorithms causing a serious downside. This brings us to the starting point of the second part of this current research. We studied the privacy concerns related to the misuse of machine learning when personal data is not protected. We focused primarily on the data collected from zero-permission sensors built into mobile devices, mainly the accelerometer and the vibration energy harvester. Recent studies reveal that mobile motion sensors are sensitive to human speech and are thus prone to side-channel attacks. Our aim here is to raise awareness about the potential risks related to these sensors leaking sensitive information about the speech. Therefore, we studied how much the privacy of smartphones and wearable users can be compromised by the extraction of voice information through multiple motion sensors. We investigated the feasibility and complementarity between the motion sensors modalities in a scenario of a device with multiple sensors. Through this work, we demonstrated that multimodal learning makes speech privacy attacks based on motion sensors easier to achieve by extracting the useful acoustic properties from the complex data structures.

2.2 Goal and research questions

Based on the discussion above, the two main goals of this research are: to improve the recognition performance in multisensory systems by combining information provided by multiple sensors and to increase public awareness of the ethical implications of the misuse of multimodal machine learning when the data is not protected. Addressing these two goals brings us to the following questions:

- When and why additional information might help in a pattern recognition problem?
- How to effectively combine data from multiple sources in order to exploit the potential they provide?
- In which learning tasks does multimodal learning shows advantages over monomodal learning?
- What are the social/ethical concerns caused by the misuse of non-protected data by machine learning algorithms?
- To what extent does combining data from multiple sources can increase the risks related to privacy violation of users?

2.3 Aspects studied during the thesis

In responding to the research questions, three major aspects were explored, principally, learning problems, multimodal learning challenges and the ethical concerns of machine learning.

2.3.1 Learning problem challenges

Among different machine learning tasks, we focused in this thesis on classification tasks. In a classification model, the learning algorithm reveals the underlying relationship between the attribute set and class label and identifies a model that best fits the training data. Then the built model aims at predicting the class labels for any unseen input objects [2]. Therefore, the objective function is learning with good generalisation capability such that the model can make accurate predictions. For the best generalisation, the model should fit the training data properly. Thus, the quality and quantity of training data are critical. However, in real-world classification applications, datasets suffer from many issues, including the noise, the imbalance ratio, the high dimensionality and the small sample size. Learning algorithms need a sufficient and representative amount of data to make generalisations about the distribution of samples [3]. Without a large and informative training set, a classifier may be

misled and may underfit or overfit on the training data. Thus, it is important to explore the dataset well and understand the learning goals in order to design a suitable machine learning approach. This thesis focuses on handling classification problems with noisy, imbalanced, and small datasets, where we explore different machine learning tools to overcome such issues.

2.3.2 Multimodal learning challenges

Real-world information can be described by various data types provided by different sensors. In this thesis, we develop different learning algorithms for different data types by exploiting their complementarity and their potential. Two problems where data is provided from two sensors are investigated. In the first application, a fusion approach was proposed to combine RGB and depth data provided by the Kinect sensor to solve different face classification tasks. This approach allows to exploit the potential of the RGB-D information effectively and at the same time to cope with the high intra-class variance of the face images. In the second application, the multimodal representation was learned from data provided by two motion sensors built-in mobile devices—namely, the accelerometer and the vibration energy harvester. The aim is to learn a useful representation by discovering through layers the high-level correlations across the two representations.

2.3.3 Ethical implications of (multimodal) machine learning

The data revolution, coupled with machine learning algorithms, is rapidly being adopted across the economy and society. Although it has many benefits, debates are raised to increase the public awareness of the potential concerns to the society such as a lack of algorithmic fairness (leading to discriminatory decisions), potential manipulation of users, privacy violation, and related safety and cybersecurity risks. Hence, we have studied, in particular the privacy risks raised from the misuse of machine learning through unwanted inferences, mainly when using data from multiple sensors. As an application, we studied the privacy risks related to data provided by motion sensors, built-in mobile and wearable devices. The goal of this work is to keep the users fully aware of how to keep themselves protected against any types of privacy threat and to encourage the responsible use of data and AI.

2.4 Organization of the manuscript and contributions

The remainder of this thesis is organised in two main parts. Before getting into the details of the proposed approaches, background information about the main notions used in this research is explored in chapter 3. Part I focused on face classification tasks using RGB-D images provided by the Kinect Sensor. Part II tackled the problem of learning acoustic information from mobile motion sensors data, principally, the vibration energy harvester sensor and the accelerometer. For each part, we started by first presenting the application background. Then, we described our approaches and used systematic evaluations. Finally, chapter 9 summarises the main findings and conclusions across all parts and chapters. In the following section, we present the outline of the thesis and an overview of each contribution as well.

2.4.1 Enhancing performance: Face analysis using RGB-D data

Preleminary study: On the usefulness of depth data in face classification In the first part of this research, the goal was to improve the classification performance by exploiting the potential provided by RGB and Depth face images provided by the Kinect sensor. As a preliminary step, we carried out a comprehensive experimental study comparing the performance of the depth images versus the RGB images and their combination to understand the usefulness of using depth information and to evaluate its relevance for three classification problems, primarily, gender, ethnicity and expressions classification tasks. We proposed a fusion scheme of the multimodal data by selecting the subset consisting of the relevant features from each modality based on the Adaboost classifier. The advantage of Adaboost is that it selects the individual features that can best discriminate among classes. In fact, by making an analogy between weak classifiers and features, each weak classifier is associated with one feature from the complete set. Thus, the set of best weak classifiers identified corresponded to the best subset of features that leads to good separability. The results obtained for the three classification tasks showed that the use of the depth data had led to competitive results compared to those of RGB data. Although we used state-of-the-art handcrafted features that are usually used for RGB data, they still able to extract relevant information in both types of data. Also, the fusion of the two information based on the Adaboost paradigm improved remarkably the results, compared to the use of one modality, the combination based on a simple concatenation and based on principal component analysis technique.

Multimodal face recognition using RGB-D data We studied the RGB-D face classification problem by attributing it to a multiple classifiers system. The set of classifiers components corresponded to the set of classifiers trained on separated extracted facial parts from the RGB and Depth faces. The goal was to address learning from local facial parts to first deal with the intra class-variation present in the RGB and Depth faces and second to compare the efficiency of the multimodal fusion from local versus the global representation of the unimodal information. We assumed that each RGB and depth facial part is characterised by a specific shape and texture properties from the other part. Thus the models trained on each facial part is diverse from the other. After creating the ensemble, the final decision is obtained using a scored fusion approach. For each component classifier, a corresponding score is computed based on its training performance measured by the ratio between the training accuracy and loss. We computed the sum of scores of the models with the same label output. The final label assigned to a given input pattern from the testing set is equal to the label of the class with the maximum scores. We examine the pertinence of our approach through three experiments. We first evaluate our approach by comparing the obtained results using the proposed ensemble with the results provided by the whole face and by an equal score combination approach. Second, we evaluate the relevance of the local versus the global potential of the unimodal and the multimodal information. Third, we tested the diversity of the different components in the ensemble.

2.4.2 Privacy concerns: Extracting speech from motion sensors

In the second part of the research, we focused on highlighting the potential risks related to motion sensors built-in mobile and wearable devices are leaking private information about speech, and the ethical implications of the misuse of advances in (deep) machine learning on non-protected data as a threat to privacy. We showcase a simple attack in which collected data from accelerometer and Vibration Energy Harvester (VEH) sensors can be used to eavesdrop on speech. We then proposed a multilevel stacked auto-encoder model that learns the time and frequency representation from each sensor. After extracting the features separately from each source, that is time and frequency, and we combine them into a joined time-frequency representation using layer-wise training strategy. Therein, the features learned in a hidden layer are used as input to the next AE to produce a new representation of the data. By representing the data through layers, we enable the learning of complex patterns across data variations. This joint representation leads to a shallow model, thereby making it difficult for a single hidden layer model to find correlations between representations that have been joined. We demonstrate the efficiency of our model with poor quality data and a very low sampling rate. We investigated three classification tasks:

gender identification (i), hotwords detection (ii), and (iii) recognition of simple phrases selected from a previously well-investigated dataset. Our experiments demonstrate the efficiency of our model and confirm that motion-sensitive sensors are a rich source of personal data, from which highly sensitive and private information about people close to the sensor emerges.

2.4.3 List of Publications

1. Safaa Azzakhnini, Lahoucine Ballihi, and Driss Aboutajdine, **A learned feature descriptor for efficient gender recognition using an RGB-D sensor**, IEEE International Symposium on Signal, Image, Video and Communications (ISIVC), (2016).

2. Safaa Azzakhnini, Lahoucine Ballihi, and Driss Aboutajdine, **Learning discriminative features from RGB-D images for gender and ethnicity identification**, Journal of Electronic Imaging, (2016).

3. Safaa Azzakhnini, Lahoucine Ballihi, and Driss Aboutajdine, **Machine perception in gender recognition using RGB-D sensors**, 13th IEEE/ACS International Conference of Computer Systems and Applications (AICCSA), (2016).

4. Safaa Azzakhnini, Lahoucine Ballihi, and Driss Aboutajdine, **Combining Facial Parts For Learning Gender, Ethnicity, and Emotional State Based on RGB-D Information**, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), (2018).

5. Safaa Azzakhnini, Ralf C. Staudemeyer **Extracting speech from motion-sensitive sensors**, 15th DPM International Workshop on Data Privacy Management, (2020)

Chapter 3

General background

In this chapter, we provide the reader with the background knowledge necessary to follow the explanations and descriptions in this thesis. We outline the related work in three sections. We first present the main machine learning definitions and challenges that have been explored in this thesis. Second, we discuss the main challenges of multimodal machine learning provided in the literature. We then provide a background on the commonly discussed concerns of machine learning and data by the research community.

3.1 The pattern recognition problem

This section introduces the main concepts of machine learning used in the thesis. It also covers the main challenges in a pattern recognition problem that have been investigated.

3.1.1 The learning problem

Machine learning is the field of research that aims at studying learning systems. According to [4], a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. In other words, a machine learning algorithm is an algorithm that is able to learn from data to perform a specific task. Then, a quantitative measure of its performance must be designed for evaluation [5].

Learning algorithms can be divided into two main types, supervised and unsupervised. Supervised learning experience a dataset containing a set of observations, but each example is associated with a label or a target. Given a dataset $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, where $x \in \mathbb{X}$ is the set of observation and $y \in \mathbb{Y}$ the corresponding target. The objective is to estimate a function $f : \mathbb{X} \to \mathbb{Y}$ that maps inputs x to target values y, given a set of n training examples D. Such type usually benefits from the construction of a model that predicts targets from a set of input values. The differences in algorithms for prediction problems typically arise from the properties of the inputs

(observations) and the properties of the targets. If the target values are categorical, the task is called classification. Here the algorithm aims at constructing a function that predicts discrete class labels called a classifier. While learning with continuous target values is referred to as regression. The second type is the unsupervised learning, where the dataset $D = \{(x_1, x_2, \cdots, x_n)\}$ do not obtain a supervised target outputs. The objective is of finding patterns in the data. It can usually be interpreted as density estimation where the objective is to find a good model of the data distribution [6]. Applications of unsupervised learning are dimensionality reduction, clustering and feature learning. Roughly speaking, the term supervised learning involves observing several examples with a corresponding value or label, so an associated value is provided by an instructor who shows the machine learning system what to do. While unsupervised learning involves observing several examples and attempting to learn some interesting properties of that distribution, so there is no instructor and the algorithm must learn to make sense of the data without this guide.

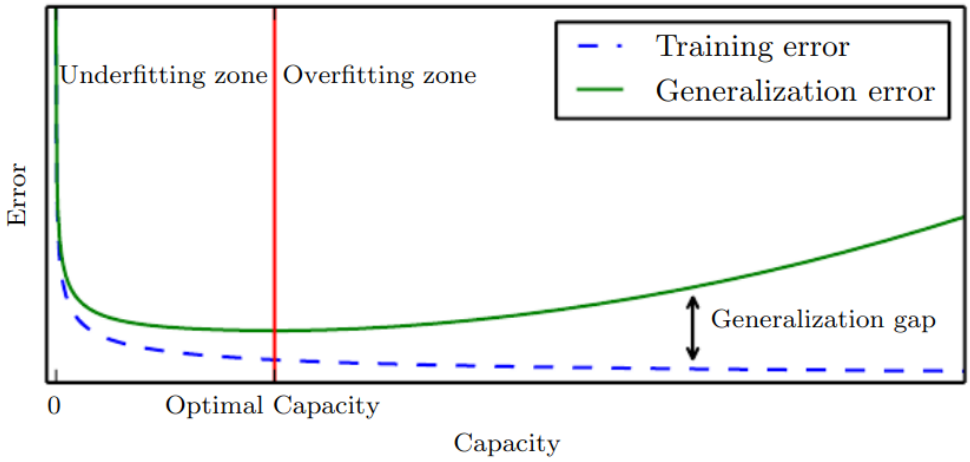


FIGURE 3.1: Typical relationship between capacity and error [7].

The main challenge in machine learning is that the learning algorithm must generalise well from the training data to any data from the problem domain. A good generalisation requires that the learning algorithm must (i) making the training error small, and (ii) making the gap between training and test error small. There is a terminology used in machine learning about how well a machine learning model learns and generalises, namely overfitting and underfitting. Overfitting and underfitting are the two central challenges in machine learning. Underfitting occurs when the model is not suitable for the task, so it is not able to obtain a low error value on the training set and to capture the underlying pattern of the data [5]. However, overfitting occurs when the gap between the training error and test error is too large. This happens when the model fits

the training data too well, so the model is viewed as a memoriser rather than a learner, since it lacks the aspect of generalisation, namely, of using observed data to predict the labels of unseen examples [7]. Controlling these two problems depend mainly on both the capacity of the model defined as its ability to fit a wide variety of functions and to the sample size. Models with small capacity are unable to solve complex tasks, and with high capacity can solve complex tasks, but when their capacity is higher than needed, they may overfit and memorise. Therefore, a machine learning algorithm performs best when its capacity and the used sample size are appropriate with the complexity of the task (as illustrated in figure 3.1).

Pattern recognition can be defined as the development of learning systems that can learn and identify the regularities in the data, going from a low level of collected data from sensors to a higher level of decision making. Learning patterns from real-world data is a challenges problem [2]. The challenges come from either the collection of representative data that can help to build a good learning model or from the suitable machine learning algorithm that can generalise well on the data. In this thesis, we explored a number of issues related to pattern recognition while studying several classification tasks that are related to the used dataset described in section 3.1.2.

3.1.2 Major issues in pattern recognition

In this section, we discuss some of the issues in pattern classification that are explored in this thesis—namely, issues related to learning from noisy, imbalanced and small datasets.

Noise and outliers Noise in the data is defined as any property of the sensed pattern due not to the true underlying model but to randomness in the world, or the sensors [2]. It occurs when the data has been corrupted by various errors such as systematic uncertainty, measurement error, human error, etc. Various types of Noise can be present, including randomness in measurements, outliers and missing data. Learning from noisy data is difficult due to many reasons. In fact, a learner may not be able to distinguish between representative cases, and noise-induced ones [2, 6]. It may also hinder extracting useful properties from the data and thus increase the complexity of the task and making the classifier less effective. To deal with noisy data, data cleaning, preprocessing and feature selection techniques are important steps in a pattern recognition system to identify and overcome outliers [2].

Imbalanced datasets Another major issue in pattern recognition is dealing with imbalanced datasets, which is considered as a crucial problem in machine learning. This problem occurs when having many more instances of certain classes than others, in which one class is represented by a large set of samples,

while the other one is represented by only a few. The degree of imbalance is represented by the ratio of the size of the sample size of the one class versus another. There are a large number of real-world applications that give rise to data sets with an imbalance between the classes such as in the medical diagnosis, image classification, biological data analysis, text classification, and fraud detection, among many others. The imbalance between the classes is a major challenge when learning from imbalanced datasets. Basically, when training a classifier, the goal is to maximise the accuracy of its predictions. When the classes are imbalanced, the classifier is affected by the majority class and tend to ignore the minority class or treat the minority samples as Noise.

Small datasets and overfitting Small data challenges have emerged in many learning problems. In pattern recognition, the collection of large size of examples is very expensive or not always possible. The issue when using a small sample size is dealing with overfitting where the dataset is too small for the chosen complexity of the model [7]. Thus techniques helping to reduce the model complexity such as unsupervised learning, feature learning and data compression, feature selection, and using of a meta-classifier are common approaches used to deal with the small sample size, and to improve the generalisation performance of the learner [5, 6].

Separability and intra-class variance In classification problems, the intra-class variance is defined as being the variance within the objects of the same class, while interclass variance is the variance between different classes. Thus a good discriminative model should be able to minimise the intraclass variance and maximise interclass variance. In many classification problems, a single class is composed of various sub-clusters where samples of a class are collected from different sub-clusters. The presence of within-class sub-clusters increases the intra-class variance and therefore increases learning concept complexity of the data set [2]. Various approaches have been proposed in the literature to mitigate the negative impact of the intra-class variance such as addressing the problem based on local features, or proposing metric learning, which aims to maximise inter-class similarity and meanwhile minimise intra-class distance [5, 6].

3.2 Multimodal machine learning

In our real-world, human interactions and learning are naturally characterised to be multimodal. We use multiple senses - hearing, touching, smelling, seeing - to explore our environment and to confirm with a unified view and to confirm our uncertain interpretations through perceived properties from each modality. In contrast to human experience, computer systems learned using information that come from different sources and through different channels. With the increasing availability of multiple sensors and daily user interactions with the internet, a

different type of information about a given task become easily affordable. Such growth gained remarkable attention from the machine learning a community to finding suitable learning algorithms that allow exploiting those multimodal databases to form a unified picture, disambiguating and increasing the system robustness [8].

Learning from diverse modalities aims at building models able to learn meaningful structures from multiple sources to gain an in-depth understanding of natural phenomena, by exploiting the complementary information and eliminating the redundancy. Multimodal machine learning has the potential to learn more generalisable representations and to improve the learning performance using the information acquired from different sources. For instance, the visual properties of a face image and the sound properties of the individual give jointly a more abstract characterisation of the given user emotional state. Similarly, combining text and visual cues lead to a certain and clear interpretation of a given image. However, learning from modalities brings many challenges for machine learning researchers [8–12]. The challenges arise from the complexity of the data to be fused and the diversity of the sensor technologies, which leads to the heterogeneity between the different information. Furthermore, Because of sensors acquisition imperfection and the application environment, the data provided by sensors is always affected by some level of uncertainty and Noise in the measurements. Thus combining the data needs to tackle the uncertainty issue and to exploit the multimodal information effectively by reducing the redundancy efficiently.

3.2.1 Multimodal Fusion

Multimodal fusion is one of the original topics in multimodal machine learning. It consists of integrating the information provided from multiple sources with the the goal of predicting an outcome to improve the learning performance and to capture complementary information. The fusion of different modalities is generally performed at two levels: feature level or early fusion and decision level or late fusion [13, 14]. In the feature level, the features extracted from each input data are first combined (often by concatenating their representations) and then fed to a single model. The goal here is to exploit the correlation between low-level features of each modality. While late fusion performs integration after each of the modalities has made a decision. It consists of the fusion of unimodal the decision values using a fusion mechanism such as voting[15]. Both the two approaches have been used in many applications including audio-visual speech recognition (AVSR) [14], multimodal emotion recognition [16], medical image analysis [17], multimedia event detection [18], etc.

For the fusion based on the feature level, feature vectors are generated separately for each modality using different feature extraction algorithms then fused to produce a single representation. A common way is to simply concatenate

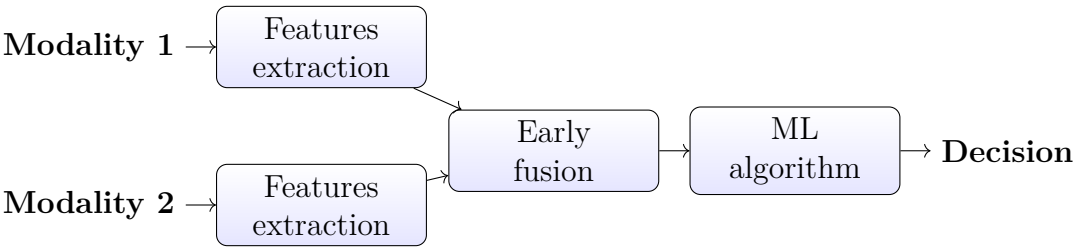


FIGURE 3.2: Early fusion

them or calculate the weighted average of the individual feature vectors[19–22]. An advantage of this approach is that it requires only one learning phase and can exploit at an early stage the correlation between multiple features provided by different types of information [23]. However, this approach has limitations to learn the correlation among heterogenous features when the number of modalities is large.

In contrast, late fusion uses unimodal decision values and fuses them using a fusion mechanism such as averaging [24], voting schemes [15, 25], a linear weighted sum or product [14, 26, 27], or a learned model [28, 29]. For the late fusion approach, the decision level fusion strategy offers more flexibility in terms of the modalities used in the fusion process when the number is large of when one of the modalities is missing [30]. This type of approach allows the use of the most suitable technique to learn from each modality. However, only a limited amount of information is available at this level where the correlations between the different multimodal variables the learning cannot be considered, also the learning process can be time-consuming [31].

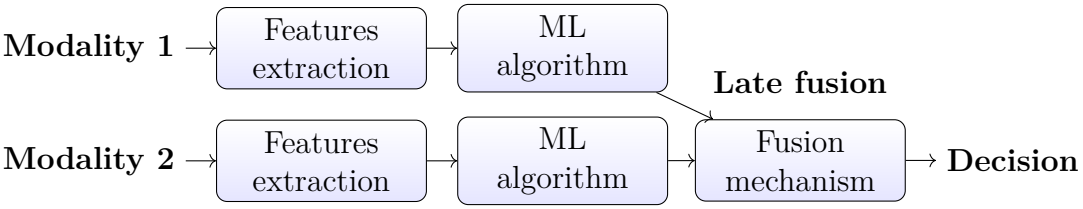


FIGURE 3.3: Late fusion

3.2.2 Multimodal representation learning

Learning representation from multiple modalities called also features learning aims at automatically discovering the good representations from multiple sources in a meaningful way that exploits the complementarity and redundancy between them. Representation learning has become a field in itself in the machine learning community after the remarkable success of deep neural network-based models to automatically discover abstract and useful correlations between variables without relying on human prior knowledge. However, learning the

good representation from many types of information still a challenging task due to the data heterogeneity, uncertainty and the different levels of Noise of each modality. According to Bengio et al.[32], a good representation should have properties such as smoothness, dealing with multiple priors about the world around and capturing temporal and spatial coherence.

There are two categories in multimodal representation : joint and coordinated [8]. The first type aims at learning the multimodal representation X_j by projecting the unimodal representations $(x_1, x_2, \ldots, x_n)$ into one space using a neural network model or a graphical model such as restricted boltzman networks as the following: $X_j = f(x_1, x_2, \ldots, x_n)$. A simple way of a joint representation is a concatenation of individual modality features. While in a coordinated representation separate representations, each unimodal representation x_1, x_2 is learned separately with two separate models $(f(x_1), g(x_2))$, but with enforcing certain similarity constraints. Then the resulting space is coordinated between them as follows $f(x_1) \sim g(x_2)$

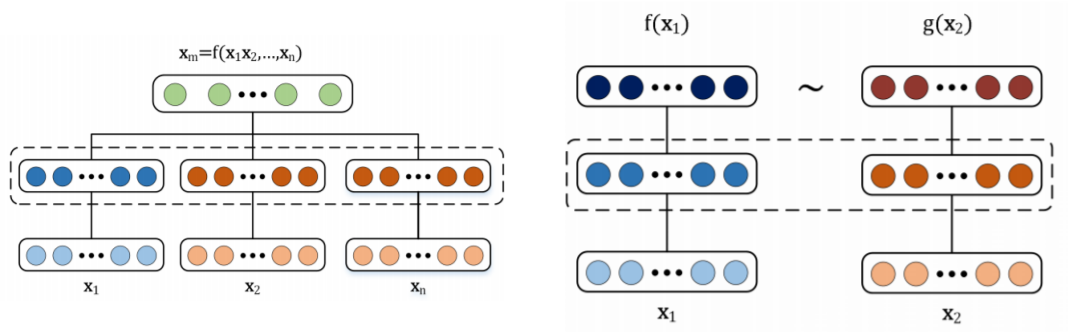


FIGURE 3.4: Joint representation (left) and coordinated representation (right)

Joint representations are learned using neural networks which have become a very popular method for unimodal/multimodal representation in various applications involving computer vision, natural language processing, and speech recognition such as video classification, event detection, sentiment analysis, and visual question answering [33–38]. Representing data using the neural network can be performed using a supervised way (trained to perform a specific task), requiring a lot of labelled training data, or unsupervised way (without integrating the supervised information). To construct a multimodal representation using neural networks, each modality starts with several individual neural layers followed by a hidden layer that projects the modalities into a joint space. Then joint multimodal representation could be passed through other multiple hidden layers used directly for prediction. By learning through layers, the heterogeneity gap of different modalities is minimised, and a more abstract representation and correlations can be discovered between the variables. Several works have been proposed in this context. In [39], the authors extended the idea of using autoencoders to the multimodal domain. Each unimodal representation was

learned separately using a denoising encoder and then fused in multimodal representation using another autoencoder layer. The features are learned in the hidden layer of the autoencoder by minimising the reconstruction loss. It is also common to fine-tune the obtained representation on a specific supervised task [40]. To exploit the intra-modality dynamics in sentiment analysis, the authors in [41] proposed to fuse language, video, and audio modalities in a tensor, which is constructed from the out product of all the modality-specific feature vectors. Probabilistic models are also used to construct multimodal joined representations. The common models are deep Boltzmann machines. A multimodal deep belief network was introduced by [42]. It was applied in many applications such as audiovisual emotion recognition in [43], gesture recognition [44] and human pose estimation [45].

Another alternative to joint representations is a coordinated representation, where the multimodal subspace is coordinated by learning separated but constrained representations for each modality. This approach has used the similarity to enforce between representations through the applied constraint, moving on to coordinated representations. Examples of constraints include minimizing cosine distance [46], maximizing correlation [47], and enforcing a partial order [48]. This way allow persevering the exclusive and useful characteristics specific for each modality [49]. One of the earliest works proposed a mapping function between images and their annotations by maximising the inner product between the image and textual features [50, 51]. A model proposed in [52] enforces a dissimilarity metric and implements the notion of partial order in the multimodal space. The goal is to capture a partial order of the language and image representations. Coordinated models constructed based on maximising the canonical correlation analysis have been used [53–57]. Deep canonical correlation analysis (DCCA) [58] was introduced as an alternative which leads to better-correlated representation space.

3.3 Ethical and societal implications of ML algorithms and data

The implementation of AI systems, particularly machine learning algorithms, is increasingly being used in the products and services that shape our daily life. Individuals and society trust smart algorithms not only to perform tasks like accounting and automatic manufacturing but also to make decisions on their behalf. Meanwhile, data, in the form of observations, permeate modern society. Every field has data. We use data to discover new knowledge, to make decisions, and to predict the future.

The growing advancement of machine learning algorithms coupled with the large availability of data is causing an explosive interest in machine learning and its applicability to all fields. The presence of AI-based system is expanding

rapidly, without adequate governance oversight, or accountability regimes. Although it enables the development of many tools with the potential of bringing good to the society, it raises at the same time many concerns and risks that pose pressing ethical impacts on society [59, 60]. Only from 2017, a range of unexpected adverse consequences has affected society at many different levels, where the lack of transparency and accountability creates many privacy issues, risk of bias, error, accountability questions and lack of transparency. Figure 3.9 present the most revealed scandals in 2018, which was a dramatic year in AI. A series of data breaches and privacy violation were reported, such as the Cambridge Analytica scandal seeking to manipulate national elections in the US and UK using social media data. On the ethical level, the project Maven revealed by Google, which aims at building AI systems for the Department of Defense's drone surveillance program. In the safety and human rights level, it was reported that a voice recognition system in the UK designed to detect immigration fraud ended up cancelling thousands of visas and deporting people in error, and IBM Watson recommended unsafe and incorrect cancer treatment. Concerns are also increasing discrimination, such as the racial discrimination reported in the criminal justice system by using Amazon's Rekognition system [61–66].

Research on the ethical and social impacts of AI have become topics of interest to both machine learning and social science communities. It gained a remarkable interest as a response to many unexpected consequences on society and societal harms that the misuse of data, poor design, or unintended consequences of AI systems may cause. The two communities started the FAT/ML organisation [67], which since 2014 has held excellent technical workshops annually on Fairness, Accountability, and Transparency in Machine Learning [65, 66, 68–78]. In this section, we provide an overview of the most discussed ethical consequences in the literature related to discrimination and bias, privacy and manipulation concerns, human-robot interaction and the effects of autonomy [60, 65, 66, 77, 78]. The goal here is to first, increase consciousness about the risks related to misuse of ML algorithms and data. Second to encourage the responsible use of data and AI.

To understand how these concerns are caused, let's first recall a simple machine learning pipeline. It consists of building a mathematical model based on sample data, known as "training data" by minimising or maximising an objective function. Then, test data is used to validate the model based on an evaluation metric. With such a machine model, the idea is that when new, not seen before, data come along, they are fed into the model, and on this new data, the model can make predictions or decisions without being explicitly programmed to do so. Thus, data and machine learning fit together, and unethical use of AI could be caused by misuse of data and/or the learning algorithm (Figure 3.5).

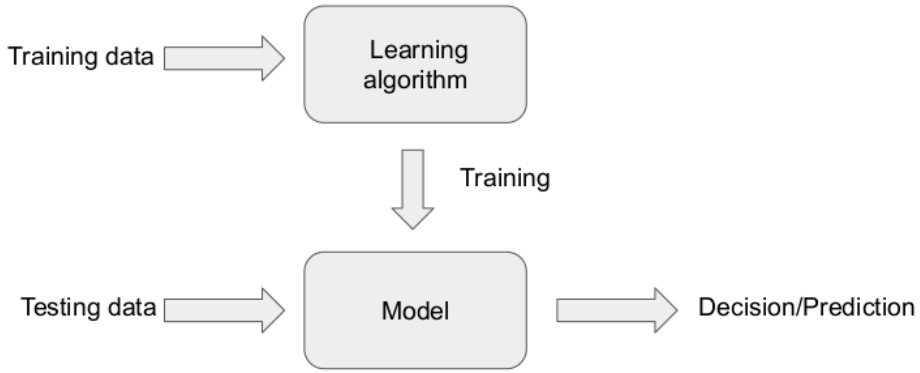


FIGURE 3.5: A simple machine learning system

3.3.1 Bias and disrimination

Bias is defined as a prejudice for or against one person or group (gender, ethnic, religious, ideology or underrepresented groups), especially in a way considered to be unfair. A bias in automated decisions means that the model we build is used to make unbiased decision or predictions. Unfortunately, there is a growing consensus that the designed algorithmic bias, increase and amplify the biases already present by nature and culture, leading to even more discrimination which can be dangerous mainly in sensitive applications such as medical, political and even justice domains[59, 60]

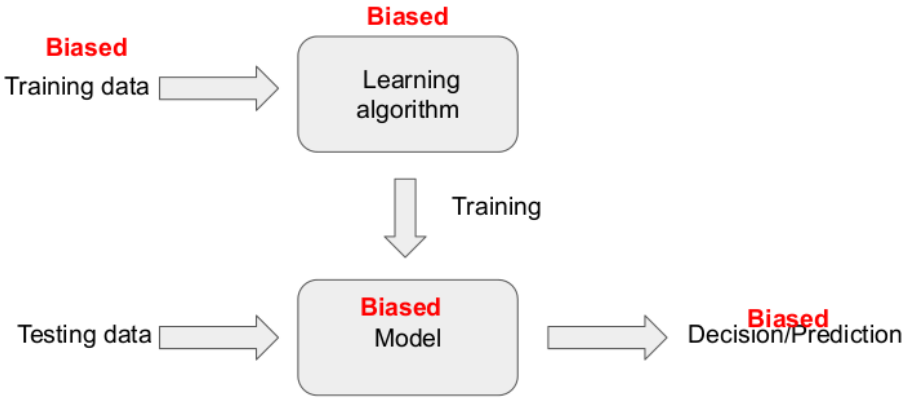


FIGURE 3.6: A simple machine learning system with bias

Potential sources of unfairness and discriminatory in machine learning outcomes arise at different points in the ML pipeline. It could arise from biases in the data or the black-box models. Concerning the first type, there are multiple ways that discriminatory bias can seep into data. For instance, using unbalanced data can create biases underrepresented groups (historical bias). It

can also obtain through the unbiased representation of certain properties in the dataset by choosing and measuring the particular features and labels of interest that be relevant to the outcome are chosen (representation bias). Referring to the above diagram 3.7, When we feed biased data into the algorithm, then we are training with biased data, and therefore, we can build a biased model that produce a biased outcome [79, 80] Bias may also come from the algorithm, reflecting political, sexual, religious, or other kinds of preferences as a result of intentional or unintentional decisions by its designers. Algorithms could be designed to take advantage of seemingly innocuous factors that can be discriminatory in tuning the hyperparameters and setting metrics at the modelling, testing, and evaluation stages that involve human choices and may have discriminatory effects in the trained moded. For example, the accuracy as an evaluation metric represents is an evaluation over the average, which may systematically discriminate against a specific minority. Finally, the outcome of a model could be biased, again reflecting the inherent bias present somewhere upstream from the final result [81, 82].

3.3.2 Privacy

Our society is increasingly relying on algorithms in all aspects of its operation. Most of these algorithms use personal data to learn from them and provide decisions. Data is connected to a single Internet, and there is more and more sensor technology in use that generates more and more data about many aspects of our lives. Thus, large and diverse amount of data are collected, enabling richer analysis of our behaviour, preferences, interests and possibly disclosing information that we prefer to keep private. The ability of machine learning algorithms to analyse complex correlations and patterns in data makes it difficult to know how the data will be used and for what purpose, and therefore the misuse or disclosure of such data can reveal serious concerns of privacy and respect of human rights [60, 83, 84].

Threats to privacy are posed by AI systems as a result of the bad design or the misuse of the machine learning model. Machine learning allows us to extract information from data and discover new patterns and can turn them into sensitive and private information. This new information could not directly be observed but is deducible from what has already been shared or inferred. Moreover, even if this information is not directly present in one source, it can be inferred by learning a combined analysis from multiple sources. Many threat of privacy has been revealed in the past years that go beyond a simple collection of data. They include the use of information to manipulate behaviour in a way that affects our personal choices and life decisions. For example, the Facebook-Cambridge Analytica "scandal" showed that AI could be used to target and manipulate individual voters by using the personal data of more than 50 million Facebook users. Sensitive information can be obtained by recovering

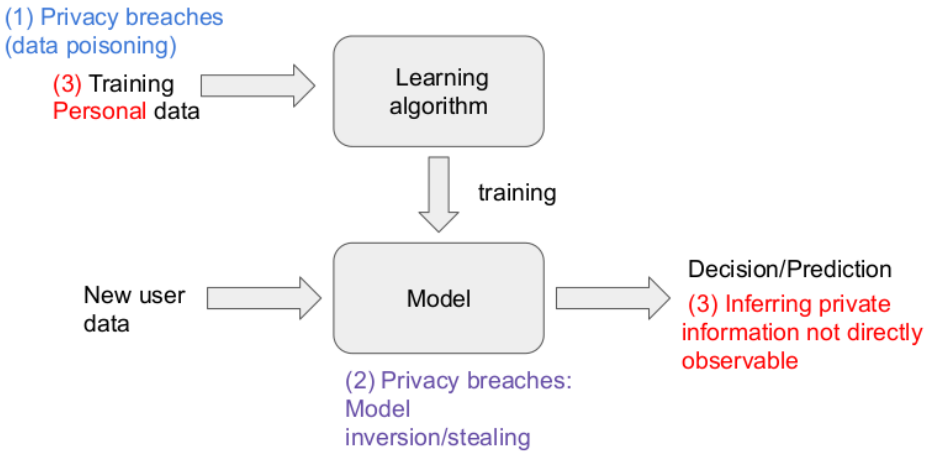


FIGURE 3.7: Privacy issues in a machine learning pipeline

the built model, or the information about data used during training [61, 62]. These attacks are a major concern as AI models represent valuable intellectual property trained on potentially sensitive data, including financial trades, medical records, or user transactions. Another privacy issue can be caused by attacks on training data called the poisoning attack. Here, the attackers able to inject bad data into your model's training pool, and hence can manipulate and change the algorithm and the way it learns [85–90].

3.3.3 Autonomy and ethics

Autonomous systems are systems that operate without human intervention, based on the interaction with their environment. New technological developments in autonomy, AI and robotics have broad applications across society. Examples of such applications include self-driving cars, drones, robots, autonomous weapons systems and software agents, such as bots, etc. These systems give rise to a range of important opportunities. However, a core concern with autonomous systems is the loss of human control and responsibility for ethical and life-and-death decisions. Autonomous systems are already revealed consequences for human safety, and well-being [91–93]. In June 2005, a surgical robot at a hospital in Philadelphia malfunctioned during prostate surgery, injuring the patient. In June 2015, a worker at a Volkswagen plant in Germany was crushed to death by a robot on the production line. In June 2016, a Tesla car operating in autopilot mode collided with a large truck, killing the car's passenger.

Recently much attention has been given to the ethical dilemmas to autonomous systems when needing to deal with life-threatening decisions. As an example, self-driving cars or vehicles gave rise to many opportunities in reducing the significant damage that human driving causes. However, there seem to

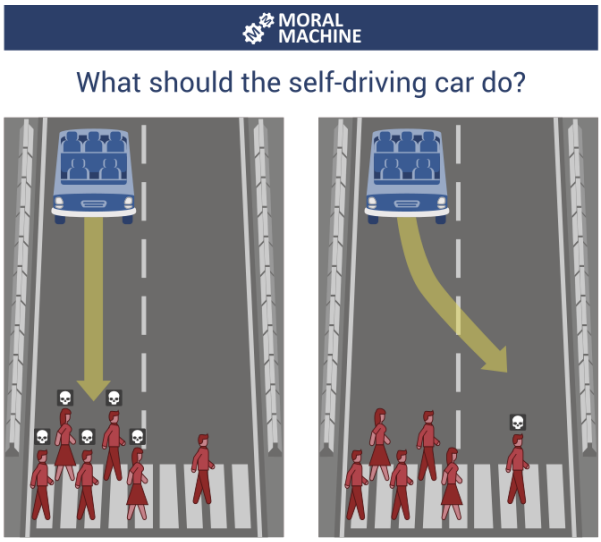


FIGURE 3.8: The “Trolley Problem for self-driving cars" [94]

be questioned on how autonomous vehicles should behave in ethical problems in driving [95, 96], not only in issues such as speeding or not keeping a safe distance but on how responsibility should be distributed in the complicated system the vehicles operate in to guarantee the public safety. There is some discussion of the canonical "trolley Car problem" in this context which raises the ethical question of whether it is better to kill one person or five [94, 97]. This problem is used as a theoretical tool to investigate ethical intuitions, mainly the difference between intended vs tolerated consequences. In the original formulation, a human has to decide. With the advent of AI-based systems such as self-driving cars, we can no longer ignore these unanswerable questions because the machine is going to have to make a decision: Does the car stay the course and kill the people standing or swerve to the right on the corner and run over the pedestrian in the crosswalk? (Figure 3.8). The loss of human control is a potential concern in the example of military use of autonomous weapons systems [98–100] . In such systems, the autonomous use of force may also give rise to significant risks of unintended, and potentially unlawful harms and consequences. Ethical issues are raised with robotics as well, such as using chatbots and robots in healthcare [93, 101–103]. In such a case, if the chatbot is giving medical advice or recommendation, how much should humans trust and what kind of ethical responsibility do we have to remind users when they are interacting with a chatbot?

Finally, AI technology has serious several social and ethical impacts if not used responsibly. As a response to these concerns, new privacy regulations have been established, such as The EU General Data Protection Regulation (GDPR) (GDPR 2018) [104]. IBM started an initiative on data transparency with Columbia University (University 2018) to enable the development of trusted

and fair machine learning systems. Besides, the Partnership on AI, an ethics-focused industry group was launched by Google, Facebook, Amazon, IBM and Microsoft by adopting principles that have been defined by the FAT/ML organisation [67] which since 2014 has held excellent technical workshops and maintains a list of scholarly papers summarised by the acronym FATE "F" for Fairness **Fairness**, which means that the models we build are used to make unbiased decisions or predictions, thus, we then need ways to detect bias and ways to remediate detected bias. "E" for **Accountability** means to determine and assign responsibility to someone for a judgment made by a machine. "T" for **Transparency**, by being open and clear to the end-user about how an outcome, e.g., a classification, a decision, or a prediction, is made. "E" for **Ethics** by paying attention to both the ethical and privacy-preserving collection and use of data as well as the ethical decisions that the automated systems we build will make [59, 65, 66, 68–78, 105].

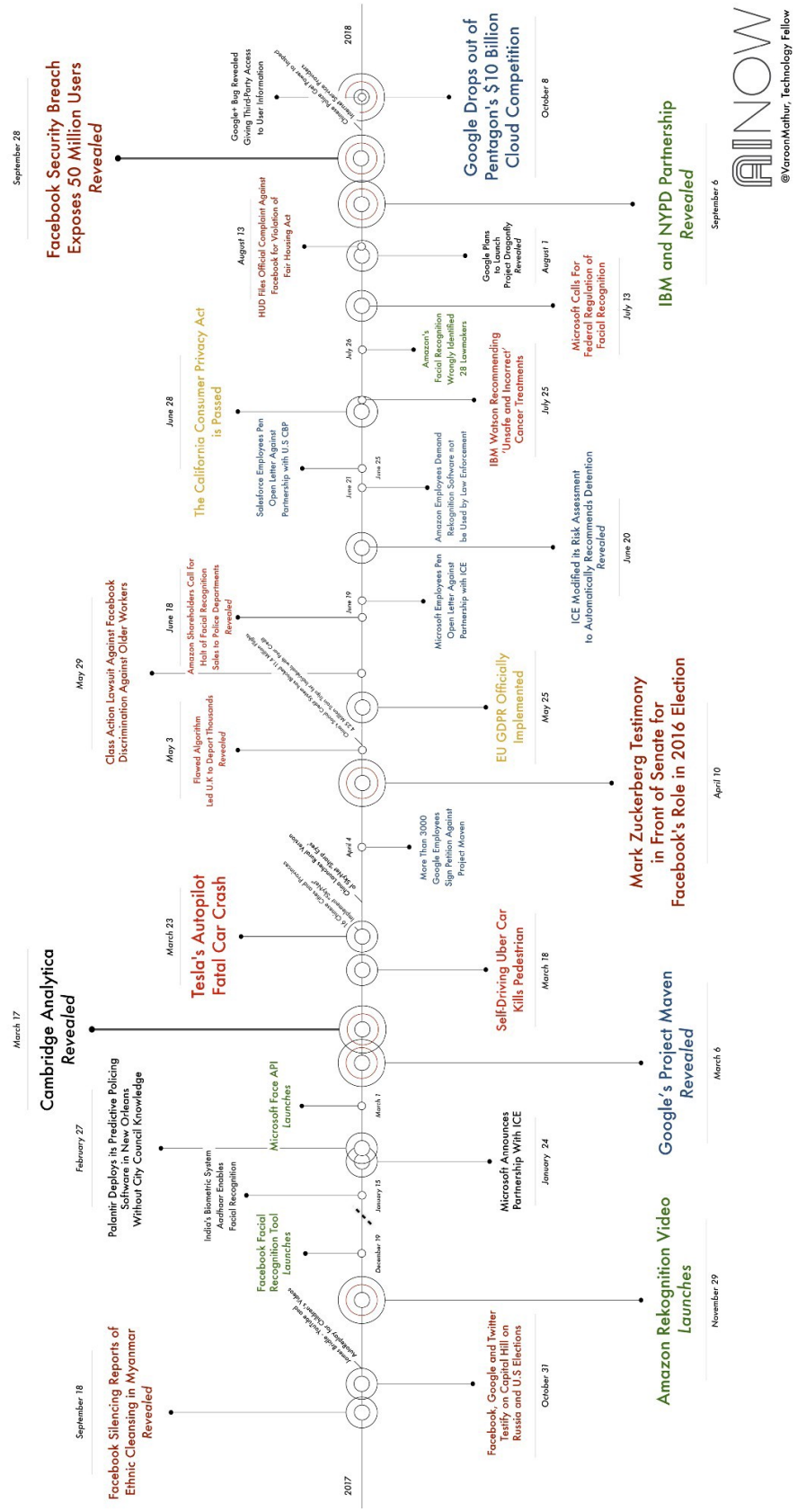


FIGURE 3.9: Timeline of news events about AI in 2018 (AI Now Institute, 2018 [66]).

Part I

Enhancing performance: Face analysis using RGB-D data

Chapter 4

Background

4.1 Introduction

The development of sophisticated sensor technologies gave rise to an interesting variety of data. With the appearance of affordable devices, such as the Microsoft Kinect, depth-maps, and three-dimensional data became easily accessible, which attracted significant attention in the vision research community. This research contributes to several face classification tasks using two types of visual data, the RGB images and the depth maps provided by the Kinect sensor. The goal is to improve the recognition performance by exploiting the potential provided by this information and to overcome limitations that the RGB information suffers from. Such limitations include the high variance in data caused by the variations in pose, expression, illumination, resolution, and occlusion. Our work consists first, on studying the relevance of depth data and how much it can be a promising source of information. Second on exploiting both the RGB and the depth information to enhance the performance of the classification system. Namely, three classification tasks including gender, ethnicity and expressions classification.

This chapter provides background information about face recognition, the Kinect sensor and the available RGB-D databases. Finally, we present related studies associated with face recognition using RGB-D data issued from the Kinect sensor.

4.2 Face recognition

4.2.1 Formulation and challenges

The human face plays an important role in transmitting visual information from one person to another. The diversity of information provided by the face enables us simultaneously to recognize another person, identify his gender and ethnicity, and estimate his age and emotional state. The characteristics of human face have long been a source of interest to a wide range of scientists and

have become a widely active research area for computer vision and machine learning communities.

Face analysis is a task that refers to a set of tasks that could be used for solving different problems related to facial recognition, classification, and detection, that humans perform easily in their daily lives. Facial related analysis has been an important research direction in the fields of computer vision for many decades. The importance of this field is mainly due to its wide range of applications such as biometric authentication, security systems, multimedia management, and advanced human-computer interaction. Even though many accurate modalities such as fingerprint and iris technologies, the need for face biometrics still important information to exploit [ref].

Face recognition with 2D images gained significant interest among the computer vision community. Nevertheless, they are still limited to variations in illumination conditions, occlusions, and facial expressions. Several algorithms that have been proposed during the last decade can achieve high accuracies. However, these approaches perform adequately only when the face is frontal and normalized. Recently, facial data was captured in unconstrained environments and algorithms were developed to tackle this problem. Some approaches have attempted to solve this problem by taking advantage of recent machine learning algorithms. Nevertheless, the general problem of recognizing faces under unconstrained conditions remains largely unsolved under illumination, and pose conditions.

While 2D images are not robust to these covariates, 3D face recognition was introduced to overcome these challenges. 3D images can capture more information about a face, thus enabling higher preservation of facial detail under varying conditions. Several 3D approaches have been proposed and discussed the use of 3D data alone for face recognition tasks, including identity, ethnicity and gender classification [106–109], or, in combination with 2D intensity images [110]. Furthermore, incorporating 3D information with colour data demonstrates the improvement compared to the use of only 2D images in real-world applications with unconstrained acquisition [111]. Nevertheless, a major inconvenient was the high cost of 3D scanners which limits their usage in large scale applications.

With advancements in sensor technology, low-cost sensors have been developed that provide 3D information in the form of RGB-D images. With the recent success of low-cost RGB-D cameras, such as the Microsoft Kinect devices, the depth information becomes affordable. RGB image provides the texture and appearance information, whereas the depth map provides the distance of each pixel from the sensor. The depth map is a characterization of the geometry of the face with grayscale values representing the distance of each point from the sensor. Compared to RGB data, which provides information about appearance and texture, depth data contains additional information about object shape, and it is invariant to lighting or colour variations. On the other hand, the quality of the depth maps is very low compared to that provided by existing 3D

scanners.

Since the appearance of the RGB-D sensors, many computer vision researchers started investigating the use of depth information mostly to preprocess the face and to assist the RGB images based systems. Particularly, depth data has been used in head pose variation to normalize the face into a reference pose or in face detection and segmentation. However, face analysis tasks using depth data were less investigated in the literature. Thus our motivation through this part of our thesis is to study the complementary properties between these two types of modalities by addressing different face classification tasks, including gender ethnicity and expressions classification.

4.2.2 Progress from hand-crafted features to deep learning

Face representation delves in the field of computer vision, which aims at understanding and analyzing images to define a mapping function from the visual input to a description of its content. Defining such a mapping is not trivial and consists of different steps. The main two crucial components in a given task first are to represent the image with a robust representation or a set of relevant features that are suitable for the task. Then, this representation can then be learned to perform the target task. It differs from image processing where the image input is an image, and the output is an image. Image processing steps often include processing rotation, contrast enhancement, and other transformations which preserve all the original information. In this section, we explore briefly how the feature extraction methodologies for face representation have evolved with time and migrated from hand-crafted way based on specific properties of the face to a learned way extracting higher abstractions based on end-to-end neural network approaches.

One of the original traditional approaches in face representation is Eigenfaces [112], which consists of using the PCA algorithm to find the principal eigenvectors, corresponding to the largest eigenvalues in the face images. Motivated from that, Fisherfaces [113] was proposed by using Fisher's Linear Discriminant (FLD) [114] learning algorithm which deals with the problem of the high intra-class variance. Some other global approaches extend Eigenfaces and Fisherfaces, such as Independent component analysis (ICA) [115]. In the early 2000s, this problem gave rise to local-feature-based approaches which achieved robust performance through some invariant properties of local filtering. Examples of such approaches include Gabor wavelets methods [116]. These methods exploit the fact that Gabor wavelets coefficients encode both facial shape and local appearance features. Following this direction, Local Binary Patterns (LBP) has rapidly been developed as one of the most efficient descriptors in face recognition systems [117] and many of their variants were proposed addressing

multi-scale/multi-resolution challenges [118]. There are also some other successful local approaches such as Haar wavelets [119] and Local Phase Quantization (LPQ) [120]. Some other local descriptors, such as Scale Invariant Feature Transform (SIFT) [121] and Histograms of Oriented Gradients (HOG) [122] have been commonly used for encoding edge or local shape information.

Despite the efforts made last decades to develop feature representations and provide useful low-level information from images, these feature representations are often hand-designed and require domain knowledge and human labour to achieve state-of-the-art performance in image classification and recognition. Recently, more advanced approaches based on deep neural network models in computer vision has shifted from hand-crafted features to automatically learned representations within an end-to-end neural network [123]. Methods for obtaining high-level representations based on deep neural network models have been exploited both supervised, and unsupervised learning [124]. The most used learning paradigm is the Convolutional neural network. This type of neural networks has achieved considerable success in face recognition, and computer vision [125]. The name "convolutional" refers to the used convolution operator during the optimization process. CNN has achieved significant progress due to their remarkable ability to learn prominent and robust features using multiple layers of processing units for feature extraction and transformation. The extraordinary success was first achieved by famous CNN architectures on the ImageNet object classification such as AlexNet, VGGNet, GoogleNet and ResNet [126–129]. As an application of in face recognition, many successful architectures have been proposed. Deepface [130], DeepID [131] and DeepID2 [132] on the Labeled Faces in the Wild (LFW) [133]. This remarkable success motivates the face recognition and computer vision communities to focus on improving CNN's either by proposing effective architecture or enhancing the loss function or activation layers to improve the CNNs performance. However, the main inconvenient is that a CNN architecture usually contains a large number of parameters, and require a large number of training samples for an accurate model fitting.

4.3 The Kinect sensor and the RGB-D face databases

Over the last years, a wide range of accessible sensing technology has been developed. Technological advances in image acquisition have known a considerable improvement. It has made it feasible to deploy low-cost alternatives to the traditional 3D scanners. More specifically, the appearance of the Microsoft Kinect has allowed the three-dimensional information to become easily accessible in the form of depth maps or 2.5-dimensional representation which opened a wide array of opportunities to computer vision research community. Additionally, the

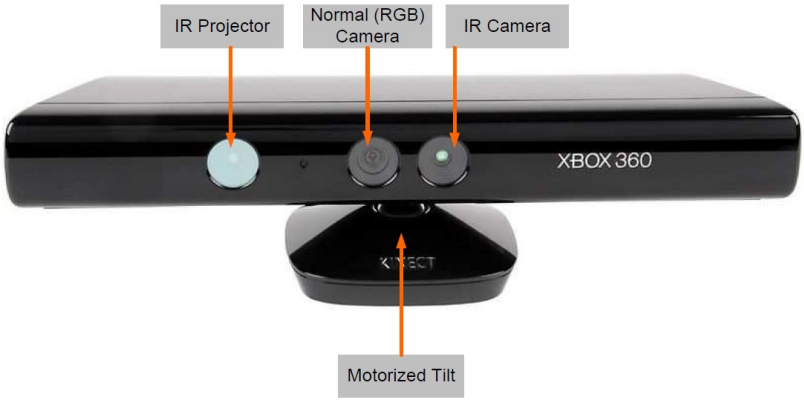


FIGURE 4.1: Kinect sensor and its components

acquired data from the Kinect has different and complementary natures, combining geometry with visual attributes that can be used to improve computer vision systems.

FIGURE 4.2: An rgb image and its depth map acquired from the Kinect sensor

The Microsoft Kinect sensor was first introduced in 2010 as a natural user interface of the Microsoft game console Xbox 360. It captures both conventional RGB images and depth maps of the scene. It has an RGB camera and an infrared (IR) emitter and camera. They are capable of capturing a coloured image and a depth scene, i.e., the distance to the observed points in the scene. However, the Kinect depth data is very noisy, and the distance computation of far objects often fails. Figure 4.1 represents the Kinect sensor with its components.

Several datasets have been built using the Kinect sensor, most of them were created in a time range from 2011 to 2014. Examples of these databases include EURECOM Kinect database, Curtinfaces database, IIT face database, FaceWarehouse database, the Aalborg university RGB-D database. We will describe publicly available RGB-D datasets mainly for face analysis applications—namely, EURECOM Kinect database, Facewarehouse database and CurtinFaces database.

EURECOM Kinect database: This database [134] contains both RGB and depth facial images of 52 subjects acquired using the Kinect sensor. There are 14 females and 38 males in the database. The people in the database belong to six different ethnicity groups (Asian, Black, Hispanics, Indian, Middle East and White). The data is captured in two sessions separated by two weeks. In each session, the facial images of each person are captured under 9 different facial variations (neutral, smile, open mouth, strong light, eyes occlusion, mouth occlusion, paper occlusion, left profile and right profile). Images of the database are cropped and manually marked facial feature points are provided 4.3.

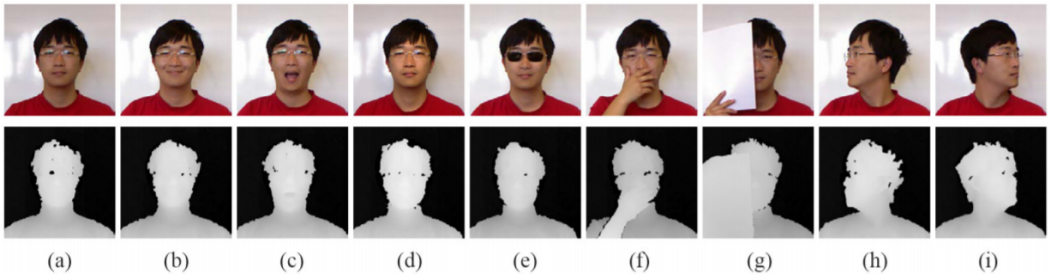


FIGURE 4.3: Eurocom Kinect database

CurtinFaces Database: The CurtinFaces database [135] contains 52 subjects of 10 females and 42 males. Three ethnic groups (Caucasians, Chinese and Indians) are included. The facial images have various variations in pose, illumination, facial expression, as well as sunglasses and hand disguise. The faces of each subject are acquired under many combinations of these challenges. For each subject, there are 97 images.

FIGURE 4.4: CurtinFaces Kinect database

The FaceWarehouse Database [136] comprises 150 individuals aged from 7 to 80. For each person, we captured the RGBD data of her different expressions, including the neutral expression and 19 other expressions such as

mouth-opening, smile, kiss, etc. For every RGBD raw data record, a set of facial feature points on the colour image such as eye corners, mouth contour and the nose tip are automatically localized, and manually adjusted if better accuracy is required. We then deform a template facial mesh to fit the depth data as closely as possible while matching the feature points on the colour image to their corresponding points on the mesh. From these fitted face meshes, we construct a set of individual-specific expression blendshapes for each person 4.5.

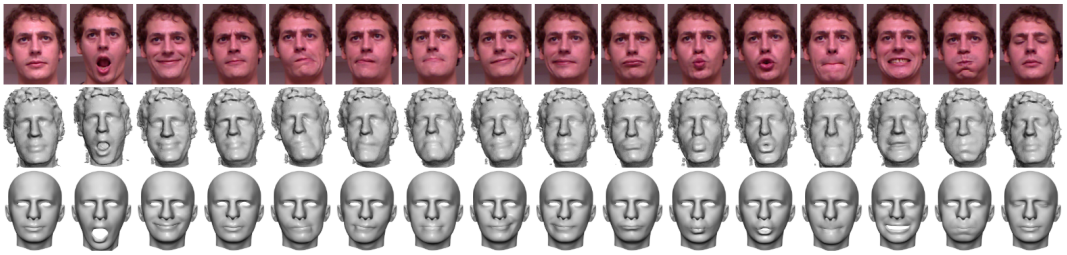


FIGURE 4.5: FaceWarehouse Kinect Database

4.4 Related works to face recognition using Kinect data

RGB-D data provided by the Kinect offer a new opportunity for computer vision and pattern classification researchers. It gained remarkable attention from the computer vision research community. Since the release of the Microsoft Kinect in late 2010, many recent papers discuss the use of RGB-D images for recognition tasks.

There are two categories of papers. The first one is related to the studies associated with face recognition using RGB-D data issued from the Kinect sensor. While the second category sees the depth information as a complement to the colour image used to improve the recognition quality. During the past decades, the performance of the combination of 2D and 3D information was heavily utilized to improve the quality of recognition [137]. Several RGB-D databases for face recognition applications have been introduced. A face RGB-D database has been proposed for recognition applications [138]. From a real-time perspective, a 3D face identification system using a depth camera has been proposed [139]. An approach based on the HOG descriptor computed on the entropy and saliency of RGB-D faces for identification applications has been introduced [140]. In the paper, it is shown empirically that depth data improve the accuracy of recognition. In another work, the problem of face recognition under different poses, illuminations, expressions, and disguise has been tackled using Kinect images [135]. The authors propose a robust preprocessing step for

estimating a canonical frontal view from non-frontal view based only the nose tip position. Another approach considers the gradient LBP descriptor for gender recognition applications [141]. The proposed approach is very efficient as it improves the recognition rate when compared to classical 2D images. A continuous authentication and monitoring system that uses 3D face images have been proposed [142]. The authors in [143] proposed the use of SURF descriptors with various enhancements on automatically generated training images using RGB-D data and a weighted score fusion for the three methods were used for making the final decision. Based on the Gabor and Gauss-Laguerre filters, the authors in [144] has developed an interesting paper to describe RGB and depth information. Authors in [145] propose a face recognition method using RGB-D data consisting of 3D face reconstruction to avoid the inconvenient of depth map corrupted by quantization noise. In a recent study, the authors explore the usefulness of depth images using four local feature extraction methods applied to identity, gender, and ethnicity recognition tasks [146]. Furthermore, in a recent work, the authors in [147] proposed the fusion of 3DLBP [148] the method with the Histogram of Averaged Oriented Gradients (HAOG) [149], a variant of HOG (Histogram of Oriented Gradients) for face recognition when Kinect is used as the 3D face scanner. Moreover, a new multi-modal approach was proposed for face recognition problem. Authors in [150] propose a method based on the entropy of RGB-D faces along with the saliency feature obtained from a 2D face. The recognition quality was performed by a tree bagger classifier. Besides, a new raw depth pose estimation and an automatic crop of facial the region was proposed in [151]. A new local descriptor (ELMDP) has been proposed by [152] applied independently on the depth and RGB images, and combined with a score fusion methodology. From a different perspective, some researchers proposed approaches based on the geometry of the face. An algorithm based on AAM methods to locate the detailed facial features from depth images has been introduced [153]. The algorithm can effectively and accurately locale facial features in various poses and complex backgrounds.

Most of these papers investigate the depth maps to improve the preprocessing task or by addressing the face recognition problem. To the best of our knowledge, by the time of our work, no research investigated the other face analysis problems from Kinect depth data such as ethnicity and expressions classification.

Chapter 5

Preleminary study: On the usefulness of depth data in face classification

Face classification problem can be formulated as a pattern recognition problem which consists of the automatic search of patterns and discovery of regularities in data to use them for data classification. Typical components of a pattern recognition system are data collection, data preprocessing, feature extraction and classification. The feature extraction and the classification are two crucial components of the recognition system, and the boundary between them is somehow arbitrarily. A good feature extractor needs to produce a representation that makes the job of the classifier easier. Meanwhile, a performant classifier must find out the relevant patterns and would not need the help of a sophisticated feature extractor. Thus, a robust pattern recognition system does not depend only on the choice of suitable approach in one of its components, but its performance relies on the interaction between the whole components [2, 6]

Before tackling a specific pattern recognition problem using multiple types of data, understanding the performance of the unimodal information is an important preliminary step to understand its usefulness for the target task. Through this preliminary work, we aim especially to answer to the following question: How much depth data can be a relevant feature using shape and texture measurements? Our goal is to gain insights into the usefulness of the depth images for face classification. To answer this question, we carried out a set of experiments comparing the performance of the depth information versus the RGB images and their combination by considering three case studies of pattern recognition, namely gender, ethnicity and expression classification as binary and multiclassification cases. Comparing the performance of the RGB images against depth maps will allow us to understand the benefits of using depth information and to evaluate its relevance to the considered task. To get a complete evaluation, we conduct experiments on every component of the pattern classification system performed on two publicly available benchmark databases. More specifically, we considered five well-known descriptors that compute the

texture and the shape properties from the data. Furthermore, we investigate three learning algorithms to find out the good learner that can recognize the extracted patterns.

To combine the information provided from both the RGB and the depth data, a feature-level fusion-based approach is proposed to deal with the different visual properties of each modality. To effectively combine the features extracted from RGB and the depth data, we used a feature selection strategy that automatically selects the most discriminative features from existing feature descriptors in a supervised way. The importance of features selected are based on AdaBoost performance on learning the relevant RGB and depth features according to the discrimination information and the target. This selectivity allows reducing the dimensionality of the feature space. Furthermore, the subset of combined features from both sources will contain a richer source of discriminative information and less redundant and noisy one.

The general procedure used for our face classification system is explained, and the techniques used are briefly described in section 5.1. In section 5.2, our experimental protocol is presented. The obtained results are reported and discussed in section 5.3. Section 5.4 concludes the paper and outlines some future perspectives.

5.1 System overview

In this section, we present our experimental pipeline comprising data collection, preprocessing, feature extraction and classification. Figure 5.1 illustrates a diagram of the components of a typical face recognition system.

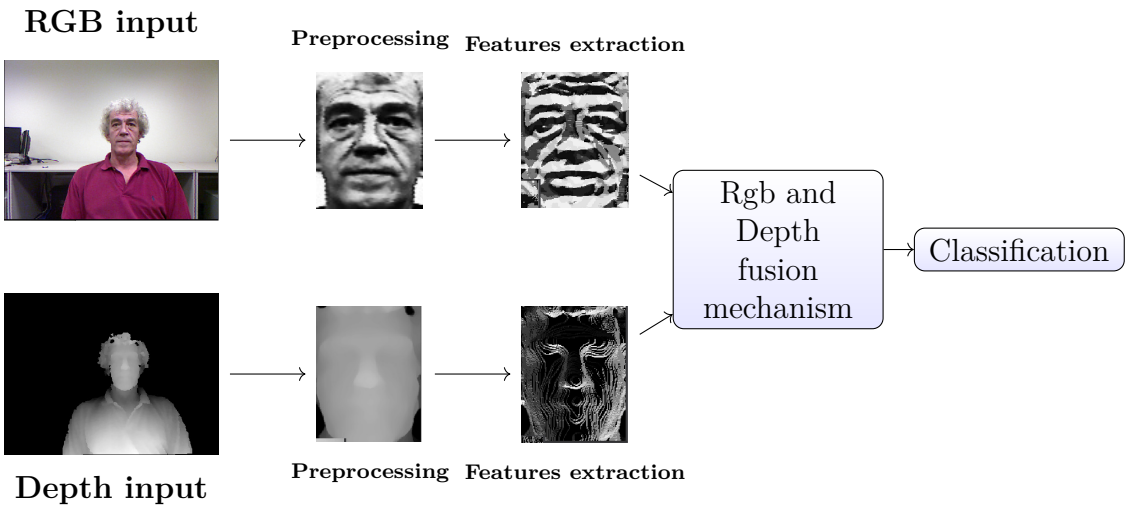


FIGURE 5.1: The general model of our procedure for face classification

5.1.1 Preprocessing

This step is essential in face analysis. It serves to normalise the face geometrically by transforming faces into a standard frame, and photometrically, based on properties such as illumination [154]. The images acquired by the Kinect sensor are characterised to be noisy and with low quality. In this framework, the preprocessing step consists first on the segmentation (or face localisation) step, which determines the location of the face in the image and segments the face area from the background. We use the Stasm package for locating a face accurately [155] which is based on the Active Shape Model [156]. The next step is face normalisation. For RGB images, Histogram equalisation is used to normalise image intensity. To eliminate the noise for depth images, we used the Median filter. An example of a preprocessed face image is presented in figure 5.1. Besides, figure 5.2 shows steps of face localisation using stasm library.

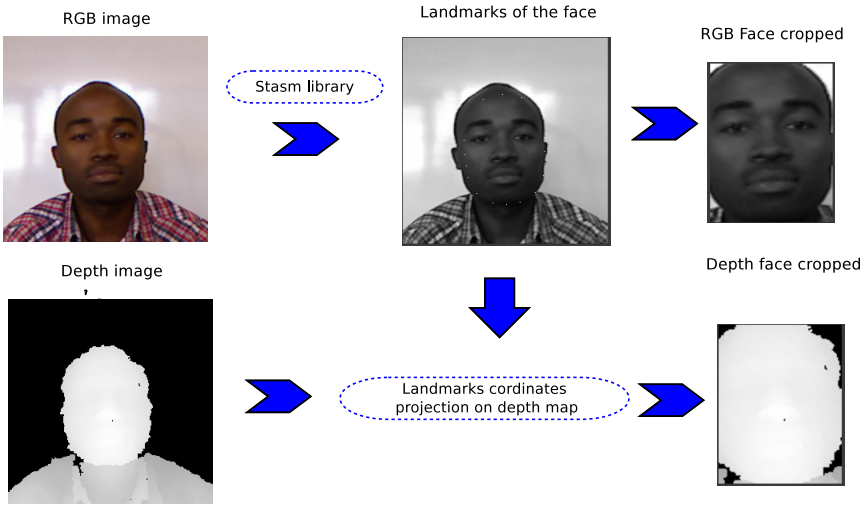


FIGURE 5.2: Face localization using stasm library

5.1.2 Features extraction

The traditional goal of the feature extractor is to characterise an object to be recognised by measurements whose values are very similar for objects in the same category and very different for objects in a different category [2]. Also, it consists of the extraction of the important information and the isolation of particular areas in the image to facilitate the processing in the next step. The choice of prominent features depends essentially on the characteristics of the problem. It must be chosen carefully. Since in this work, we are interested in face classification tasks, we consider especially texture and shape properties. Texture and shape are the most promising features for characterising the face. For this reason, we suggest applying four of the most successful descriptors used for face representation for measuring these characteristics. Namely LBP

operator and its variant for 3D images 3DLBP, the HoG descriptor, SIFT and Gabor descriptor.

Local Binary Pattern (LBP) LBP is an efficient texture operator which labels the pixels of an image by thresholding the neighbourhood of each pixel and considers the result as a binary number [157]. The computational simplicity of LBP makes it a very efficient choice for real-time applications. In LBP, each pixel is compared with its eight neighbours in a 3×3 neighbourhood by subtracting the center pixel value. The resulting strictly negative values are encoded with 0, the other values are encoded with 1. For a given pixel, a binary number is obtained by concatenating all these binary values in a clockwise direction, which starts from one of its top-left neighbours. The corresponding decimal value of the generated binary number is then used for labelling the given pixel.

3DLBP (3D Local Binary Patterns) Motivated by the original LBP, The authors in [148], proposed 3D Local Binary Patterns (3DLBP) operator based on both global statistics of geometrical features and local statistics of correlative features of 3D facial surfaces. In 3DLBP, the information of depth differences (DD) is encoded into binary patterns. The authors observe that more than 93 % of the DD between points in the radius $R = 2$ are smaller than 7. Hence, they use just three bits to represent the DD. Three binary units can characterize the absolute value of DD from 0 to 7. for each pixel surrounding the center point, there are four bits representing that position $\{i_1 i_2 i_3 i_4\}$ where $i_1 i_2 i_3 i_4$ represents the absolute value of the DD and i_1 represents the sign (encoded as the original LBP). The four bits are then separated into four layers. Then, for each layer, the corresponding bits of all the DD from the surrounding pixels are concatenated and generate one LBP code. In total, there are four LBP codes P1, P2, P3, P4, where the first LBP code is the same as the original LBP. They are called 3D Local Binary Patterns (3DLBP). For matching, the histogram of each LBP code is computed, then the four histograms are concatenated to form a unique descriptor for the image.

Histogram of Gradient (HOG) The histogram of oriented gradients (HOG), is a feature descriptor used in computer vision and image processing mainly for the purpose of object detection [158]. This technique counts occurrences of gradient orientation in localised portions of the image. The basic idea behind HoG is that an object appearance and shape can be characterised by the distribution of local intensity gradients or edge directions. To compute the HoG descriptor of a given image I, the gradients are first obtained at each pixel by computing two 1D derivatives in both horizontal and vertical directions.

Scale-invariant feature transform (SIFT) SIFT is a feature extraction technique used in computer vision to detect and describe local features in images [159]. This approach transforms the image into scale-invariant coordinates relative to local features. The scale-invariant features are efficiently identified by using a staged filtering technique. First, the key locations are identified in scale space by looking for locations that are maxima or minima of a difference-of-Gaussian function. Then, Each point is used to generate a feature vector that describes the local image region sampled relative to its scale-space coordinate frame. The features achieve partial invariance to local variations by blurring image gradient locations.

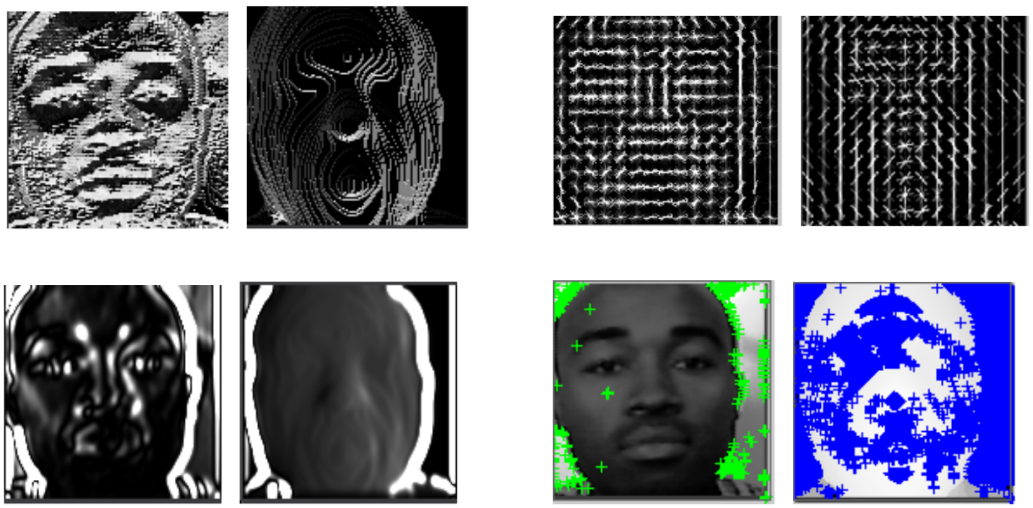


FIGURE 5.3: Example of image outputs using the four descriptors: LBP (right top), hog(top left), gabor(right bottom), SIFT(left bottom) . The rgb image in left and the depth image in right

5.1.3 RGB and Depth fusion

Since we want to study the usefulness of the combination of the depth with the RGB data for different pattern recognition problems, we propose a fusion scheme of the multimodal information by selecting the subset consisting of the relevant features from each modality. Here by "relevancy", we mean the importance of the features for a specific classification task.

We select the pertinent subset based on Adaboost classifier [160]. AdaBoost is a supervised binary learning algorithm based on the boosting paradigm. The advantage of Adaboost is that it selects only those individual features that can best discriminate among classes. Each data point is given an associated weighting parameter. At each stage of the algorithm, AdaBoost trains a new classifier

using a data set in which the weighting coefficients are adjusted according to the performance of the previously trained classifier so as to give greater weight to the misclassified data points. Finally, when the desired number of base classifiers have been trained, they are combined to form a committee using coefficients that give different weight to different base classifiers. Making an analogy between weak classifiers and features, each weak classifier is associated with one feature from the complete set. Thus, the set of best weak classifiers founded correspond to the best subset of features. AdaBoost is an effective procedure for searching out a small number of good "features" which nevertheless have significant variety and lead to good separability [161]. In our implementation, we used the decision tree as a base classifier for the boosting paradigm. The depth of the tree is set to 1 to guarantee that each tree is only a decision stump, so only one feature is contained in each weak learner.

Formally, let x_{ir} and x_{id} be the feature vectors of a the i_{th} face image obtained from an rgb and depth component respectively, where $i \in \{1, 2, \cdots, N\}$ and N is the size of the training set. y_i^k where $k \in \{1, 2, 3\}$ is the set of the corresponding labels for the three considered classification tasks. The input to Adaboost algorithm is the set of training samples $X_{tr} = [x_r x_d]$ and targets y_{tr}. The aim is to choose from the set X of n features, a subset S of m features such that S contains the information in X relevant for the classification task y. We fixed the number of features to select according to classification performance. As AdaBoost is a binary classifier, we need to extend the fusion approach for the multiclassification case. To do so, we get the set of features by using One Versus the Rest approach [162]. Given M-class classifiers, it is common in machine learning to construct a set of binary classifiers $f_1, f_1, \cdots, f_m$. We trained each binary classifier to separate one class from the rest and obtain the set of the selected features for each. The final subset is obtained by combining the obtained subsets and removing the redundant indices(alg.1,2).

Algorithm 1 Feature selection for the binary case

```
Input: X_tr, y_tr
training
for i ← 1, weaklearner do
    for j ← 1, nodeinthetree do
        get the variable index
Return S
```

Algorithm 2 Feature selection for the multiclassification case

```
Input: X_tr, y_tr
construct a set of binary classifiers
for i ← 1,binary classifier do
    for j ← 1, set of selected features do
        combine the obtained indices
        eliminate the reduduant features
Return set of indices of selected features
```

5.1.4 Classification

The goal of the classifier is to use the vector provided by the feature extractor to assign the object to a category. Techniques used for classification are useful for recovering the model that generates the pattern accordingly to the pattern complexity and depending on the type of candidate models themselves [2]. In this work, we compared the performance of three learning paradigms to recognise the binary and the multiclass patterns in the extracted features. Namely margin-based classifiers, bagging and boosting methods. Four popular classifiers are used. Namely, Support Vector Machines, Random Forest, Adaboost, and Gradient Boosted Trees [161, 163–165].

5.2 Experimental study

In this set of experiments, we compare the obtained results for the three classification tasks using information extracted from only one modality, RGB or Depth and their combination. We performed the classification by comparing three learning paradigms. We first introduce the dataset used in our experiments. Details about our experimental procedure are presented. Furthermore, the obtained results are reported and discussed.

5.2.1 Datasets

Our approach is tested on the combination of the two benchmarking RGB-D databases: EurocomKinect Database [134] and Curtinfaces Database [135]. In fact, the most existing RGB-D databases contain a small number of individuals, and many of them are biased. In addition, we do not have access to other private databases such as FaceWarehouse 4.5. A good approach to avoid these drawbacks might be to combine multiple Databases. Nevertheless, the two databases contain variations in illumination, different poses, and various facial expressions. Thereby, the combination led to a rich RGB-D dataset, which allows a good evaluation for recognition systems. The resulting dataset consists of 572 images of 104 individuals with variations in gender, ethnicity, expressions

and illumination. The images in the datasets contain three basic facial expressions: neutral, smile, disguised/angry, and three ethnic groups (white, Asian and others) and different illumination variations (strong and poor illumination) (figure 5.4). While the two databases have different acquisition settings, only useful images for the studied tasks were considered.

FIGURE 5.4: Some individuals from the used database

5.2.2 Experimental setting

As described in section 5.1.1, we used the Stasm library to locate faces for RGB images. Stasm is a toolkit for finding features in faces. The software takes a facial image in input and returns the positions of the facial landmarks. For depth images, faces are obtained by projecting the face coordinates from the RGB image on the depth image. After the face location step, faces are resized in a uniform size (77×99). We use five face representation techniques, HOG, LBP, 3DLBP, SIFT and GABOR. The resulting vector is normalized and passed to the classifier. For AdaBoost feature selection, we limit the number of the weak classifiers to a thousand as the optimal found size in our experiments. In PCA, we use the total variance of data of 95%. The classification was performed by comparing three types of classifiers. Principally, Support Vector Machine (SVM) using the RBF kernel with $\gamma = 0.001$ and $C = 100$, the Random Forest algorithm uses a population of 1000 trees and Boosting classifiers uses 1000 weak classifier. The optimal hyperparameters were optimized by cross-validated grid-search over a parameter grid. To evaluate the overall performance, we employed the receiver operating characteristic (ROC) curve. It is plotted by calculating sensitivity (True Positive Rate; TPR) against the specificity (False Positive Rate; FPR) under different thresholds. The area under the ROC curve, termed as AUC, is often used as a metric to access the overall performance. The AUC score ranges from 0.5 to 1. The larger score a predictor achieves, the better performance it has. Here, the stratified variation is used as a variation of KFold cross-validation where the folds are made by preserving the percentage of samples for each class with $k = 10$. For this, we divided the data into k equal

folds (portions). We then trained the model on the $k-1$ folds and tested it against the remaining folds. That process was repeated $k = 10$ times. The final performance after that corresponds to the average of the obtained values. We used cross-validation analysis to ensure that all data was used for both training and test.

5.3 Results and disscussion

The results are presented for the three classification tasks. Figures 5.5, 5.6, 5.7 represent the obtained ROC curves for gender, ethnicity and expressions classification tasks. Three classifiers are used for this classification. Namely, SVM, Random forest (RF), AdaBoost (AB) for the binary case and the Gradient Boosted Tree classifiers (GBT) for the multiclassification case. In this set of experiments, we compare the obtained results using the information extracted from only one modality, RGB or depth for each task. Single modality and bimodality experiments have been executed with four features extractors and three classification algorithms.

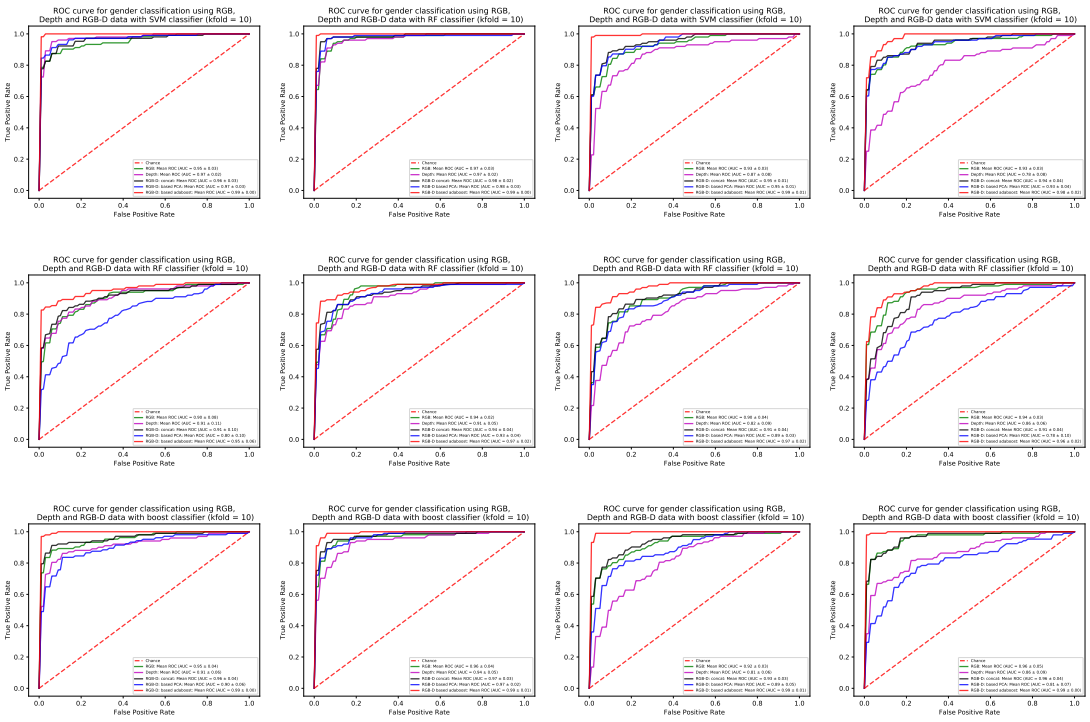


FIGURE 5.5: Gender classification with SVM (top row), Random Forest(middle row) and boosting classifiers(bottom roow) using LBP, HOG, SIFT, GABOR feature extractors from right to left respectively

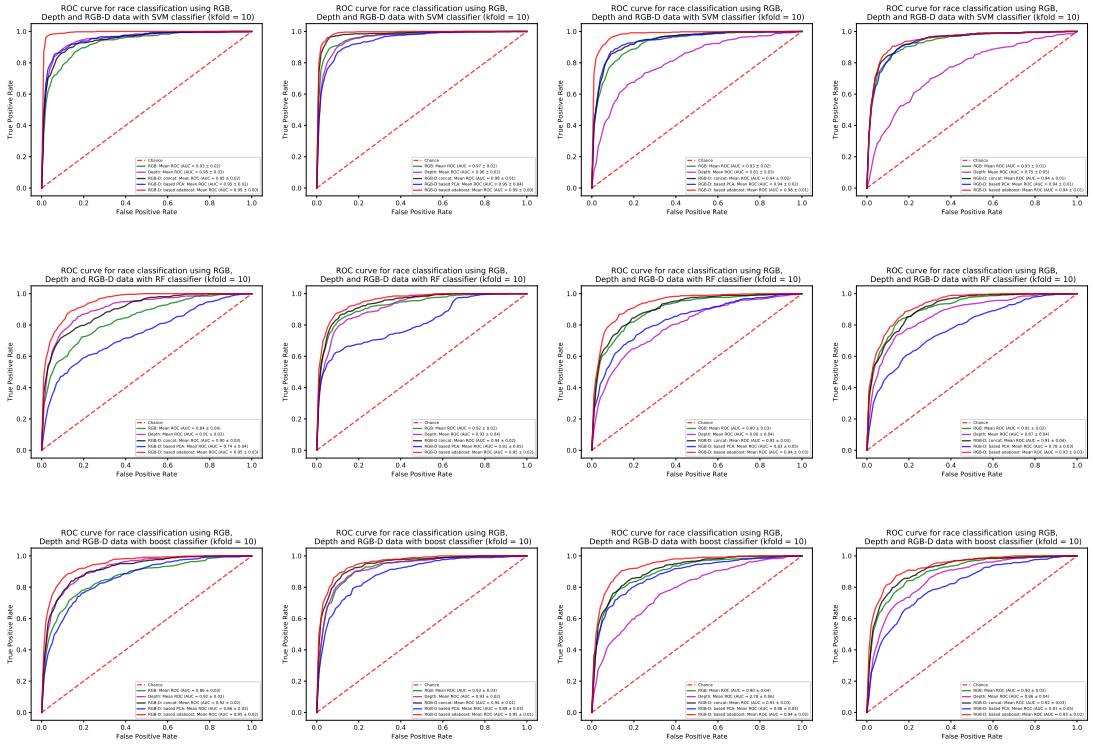


FIGURE 5.6: Ethnicity classification with SVM (top row), Random Forest(middle row) and boosting classifiers(bottom roow) using LBP, HOG, SIFT, GABOR feature extractors from right to left respectively

5.3.1 Single modal comparison

In this set of experiments, we use information extracted from only one single-modality. The aim is to compare the performance of the recognition system using only RGB or depth information for different classification tasks. Thus, we reviewed four different types of shape and texture-based feature representation encoding algorithms in order to investigate which features are more informative. The obtained results show that the use of the depth data had led to competitive results compared to those of RGB, mainly when using LBP/3DLBP and HoG features. For example, for gender recognition, using the LBP descriptor achieved an AUC of 95% and 97% for RGB and depth images, respectively using the SVM classifier. Likewise, as for the HoG feature, the achieved performance was 97% for both modalities. These observations are similar for ethnicity and expressions classification tasks. As for the other classifiers, the SVM, RF and AB are competitive with each other, while the performance achieved by the SVM classifier is generally higher than using other classifiers among the five compared descriptors in terms of AUC. However, when using sift and Gabor features, the results of the depth still not good enough compared with RGB data

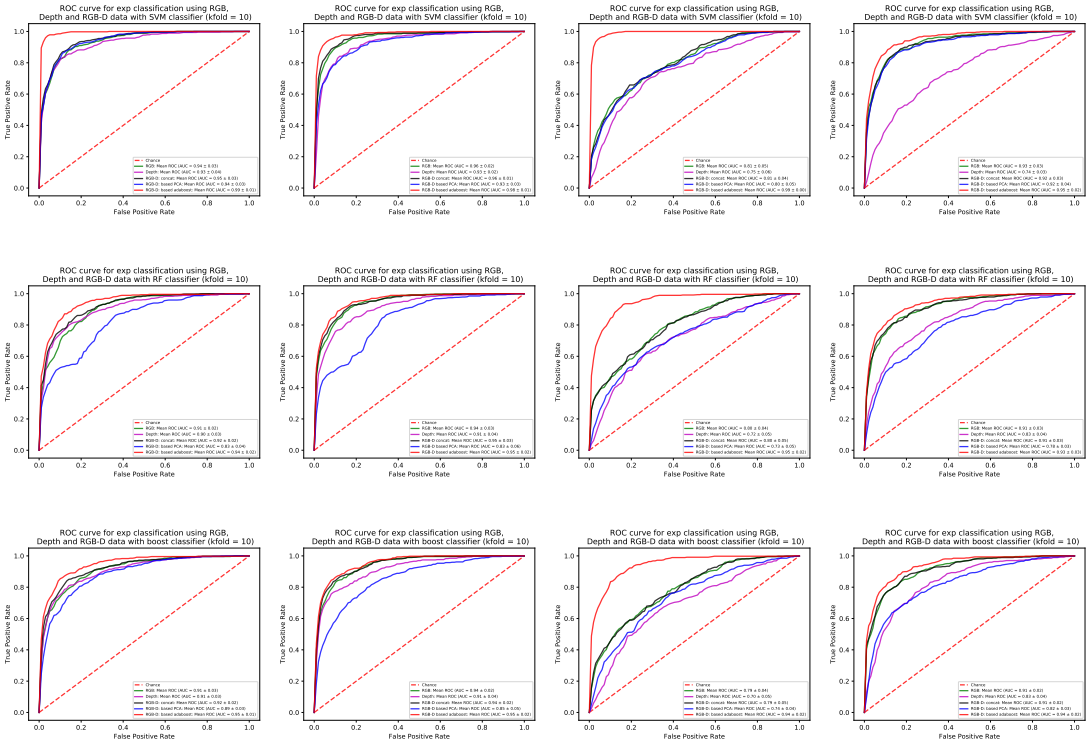


FIGURE 5.7: Expressions classification with SVM (top row), Random Forest(middle row) and boosting classifiers(bottom roow) using LBP, HOG, SIFT, GABOR feature extractors from right to left respectively

which shows that those descriptors are not suitable to extract useful characteristics from depth data. This shows that relevant information can be extracted from the depth images alone, despite the fact that we used state-of-art descriptors that are usually used for RGB data.

5.3.2 Bimodal performance

In this set of experiments, we use information extracted from both modalities, RGB and depth. As described in section 5.1.3, multimodal data is fused on the feature level for the different types of feature sets. We used the best RGB and depth features for combination, i.e. the selected features using AdaBoost-based for each classification task. We first extract all RGB features for depth features and concatenate them. Then we apply AdaBoost to select the best 1000 features. This choice is based on the fact that the best performance of AdaBoost is achieved with 1000 features. Figures 5.5, 5.6, 5.7 shows the ROC curves for the obtained performance compared with a simple concatenation and with PCA based dimension reduction approach. These results indicate that integrating

information from both RGB and depth data leads to a significantly improved performance over single-modal. When comparing the two different types of fusion, the feature level fusion by the AdaBoost-based selection performs better well for the three classification tasks and for the four descriptors. The PCA-based features resulted in lower performance, mainly when using Random Forest classifier and achieved a good performance only when using the SVM classifier. This is not surprising since a supervised feature-level fusion is expected to take advantage of the correlation between the two modalities which is stronger and select the best features sets from each modality that contains the most informative features for the task. In addition, it allows of a smaller set of the most informative and correlated features from each modality.

5.3.3 Features analysis

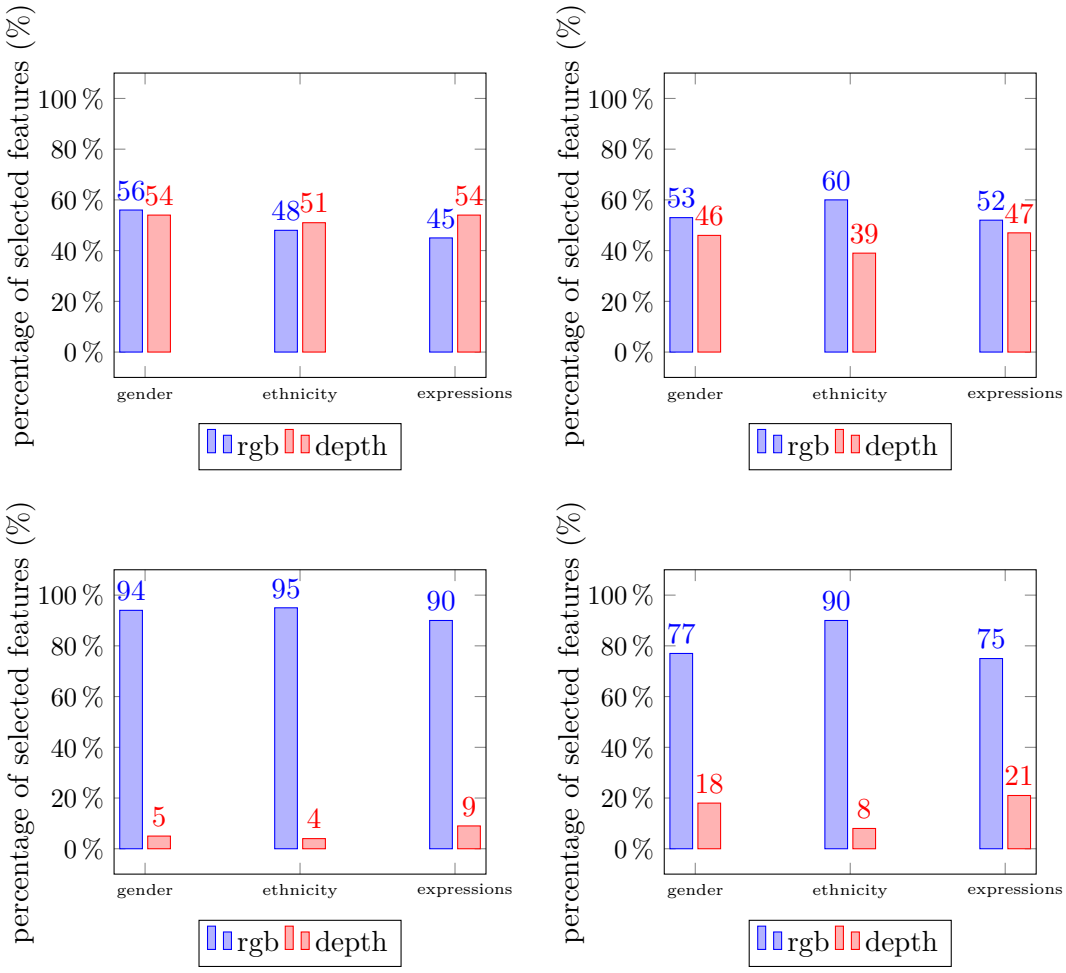


FIGURE 5.8: Distribution of the selected rgb depth feature for each descriptor for the three classification tasks

As the concatenation of RGB and depth by AdaBoost performs better when the best features set useful for each task from each modality are used. Therefore, it is worthy of investigating which RGB and depth features are selected. Figures 5.8, show the distribution of the AdaBoost-selected RGB and depth features and the importance of the features from the various descriptors for each classification task respectively. The features that have been most frequently selected by the AdaBoost from either RGB and depth features are those belonging to the LBP and HOG features. The depth features being least frequently selected are those from the SIFT and Gabor feature sets, and the dominant features that belong to the RGB data. This can be explained by the fact that these descriptors are not suitable to extract informative representation from depth data. Furthermore, the selected features vary according to the discriminative relevancy of the features to perform the classification.

On the other hand, the state-of-art papers discuss the usefulness of the depth with a focus on feature extraction only for face recognition. In contrast, our work tries to address the classification task. As far as we know, we are not aware of a similar study for RGB-D images.

5.4 Conclusion

In conclusion, we investigated in this preliminary work the usefulness of depth information for gender, ethnicity and expression classification. Five state-of-the-art descriptors, three classification approaches were investigated and tested for RGB, depth, and their fusion. The results show that depth data can be a rich source of information which has promising and beneficial uses in face recognition applications. Moreover, the results show that the accuracy of a recognition system depends on the choice of the most suitable combination between the classifier and the feature extraction technique. Mainly, using HoG and LBP showed promising results with both SVM for binary classification and muti-classes case. The results also suggest that integrating the information from both modalities lead to improved reliability over single-modal approaches. The experimental results have demonstrated that the proposed fusion approach improves the system performance and shows the effectiveness of the candidate features selected for fusion according to the studied classification task. We consider the content of this paper to be a preliminary work on this topic. Perhaps it can be a promising direction to use as a starting point for the next investigations.

Chapter 6

Multimodal face classification using RGB-D data

The obtained results in the previous experimental study (chapter 5), showed that the additional information carried by the depth modality is promising and beneficial for solving the face classification problem. We extend our work by studying how can we select an optimal strategy for the face classification system that can better exploit the texture and shape characteristics provided by the two modalities, and therefore improve the recognition performance. In this chapter, we first, present the motivation of this research in section 6.1. Section describes 6.2 the proposed approach. Our experiments are detailed in section 6.3. The results are reported in section 6.4 and disscussed in section 6.5. Finally, conclusions and future directions are presented in section 6.6.

6.1 Motivation

Face analysis is a long-standing challenge in the field of pattern recognition. The goal is to not only to create systems that detect, verify and recognise faces but also to understand characteristics of the human face and facial diversity.

The human face presents great potential containing a wide variety of features. The diversity of information provided by the face enables us to simultaneously recognise another person, identify his gender and ethnicity, and estimate his age and emotional state. However, we cannot say precisely how we make that judgment. In the literature of psychology and neurophysiology, many studies have made significant progress at understanding the effectiveness of the global versus the local discriminative information in visual processing such as gender identification and expression analysis. These studies focus on the understanding of what kind of features are more distinguished in the perception of gender [1], [2], [3], [4]. For instance, to figure out how human brain discriminates between males and females, several approaches have been used in the previous studies, based on presenting in isolation, masking or replacing within a full image individual features (brows, eyes, nose, mouth, and chin) or pair ones (brows and eyes, eyes and nose, nose and mouth, mouth and chin). These studies have

produced varying results indicating the influence of each part in sex discrimination. In general, the results showed that the eye region and the face outline (particularly the jaw) has the highest load in judging the gender.

From a pattern recognition point of view, face classification is a challenging problem. The challenges mainly come from the large variations in the visual stimulus caused by the changes in illumination conditions, gender, ethnicity, expressions and age variations etc. This wide range of variations leads to a high degree of intra-class variance. One approach to handle the problem of the large intra-class variation is to address the problem based on local features. Existing face representations methods consist into two categories: global-based approaches where the feature vector contains some holistic characteristics about the face and local-based approaches corresponding to a particular local region in the face and encoding the specific traits within this specific area. While global-based face representations have shown a good performance, however, local approaches are believed to be more robust to the intra-class variations to some extent. Based on these observations, we aim here at studying the RGB-D face classification problem, by first understanding the facial diversity in the presence of the unimodal and multimodal information, and second by studying an optimal strategy that allows the system to better exploit the multimodal information. Therefore, this problem is addressed as a multiple classifiers system where the set of classifiers is trained on separated facial parts from the RGB and Depth faces. We assume first that extracting features from local face areas encode more detailed local features. Second, that each RGB and depth facial part is characterised by a specific shape and texture properties from the other part, and therefore they contain diverse information. Then, in order to exploit this diversity all these classifiers are combined as an ensemble classifier system to predict the final decision based on the proposed sum rule based on each model performance. Three classification tasks were considered gender, ethnicity and expressions classification. In summary:

- We aim at finding out which are the most discriminative facial parts that play an essential role in recognising the gender, ethnicity, and the expressions using the unimodal information and the combined RGB-D data.
- We address learning from local facial parts to deal with the intra class-variation present in the RGB and Depth faces and to evaluate the local representation of the RGB versus the Depth information and their combination. We further explore the imbalance issue and noise issue present in the dataset jointly.
- We propose an ensemble classification approach based on combining the complementary and diverse information learned from each separated facial part as a component classifier. Then, to exploit the facial diversity, the final decision is rendered by a proposed scoring rule that combines all components decision based on their training performance.

6.2 Proposed approach

In this section, we first present the global mathematical framework of the proposed approach, next a detailed description of each component of the ensemble system, is provided.

6.2.1 Mathematical formulation

Considering a face classification problem, where patterns have to be assigned to one of the possible m classes $(y_1, y_2 \cdots y_m)$. Let us assume that for a given face F, we extract a set of n parts denoted $P = \{P_1, \cdots, P_n\}$, where $F = \{\cup_{i=1}^{n} P_i\}$. Each part P_i is representing the given pattern by the input rgb and depth features space denoted respectively x_i^r and x_i^d.

Given a set of training examples, the learning algorithm outputs a set of hypothesis (classifiers) denoted $h = \{h_1, h_2, \cdots, h_n\}$ trained on each element from P. Let $a = \{a_1, a_2, \cdots, a_n\}$ and $l = \{l_1, l_2, \cdots, l_n\}$ be respectively the set of (balanced) training accuracies and missclassification loss. A score S_i is assigned to each part based on its training performance as the following:

$$S_i = \frac{a_i}{l_i} \tag{6.1}$$

For a given input sample I from the testing set, let $D_I = \{d_1, d_2, \cdots, d_n\}$ be the set of the decisions corresponding to the label outputs (decisions) predicted by h. We compute first, the sum of scores of the models with the same label output $j \in \{1 \cdots m\}$:

$$Score_j(I) = \sum_{i=1}^{n} S_i \quad \text{with} \quad d_i = d_j \tag{6.2}$$

Then the class which receives the highest score is then selected as the final decision assigned to the sample.

$$d_{Final} = d_i \quad \text{with} \quad Score_i > Score_j \tag{6.3}$$

6.2.2 Ensemble components description

A multiple classifiers system consists of constructing a set of classifiers in the training step, also called an ensemble of classifiers. The predictions made by this collection are aggregated to predict class labels in the testing set. Building an ensemble of classifiers consists of two phases—the creation of the ensemble and the combination of decisions. In this section, we describe in details the used techniques in each step.

Creation of ensembles

Two architectures can be used, a serial or a parallel one. In the proposed approach, the set of classifiers are trained in parallel, and their output is combined afterwards to give the final decision. Before combining the decision, the system is designed on three levels (Figure 6.1), the data level, the feature level, the classification level:

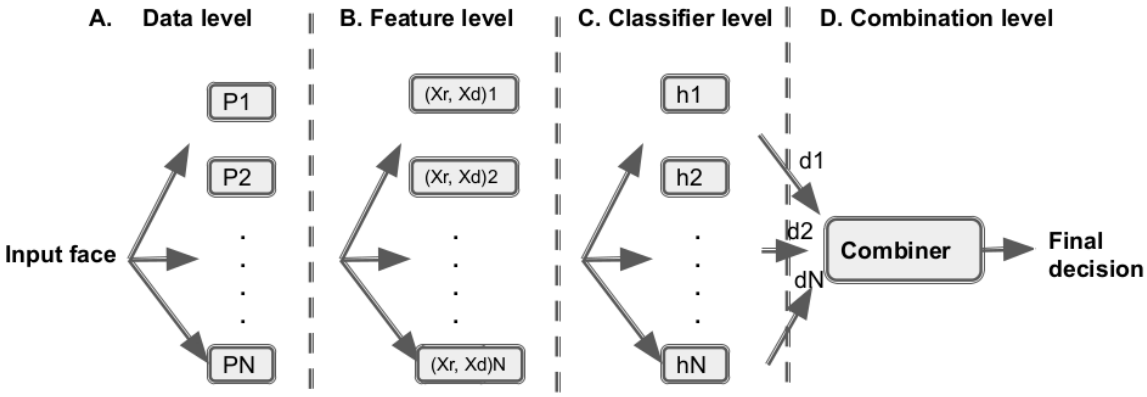


FIGURE 6.1: Global architecture used for the ensemble

Data level The set of base classifiers is generated on the data level. We divide the input set of samples into a different subset of samples. As mentioned previously, the studies done on understanding the face perception showed that each facial part represents different properties and representations of patterns. Thus, different subsets are constructed by dividing the face into separated facial parts. By using this prior knowledge about the problem, we aim at by this way at creating diverse classifiers that are good to combine. We automatically extract six face parts from the RGB and depth face images. Namely, eyes, cheek with the nose, nose with the mouth, chin, chin with mouth and jaw, nose.

Feature level On the feature level, we use different methods for the preprocessing and extracting the features vectors from the constructed subsets:

Pre-processing: To preprocess the RGB images, we apply the Histogram equalization in order to normalize the image intensity. Further, to eliminate the noise for depth images, we used the Median filter.

Feature extraction: In order to extract the distinguishing features from the images, we suggest applying two of the most state-of-art successful descriptors used for face classification. Namely, LBP [157] and 3DLBP [148] operators. Moreover, we investigate deep learned features, which are automatically learned from the images differently from handcrafted feature representations. As we are dealing with a small dataset, we choose the autoencoders as an unsupervised

learning algorithm. Our primary goal is not to propose a good descriptor approach but to study the usefulness of the depth data through using separated facial parts. Therefore, we selected some of the best performing state-of-the-art feature extractors to test our ideas.

Rgb and depth fusion: The multimodal fusion of RGB and depth modalities is performed using the previous method based on AdaBoost algorithm as a feature selection method.

Classifier level On the classifier level, we apply the same learning paradigm to train each subset. After feature extraction, we obtain feature vectors from each facial part.

Combination of decisions

After creating an ensemble, many rules can be used to obtain the final decision from the classifiers combination. In order to effectively combine the individual learner in the context of classification, we aim at selecting an appropriate way to fuse the outputs that can highly improve the performance of the system. Thus, we defined a score associated with each classifier based on its discriminative performance in the training step (equ. 6.1). The ensemble constructed with classifiers is combined using a sum rule as the following. We add the scores provided by each base classifiers and which predict the same output (equ. 6.2). Then the class label with the maximum score is assigned to a given input pattern (equ. 6.3).

6.3 Experimental study

In this section, we are describing our experiments in more detail. The approach is tested on the same dataset used in the previous work 5.2.1. Therefore, we start by explaining how we preprocessed it and performed the feature learning and classification.

6.3.1 Pre-processing Pipeline:

To extract the facial parts from the face image, the Stasm library is used to locate face parts for RGB images [155, 156]. Stasm is a software library for finding features in faces. Based on the Active Shape Model (ASM), it takes a facial image in input and returns the positions of 77 facial landmarks. For each input rgb image, 77 face key-points using are extracted. These landmarks are used to comprise the measures which characterise all the horizontal and vertical dimensions which localise each facial part (Fig. 6.2) For depth images, faces are obtained by projecting the face coordinates from the RGB image on the depth

image. All parts are normalised in a uniform size. Figure 6.2 shows the steps of face parts extraction using a Stasm library.

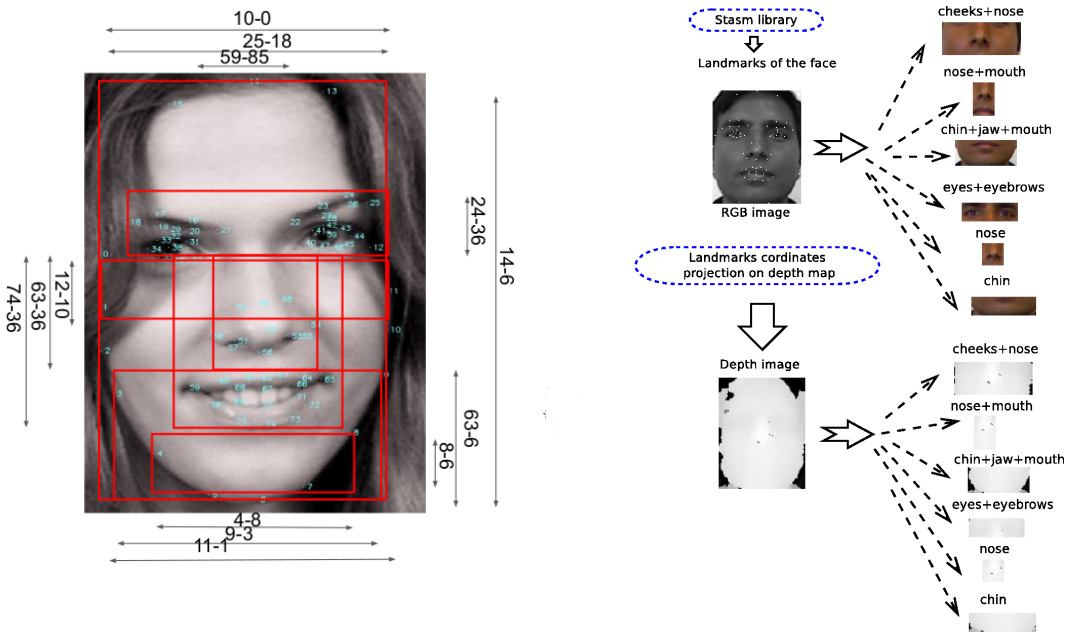


FIGURE 6.2: In left, the 77 key-points extracted using STASM from each face (small enumerated dots). In right, the preprocessing steps to extract facial part from rgb and depth faces

6.3.2 Experimental setting and evaluation

In all the experiments we performed, the data set was randomly divided into two sets: 70% as the training set and the remaining 30% as the testing set. This process was repeated 25 times. Then the results obtained are averaged in order to get the final results. In order to assess if the performance of the ensemble is statistically significant than using the whole face image, t-tests have been conducted with 95% confidence interval. There are three stages of our experiments. In the first stage, we evaluate the training performance of the extracted parts for gender, ethnicity, and expression classification tasks in the context of RGB-D information. Training performance is evaluated using the stratified 10-fold cross-validation. To compute the score used for each facial part, we calculated the balanced accuracy in both binary and multiclass cases to deal with imbalanced datasets. It is defined as the average of recall obtained on each class (equ. 6.5). The misclassification rate is calculated as the percentage of incorrectly classified instances. The classification was performed using the Support Vector Machine (SVM) using the RBF kernel with $\gamma = 0.001$ and $C = 100$. For rgb and depth fusion based on AdaBoost, we limit the number of weak

classifiers to 1000 as the optimal number. Three evaluation metrics are used. The F1-measure (equ. 6.6), the balanced accuracy and the misclassification rates.

$$Precision = \frac{TP}{TP + FP} \tag{6.4}$$

$$Recall = \frac{TP}{TP + FN} \tag{6.5}$$

$$F1 - meseaure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6.6}$$

where, TP and TN correspond respectively to true positive and true negative samples, while FP and FN represented the false positive and false negative errors.

6.4 Results and analysis

In order to have a general idea about the behaviour of the ensemble and be able to account for questions that we have set out, we conducted a series of experiments and have divided this section into three main points according to each one of the goals of the study and a final one where we discuss and sum up the results obtained. First, we examine the performance of the ensemble. Second, we evaluate the relevancy of the local versus the global unimodal/multimodal information. Third, we evaluate the diversity of the different components. The experiments were performed for three classification tasks, including binary and multiclassification cases.

6.4.1 Ensemble performance

Through this experiment, we carry out a comparison of the proposed approach with both the performance obtained using the whole face and with individual component classifiers (facial parts). The aim is two-fold. First, to determine if the proposed ensemble improves the results and second to investigate the approach improvement w.r. to using one single model. That is whether the tradeoff between complexity increment and performance enhancement is justified or not. We have compared our approach with a simple combination without considering the score affected to each base classifier as well. Tables 6.1, 6.2 and 6.3 show the general results for respectively, gender, ethnicity and expressions classification tasks.

Compared to using the global face information, the results showed a considerable improvement for the three classification tasks in term of the balanced accuracy and the F-measure. For gender classification (table 6.1), an improvement of 5% is shown when using RGB only, or depth only, enhancing mainly

the performance evaluated by the F-measure metric. The noticed difference between the values obtained by both evaluation metrics is mainly due to the imbalance in gender classes. However, the results have remarkably increased using both modalities, from the accuracy of 91% and an F-score of 84.71% to a stable value of $\sim$ 94% for both evaluation metrics. For ethnicity classification (table 6.2), the results have improved with approximately 4% using the RGB and the RGB-d information achieving a recognition performance of 89%. Using only the depth information, the performance is similar. For expressions classification (table 6.3), the results have increased especially when using the RGB information and the combined RGB-D information with $\sim$ 4%. When using the depth information, the results reached an improved with 2 % over the evaluation metrics.

Compared to the individual facial parts, the performance of an ensemble is much better than any of the individual classifiers in the ensemble. As a first observation, the performance of the ensemble has increased significantly when using RGB-D information. An improvement of 6 % and an 11 % in term of respectively the accuracy and the F-score was achieved for gender classification achieving a stable recognition rate of 94 %. For ethnicity classification, an improvement of 11 % was obtained for both metrics. For expressions classification, the results were increased with 7 %. In the case of using one modality, the proposed fusion approach has improved the classification performance from the individual base learners, especially for the depth modality, which shows the diversity among individual learners. A remarkable improvement of 5 % and 11 % over the obtained accuracy for respectively gender and ethnicity classification tasks. Moreover, an improvement if 9 % and 12 % was achieved over the F-score metric for the two tasks, respectively. Meanwhile, when using the RGB information only, the ensemble performance for gender recognition reached the same accuracy of the best individual learner; however, the f-score score was increased with 6 %. For ethnicity classification, the recognition performance increased with 3 % and 8 % over the accuracy and the F-score. However, for expressions classification, an improvement of 4 % was obtained using the RGB and depth information in term of both metrics

Finally, the results obtained by combining the predictions using the proposed score are much better than a simple combination (non-scored fusion), which shows the effectiveness of the approach.

6.4.2 Global versus Local combiner

Herein, we evaluate the approach for each modality individually, and the bi-modal combination with using the whole face. We compare the local(facial parts) versus the global(whole face) potential using the unimodal and for the bi-modal information. First, we discuss the discriminative relevancy of each facial part for each task. Second, we discuss the local and the global relevancy of

each modality to determine how much using the local cues of each information is more useful in exploiting the potential that the data provided.

TABLE 6.1: Performance of the ensemble for gender recognition using rgb, depth and rgb-d information

	rgb		**depth**		**rgb-d**	
	balanced accuracy	**F-score**	**balanced accuracy**	**F-score**	**balanced accuracy**	**F-score**
cheeknose	70.90	57.14	73.82	59.64	86.56	77.27
chin+mouth+jaw	79.62	65.51	78.19	69.23	84.09	76.15
eyes	84.30	72.41	61.39	52.29	84.71	77.70
chin	67.59	46.66	69.27	52.83	74.32	58.97
nose	67.59	46.66	58.82	45.33	77.99	66.33
nosemouth	72.17	53.33	65.18	45.83	84.75	76.75
face	79.92	74.41	79.54	73.46	91.08	85.71
non-scored fusion	79.41	74.41	74.09	69.77	92.20	91.59
scored-fusion	84.50	79.16	83.37	78.87	94.75	94.33

TABLE 6.2: Performance of the ensemble for ethnicity classification using rgb, depth and rgb-d information

	rgb		**depth**		**rgb-d**	
	balanced accuracy	**F-score**	**balanced accuracy**	**F-score**	**balanced accuracy**	**F-score**
cheeknose	64.72	63.40	69.75	69.69	78.31	78.37
chin+mouth+jaw	62.45	58.47	63.55	63.14	66.70	66.87
eyes	74.37	72.34	71.14	71.30	77.77	77.71
chin	54.25	49.40	51.46	51.12	61.36	61.59
nose	64.06	60.98	62.63	62.34	69.80	69.62
nosemouth	69.00	68.69	65.13	65.14	68.67	68.47
face	78.40	76.22	82.81	82.81	85.99	86.12
non-scored fusion	74.07	72.70	75.11	75.84	85.62	85.69
scored-fusion	81.79	80.05	82.15	83.01	89.36	89.57

Regarding the facial parts relevancy for each classification task, the results show that the most informative part using the RGB information to recognise the gender is the eyes region and is the lower part of the face when using the depth information. By combining the RGB-D information, almost all part achieved a good performance except the nose and the chin parts. For ethnicity classification, the eyes are the most discriminative local parts for the cases when using the RGB information or the Depth information. For expressions classification, the middle and the lower parts of the face are the most informative using both the unimodal and multimodal information.

Next, we report the potential of local versus global (whole face) information for each type of information. We see from the results that the relevancy of

TABLE 6.3: Performance of the ensemble for expressions classification using rgb,depth and rgb-d information

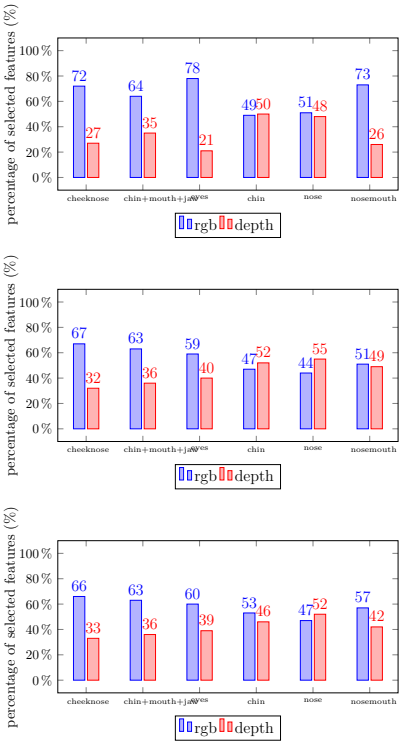
	rgb		depth		rgb-d	
	balanced accuracy	**F-score**	**balanced accuracy**	**F-score**	**balanced accuracy**	**F-score**
cheeknose	70.66	72.01	71.43	73.61	76.31	77.90
chin+mouth+jaw	69.36	70.15	65.20	66.68	72.08	73.59
eyes	51.16	51.85	45.01	44.81	52.89	52.93
chin	44.00	43.54	43.88	44.02	55.76	57.79
nose	56.31	57.03	60.39	60.51	60.03	61.89
nosemouth	60.55	61.67	57.15	57.71	70.15	71.47
face	71.09	72.72	72.95	75.30	79.36	80.20
non-scored fusion	66.91	70.11	62.71	66.31	71.49	75.43
scored-fusion	74.67	76.88	74.95	77.17	83.21	83.99

each modality depends on the considered classification task. For the case of gender classification, the RGB information is more relevant in using local information (eyes region) than the global RGB face. The classification accuracy of the ensemble remained of a similar value than the best individual learner (eyes region), while it performed significantly better in terms of F-measure. The global depth information is more effective. However, the combination of local parts showed a considerable improvement in the accuracy compared with global information. the RGB-D fusion for the global face performs better than individual learners.Nevertheless, the ensemble approach as a combination of the local parts achieved a significant and stable performance for both metrics. In the case of the ethnicity classification task, even though global information performs better when using the RGB, depth and their combination, the combination of the models built from local parts improves and achieved a more stable performance especially when using fusing the RGB and the depth. Finally, for expressions classification, the local parts perform better than the global face. The performance of the ensemble improves classification performance. Using the RGB-D features lead to obtaining a more stable performance for both evaluation metrics.

Features analysis: In this section, we present the selected features for each facial part for each classification task. The aim is to visualise the features that have been frequently selected by Adaboost from either the RGB and depth information. The distribution of the selected features is presented in figure 6.3.

From the results, we see that the percentage of selected RGB and depth features differ for each classification task. In general, the RGB features are the most selected mainly for the eyes region, the cheeknose and the lower part(mouth+jaw) of the face. For the chin and the nose parts, the number of selected features are equal. For the nosemouth part, the RGB cues are more

selected than depth for gender classification. However, for the ethnicity and expressions tasks, both types of information are equally selected.



	rgb	depth	rgb-d
cheeknose	70.90	73.82	86.56
chin+mouth+jaw	79.62	78.19	84.09
eyes	84.30	61.39	84.71
chin	67.59	69.27	74.32
nose	67.59	58.82	77.99
nosemouth	72.17	65.18	84.75

	rgb	depth	rgb-d
cheeknose	64.72	69.75	78.31
chin+mouth+jaw	62.45	63.55	66.70
eyes	74.37	71.14	77.77
chin	54.25	51.46	61.36
nose	64.06	62.63	69.80
nosemouth	69.00	65.13	68.67

	rgb	depth	rgb-d
cheeknose	70.66	74.43	76.31
chin+mouth+jaw	69.36	65.20	72.08
eyes	51.16	45.01	52.89
chin	44.00	43.88	55.76
nose	56.31	60.39	60.03
nosemouth	60.55	57.15	70.15

FIGURE 6.3: The distribution of the selected features from rgb and depth for each facial part for the three classification tasks respectively. In right the correspondng obtained accuracy for each case.

6.4.3 Diversity and Ensemble Performance

Tables 6.1, 6.2, 6.3 do not provide information about how individual patterns are diversified and has committed to influence the ensemble performance. In this section, we aim at understanding the diversity among the members of the ensemble being an important factor in classifier combination. Diversity is associated with the level of dependence among the base learners, which form the whole classification system. We first present a didactic example where the ensemble performed successful patterns to classify 16 unknown subjects from the testing set for the three classification tasks. We also provide the output labels predicted using the whole face image, and we compare them with the ground truth labels. Furthermore, we compared the performance of the ensemble by changing the ensemble size based on the greedy forward selection technique. In each iteration, a model is added to the ensemble according to its performance. We consider as performance both the balanced accuracy and the

F-score. Moreover, we compared the percentage of the misclassified samples in the test set. Figures 6.5 show the ensemble testing errors, accuracy and F-score as a function of the ensemble size for respectively, gender, ethnicity and expressions classification tasks.

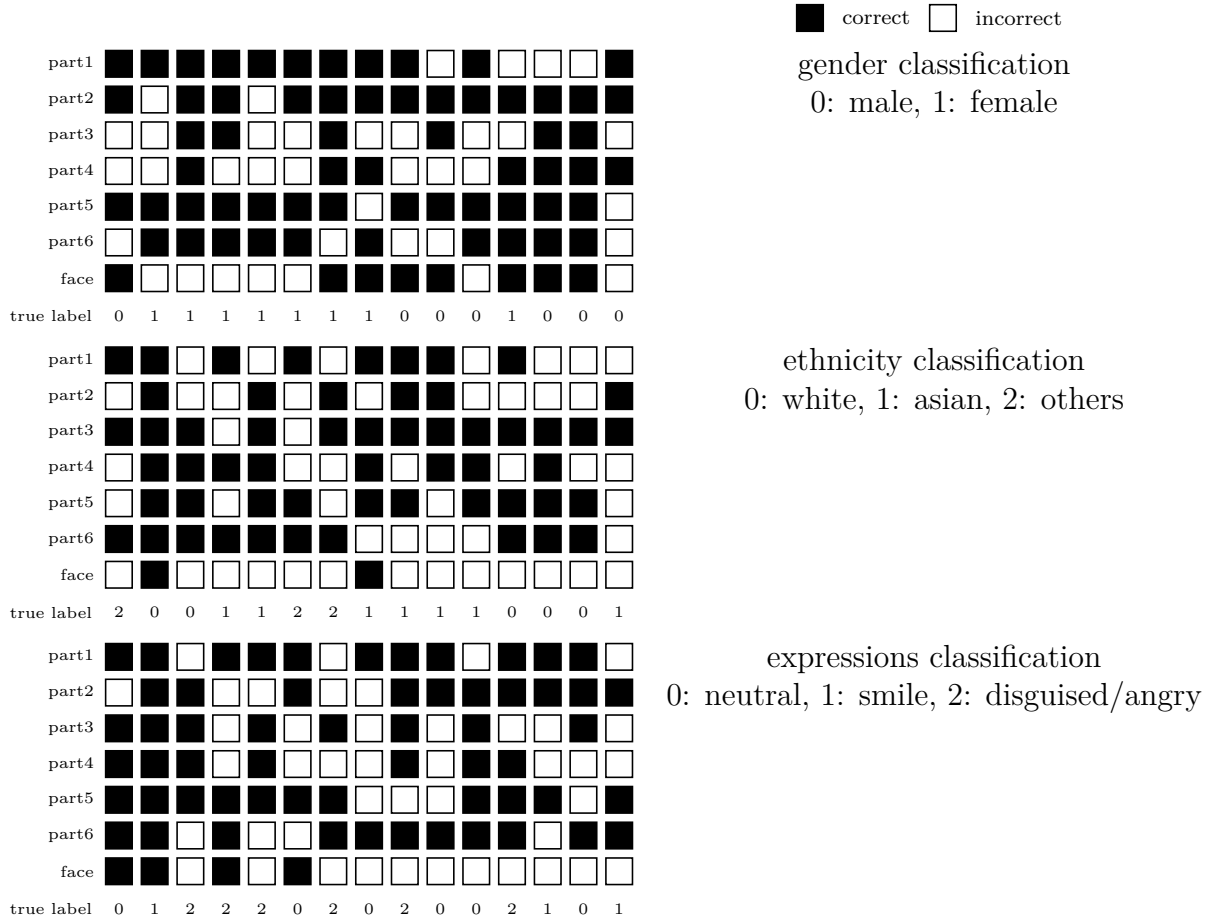


FIGURE 6.4: Ensemble diversity example

Figure 6.4 illustrates an example of the patterns that have been correctly and incorrectly classified by the individual learners and how the ensemble can produce correct prediction from their predictions. By this example, we see that the individual classifiers built from separated facial parts perform different and uncorrelated errors which shows that they are independent and not identical. Intuitively, an ensemble model is not efficient when the ensemble members are correlated. In this example, we see how the output of the learners are correlated mainly when they predict the correct output. For the other cases, uncorrelated ensemble members contributed to predicting the correct pattern. Also, even the worst-case scenario (only two individual learners are correct), the parallel approach is very positive, as the decision- making process has the advantage of

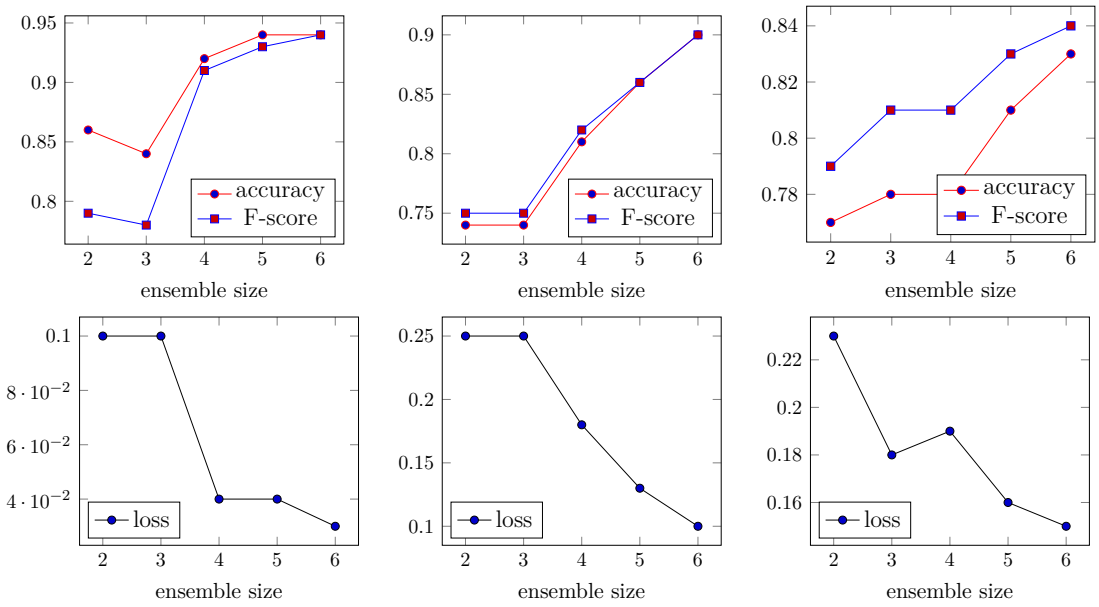


FIGURE 6.5: The performance and errors vs ensemble size for gender, ethnicity and expressions respectively from right to left

considering the score calculated before finally rejecting or making a misclassification. By this example, it is clear that when independence is assumed, we can have incorrect or correct classification independency, which contributed to a performance increase. Furthermore, compared with the outputs predicted using the whole face image, we can see that in the case of gender recognition, the incorrect predictions are mainly those belonging to the minority class which shows how learning from different errors by a variety of classifiers can help to overcome the imbalanced problem in the data.

The plots in figure 6.5 show that when the size of the ensemble increases, the overall performance of the ensemble increases and the misclassification rate decreases. In addition, we see that for all the classification cases, mainly the gender recognition task, the obtained performance between the two metrics become identical. Therefore, the parallel combination of decisions leads to a stable performance.

6.4.4 Handcraftred versus learned representations

Here, we investigate learned features, which are automatically learned from the images differently from handcrafted feature representations. Our goal is not to propose a good descriptor approach but to study the effectiveness of the proposed approach using different feature extraction method. Therefore, we simply choose to apply stacked autoencoders neural network as an unsupervised learning technique without incorporating any prior knowledge (see section 8.4.2). Table 6.4 present the obtained accuracies, f-scores and loss for each learner

and their combined decision compared with the performance of the whole face. The results are presented for the three classification cases using the combined RGB-D features.

	gender classification			**ethnicity classification**			**expressions classification**		
	accuracy	**f-score**	**loss**	**accuracy**	**f-score**	**loss**	**accuracy**	**f-score**	**loss**
p1 (cheeknose)	82.79	68.57	0.15	60.95	59.13	0.38	69.18	70.66	0.23
p2 (chin+ mouth+jaw)	90.36	77.77	0.11	64.95	64.46	0.34	65.29	64.23	0.30
p3 (eyes)	85.94	76.19	0.10	0.73	0.74	0.22	57.15	56.82	0.39
p4 (chin)	81.14	65.75	0.17	59.50	58.33	0.4	57.99	52.29	0.46
p5 (nose)	85.14	71.64	0.13	67.73	65.20	0.34	67.55	66.50	0.27
p6 (nosemouth)	79.09	63.76	0.17	65.27	65.14	0.32	71.46	71.41	0.23
face	93.51	86.15	0.06	76.34	74.76	0.24	74.17	74.95	0.21
scored decision	96.74	94.90	0.02	81.28	80.62	0.18	77.35	77.58	0.18

TABLE 6.4: Obtained results using stacked auto-encoders

Similarly to the previous results, we see from the table above that the best performance was obtained using the combined decision by individual learners achieving a stable performance and low misclassification rate. The combined performance improve remarkably the performance of the individual learner, which shows the effectiveness of the approach to exploit the diversity among them. Furthermore, it outperformed the performance obtained using the whole face image. Therefore, the results confirm our hypothesis for different used features sets.

6.5 Disscussion

From the comparative results, we conclude that the combination of multiple facial parts using the bimodal information into multiple classifier system is effective in enhancing the recognition performance. Based on the reported results in section 6.4, we can draw the following conclusions and summaries:

The fusion based on combining multiple classifiers built from local facial parts improve the bimodal performance considerably for the three classification

cases, including gender, ethnicity and expressions face classification. The improvement in performance is arising from the reduction in variance achieved using the divide and conquer strategy. Learning from the local facial parts, reduces the images noise and lead to the effective exploitation of the potential of the unimodal and therefore the bimodal information.

Furthermore, the averaged results of different base learners have demonstrated that this approach can achieve higher F-score rate and produce more stable over the two evaluation metrics. This show that the proposed ensemble strategy is much more useful than individual learning methods on imbalanced data learning, especially for the case of gender recognition, where the class distribution is highly skewed (80:20). In such case, a learner can achieve good accuracy and low loss simply due to a high bias of the classification results towards the majority class where the minority class instances might be identified as noise, discarded by the classifier and misclassified. Therefore, choosing the classifier that gets high F1 scores on both classes, as well as a high accuracy is more reliable as the F-measure emphasises both accuracies on positive and negative classes. The comparison with the face performance also shows that an ensemble approach is more robust than single classifier and more effective to deal with the imbalance problem for both binary and multi-class imbalance problems.

Another important point is that the efficiency of the multi-classifier system to solve the learning challenges and enhance performance is mainly since individual learners are independent of each other. Such independence among classifiers is known as the diversity within the ensemble. The diversity is considered one of the essential properties of a compelling ensemble. It is important to have independent and uncorrelated estimation errors; otherwise, the combined decision will not provide any improvement to the recognition process. From the experiments, we can conclude that as long as each facial part is different from the others due to the high diversity present in the human face image, dividing it into separated facial parts approach can generate several datasets which may have different distribution such that ensemble members may work better. Therefore, each classifier will realise a different model which will help to build a diversified multi-classifier system. In addition to the diversity provided from the input space, the fact that the combination approach is based on the use of discriminative criteria leads to a consistent recognition improvement over a variety of tasks. Combining the decisions by assigning discriminative scores is a good way to take into account the relative strengths and weaknesses of the participating models, and directly manipulating the class distribution of the data. Moreover, as each modality offer a specific characterisation to the face and the separated facial parts, Combining both RGB and depth features provide the classifiers with complement and informative features from both shape and texture characteristics. Hence, their performance will be higher and maximising

the diversity term. Thus, understanding facial diversity leads to building informative and diverse individual learner, so their combination allowed to effectively enhance the overall recognition capability from the individual estimators and both types of information.

To sum up, the understanding of the relationship between data issues, such as the high variance and the high imbalance ratio and learning model complexity, will be useful to provide fundamental insights and critical technical tools to overcome such issues and therefore enhancing the performance. Furthermore, the use of context and a priori information is very important in designing effective practical classification systems. Thus, in this context, understanding facial diversity is vital to justify the choice of algorithms used to exploit the used information better and build a good recognition system.

6.6 Conclusion

This part of the thesis has been focused on the analysis and development of ensembles of facial parts using the RGB-D information to solve classification problems. Throughout this research, we designed and implemented an approach using different tools from the domain of machine learning, such as ensemble classification algorithms, and feature selection methods to evaluate not only the effectiveness of ensemble learning methods but to better to exploit the different used types of information as well as to present a more in-depth insight into the underlying data. Furthermore, to alleviate issues such as difficulty-of-learning, class imbalance, high dimensionality, noisy data, and small sample size. Through this approach, it has been demonstrated that the generalisation capability of the ensemble is high than using a single classifier. The results, along with statistical analysis, show that the combination of local information outperforms using a global classifier and maximise the classification performance in most scenarios. Dividing the face into local parts reduce the presence of noise, and the high intraclass variance present mainly in RGB data and therefore exploit better the used information.

Part II

Privacy concerns: Speech inference from motion sensors

Chapter 7

Background

7.1 Introduction

Privacy is increasingly a concern in today's digitally connected world. Personnel data is being collected and stored in a number of ways and mediums such as smartphones, mobile devices, wearables, multiple social networks, internet providers services, government, etc. When data is collected, particularly personal data, it can be hard to know the purpose for which it will be used for or by whom it will be used. Personal information may be obtained in different ways: being provided by the person, observed by the person, derived from the observed/provided data or inferred from individuals through learning from such data. Thus privacy involves controlling the collection, the use and the share of any personal data that may give rise to concerns about the privacy of the individuals and making it feasible to infer private information.

Machine learning is a field of research that became a core component of many real-world applications in many domains including health, transportation, energy, education, banking, biometrics to cite a few. The increasing availability of AI-based systems which contain a rapidly growing availability of multiple sensors and the daily user interactions with the internet, give rise to multiple types of information. This diverse and large availability of data, coupled with rapid technological advances in machine learning algorithms (which learn from data), is changing society markedly. Although it enables the development of many tools with the potential of bringing good to society, their misuse might also generate or inflate risks that harm society. Wider debates are raised from the increase of information technologies, and more specifically the development of machine learning models and have warned about their ethical and social concerns [166–169].

In the second part of our thesis, we are mainly focusing on the privacy concern raised when the combination of unprotected data collection and advanced learning algorithms may lead to non-transparent inferences and manipulation of the private life of individuals, for example, their personality, intentions, recommendations, tracking, religious activities, political affiliations and so on (e.g the Cambridge Analytica scandal [61]). Moreover, by combining data sources

we can learn more than would be the case from analysing single source independently, and more accurate predictions may be inferred which increases issues about privacy or decreases the privacy guarantee.

To address these issues, we focus on a specific application around this phenomenon: Analysis of private information, mainly the speech information, that is inferred from the data provided by zero permission motion sensors built-in both mobiles and wearables, as a promising source of data, using advanced machine learning models. Mobile devices and warables are one of the commonly used technologies which attract people to improve their quality of life. These devices are equipped with rich sensors that provide an advanced and comprehensive user experience. However, it is a well known problem that the presence of numerous sensors is of major concern to the privacy of users and their social environment. We propose a deep neural network model which learns the acoustic information from two kinds of motion sensors . We investigate the obtained results to analyse to what extent private information can be derived from such personal unprotected data. Our goal is to raise awareness about the misuse of machine learning as a threat to privacy by inferring private information about the individuals speech from unprotected motion data collection as vulnerable sources. Moreover, our study shows that combining data from multiple sensors can lead to more accurate inferences and therefore increases the privacy risk.

7.2 Mobile motion sensors

In this section, we provide a background information about the investigated sensors in our work.

7.2.1 Vibration Energy Harvester (VEH)

A VEH is a transducer that converts kinetic energy from vibrations to electrical power. For low-power electronic devices in specific environments they can harvest enough energy to actually operate the device [170–172]. A VEH can be seen as having three parts: the transducer to convert the kinetic to electrical energy, a power-electronic interface, and some electrical energy storage, like a battery [173]. Common VEH transducers are piezoelectric, as this type has shown the highest potential for harvesting energy [174, 175].

Suitable vibration sources are diverse, as for example human motion, waves, wind, or vibrations of machinery. A typical piezoelectric element as used in VEHs is illustrated in figure 7.1 where one end of a cantilever beam is fixed to the device, while the other is set free to oscillate (vibrate). When the piezoelectric is affected by vibrations, an AC voltage is generated by the accumulation of positive and negative charges on the two opposing sides. The AC voltage generated in general is proportional to the applied stress.

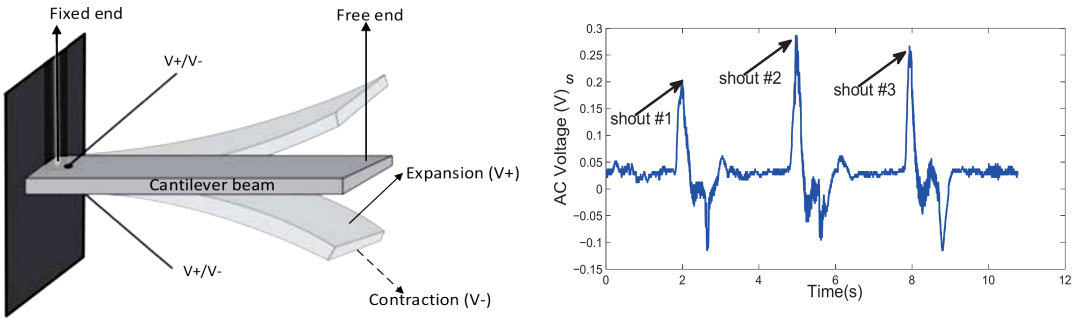


FIGURE 7.1: Piezoelectric transducer [176] (on the right) and the effect of shouting on VEH piezoelectric cantilever beam [176] (on the left)

Acoustic effect Sound waves, when emitted, are moving through air and cause pressure on the cantilever beam. Experiments [176] have demonstrated this effect by having a person shout three times while physically near to the piezoelectric part. The generated signal (see figure 7.1) shows how the VEH's voltage peaked with each shout.

7.2.2 Accelerometer

Accelerometers turn acceleration into an electrical signal based on the same operating principles as VEHs. The acceleration in different dimensions can be translated to changing positions. Raw gyroscope data consist of three values indicating the acceleration along the x-axis, y-axis, and z-axis (usually corresponding to the up-down, right-left, and front-back movement respectively).

Acoustic effect Recent work [177] showed that accelerometers are sensitive enough to draw conclusions about human speech. The authors there recorded sensor output while a speaker was spelling the vowel "A". The spectrum analysis of the output signal shows a considerable variation of the accelerometer readings during speech. They reported that the human voice has sufficient sound pressure to have detectable impact on smartphone accelerometers.

7.2.3 Gyroscope

The gyroscope is mainly used to measure the motion along an axis by measuring rotation across a given axis. This includes pitch, yaw, and roll motions. The obtained values correspond to the angular motion in terms of degrees-per-second about the x, y, and z axes[18]. In recent research works, it is observed that gyroscope is sufficiently sensitive to acoustic signals. Acoustic signal affects the gyroscope readings by vibrating the driving mass in the sensing axis (axis which senses the Coriolis force). The acoustic signal has a strong effect on gyroscope measurements when its frequency is close to the resonance frequency of vibrating mass. But acoustic signals with frequencies lower than the resonance frequency

also have the measurable effect on gyroscope readings that makes it possible to reconstruct the acoustic signal

7.3 Related works to speech inference from motion sensors

Sensor data has already received remarkable attention from the security research community, as to better understand the potential impact of this data on user privacy. Projects have investigated opportunities to identify and track users [178–181]. This was investigated in particular with the acceleration sensor [182–189]. These studies did also show that the motion sensors included in smartphones are sufficiently sensitive to allow the identification of acoustic information based on the readings induced by sound waves. An according investigation of data provided by gyroscopes was performed by [190]. This paper was inspired by works discussing such effects of acoustics on gyroscope measurements [191–193]. The authors there demonstrated by a rich experimental study that gyroscope data is sufficiently sensitive to extract information about the original audio signal. This included the identification of the speaker's gender as well an isolated hot-word.

In [177] the authors investigate accelerometer data for hot-word detection. The main motivation is to enable accurate low-energy and low-cost implementation of voice control by using the accelerometer instead of the microphone. The obtained accuracies were competitive with voice control applications such as "Google Now" and "Samsung S Voice". However, mobile operating systems limit the sampling rate (usually to 200Hz). Low sampling rates pose a hard limit on the available data and therefore are a significant challenge to speech reconstruction.

To overcome this challenge, a recent work [194] has proposed an eavesdropping attack by leveraging a distributed form of time-interleaved analog-digital-conversion to approximate a higher sampling rate. Combining the data provided by a geophone, an accelerometer, and a gyroscope they were able to reconstruct intelligible speech. A threat analysis of extracting speech signals from motion sensors of smartphones is provided by [195]. The authors there examined the presence of speech information in accelerometer and gyroscope data by studying many possible attack scenarios and analysing the behaviour of these sensors.

Furthering this track of investigations a recent publication explored vibration energy harvesters (VEH) and whether they can be used like a sensor [176]. VEHs convert physical movement into electric energy, often to extend battery life. Because of the high availability of vibration sources, VEHs are considered an effective energy harvesting option for low-power mobile devices, like for the Internet of Things (IoT). The authors there also notice that VEHs can be sensitive enough to detect hot-words in speech.

Chapter 8

Extracting speech from motion-sensitive sensors

8.1 Introduction

The increasing presence of wireless sensor networks and the blanket re-use of the resulting data volumes by AI-based systems raises pressing ethical questions about the impact of these technologies on our society. One of the commonly used technologies are Smart Phones and similar mobile communication devices which attract people to improve their quality of life. These devices are equipped with rich sensors that provide an advanced and comprehensive user experience. However, it is a well known problem that the presence of numerous sensors is of major concern to the privacy of users and their social environment. Specifically previous studies already revealed that motion-sensitive sensors actually react to human speech. In this regards Deep Neural Networks (DNN) proved very successful to model high-level abstractions in data. Our main focus is highlighting (i) the potential risks related to these sensors leaking private information about speech and (ii) the ethical implications of advances in (deep) machine learning as a threat to privacy. In this work we showcase a simple attack in which collected data from accelerometer and Vibration Energy Harvester (VEH) sensors can be used to eavesdrop on speech. We propose a multistage stacked auto-encoder model that learns the distinctive time and frequency characteristics independently without user interaction. We demonstrate the efficiency of our model with poor quality data and a very low sampling rate. We investigated three classification tasks: gender identification (i), hotwords detection (ii), and (iii) recognition of simple phrases selected from a previously well investigated dataset.Our experiments demonstrate the efficiency of our model and confirm that motion-sensitive sensors are a rich source of personal data, from which highly sensitive and private information about people in close proximity to the sensor emerges.

8.2 Motivation

Mobile devices are often equipped with sensors to provide services based on the according sensor readings, like location, movement, temperature, and alike. The data from such sensors, however, can also be used for other purposes. Here we investigate how data created by a movement sensor and an energy harvesting component can be used to extract information about human voice communication.

Advances in machine learning, particularly deep neural networks (DNN), achieved remarkable success in complex classification tasks and pattern recognition problems [196–198]. In comparison with handcrafted methods, deep neural networks have been successfully to learn high-level features from complex patterns in data [32]. However, learning the acoustic information from the motion data remains a complex and difficult challenge due to many factors. In fact, the data obtained from these sensors are correlated with multiple sources which leads to challenges in extracting useful features and increases the level of uncertainty. A commonly used approach to overcome this problem is to combine data from multiple sources of information. Compared with that of single sensor data, multi-source information fusion often contain redundant and complementary information and lead to more reliable and accurate information representations. Basically, the fusion of multiple modalities is performed using two levels of fusion. Namely early fusion (feature-based) which consists of combining the features before the learning process and late fusion (decision-based) which performs integration after each of the modalities has lead to a decision. In the latter the decisions are combined using what is called a decision fusion [199]. DNN based models have been proposed to learn from multimodal data whereas auto-encoders have shown a high efficiency in learning abstract joint representation from multiple sources of information [39, 200, 201]. As mobile devices are often equipped with more than one sensor to acquire a variety of possible information, multimodal deep learning approaches thus can be used by an attacker to eavesdrop on the contents of human voice communication [202, 203].

In this work we propose to use DNNs on sensor data to increase the accuracy in recognising voice patterns. We want to determine the potential risks related to privacy when such data is not protected. Here we focus on the data collected from a VEH and an accelerometer while users were speaking as collected in a preceding study by [176].

Our intentions are twofold: Firstly we want to highlight the improved ability of extracting acoustic patterns from sensors not primarily used for acoustic information by using DNNs. Secondly we want to explore if the combination of data from different sensors can significantly improve the detection rates. Our aim is to raise awareness about the dangers related to privacy when unwanted data collection through zero permission sensors data is combined with the intransparent classification capabilities of deep learning models.

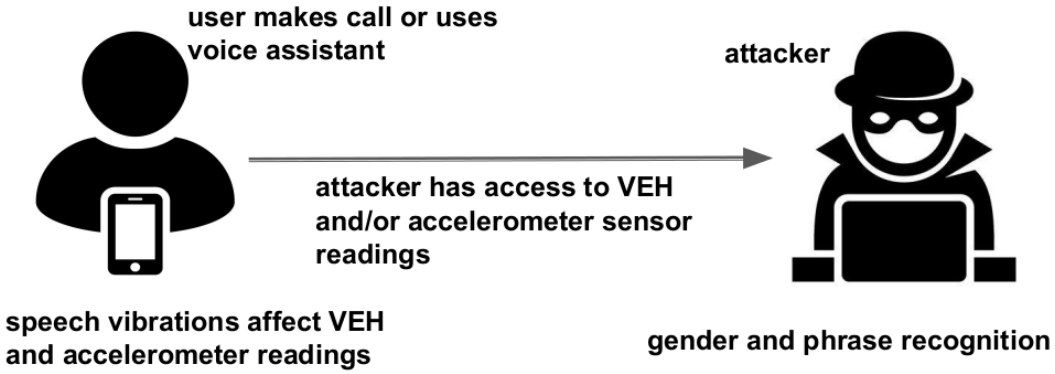


FIGURE 8.1: Example for an attack scenario

We perform an experiment by using a DNN based on a stacked auto-encoder upon the collected data. Our model extracts acoustic features by exploring both time and frequency representations. The extracted features are then used in supervised classification to identify the speaker's gender, detect a simple hot-word, and distinguish it from short sample phrases.

We show how the combination of the data provided by the VEH with the data of the acceleration sensor significantly improves the recognition rates in comparison to [176]. To that end we train the DNN with both sources.

To the best of our knowledge the effect of speech on motion sensors has so far only been performed using manual feature extraction techniques or only one way sensor data [176, 177, 190, 195, 204]. We, therefore, assume our approach of using deep neural networks in this context is novel.

8.3 Threat model

When sensor readings expose voice communication additional threats to privacy become apparent. The threat model changes as an attacker then only needs access to sensor readings instead of the microphone directly. The attack vector thereby is extended to any application having access to the readings of relevant sensors, as for example on the user's device in figure 8.1.

The attacker here can identify acoustic patterns in the accelerometer and VEH readings. With the sensors used in the experiment the user needs to be physically close to the device [176]. However, this seems very likely when using a mobile phone or a smart watch, which also happen to be the devices where accelerometer and VEHs are (to be) used. Furthermore the attacker does not necessarily require direct access to the sensor data. We assume that access to locally cached sensor readings may be sufficient to allow offline attacks.

We consider two scenarios in our study: In the first scenario, the attacker only has access to the data of one source. With the available data we can

compare having access to only the accelerometer or only the VEH. We determine how well the DNN identifies the speaker's gender, detect the hot-word, and identify short sample phrases for each of those data sets. In the second scenario, we assume the attacker to have access to both, the accelerometer and the VEH. Here we combine the data sets of the accelerometer and the VEH to train the DNN. Again we determine how well the DNN identifies gender, hot-words, and phrases, so that we can compare these results with those derived by using a single data source.

This chapter is organised as follows. Section 8.4 explains our approach in detail and background information about the used the auto-encoder based model. In section 8.5 we present the results of our experiments. We disscuss our findings in section 8.7. Finally, the conclusion is provided in section 8.8.

8.4 Learning Acoustic Information

Here we present an overview of our proposed approach and discuss details, before presenting the experiments in the next chapter.

8.4.1 Classification task

Classification tasks are tied into information representation. The learning process on how to represent data is a critical step that on the one hand should preserve as much information as possible from the input data. On the other hand this process should eliminate redundancies to foster the extraction of structures and properties. Motion sensors are designed to respond to movement. Their output signals originate from physical movement of an accordingly designed part of the sensor. They are particularly used for tasks related to motion recognition, such as identifying physical activity. Modelling motion sensor data to perform sound recognition is an especially challenging task, because the sensor‘s sensitivity is optimised towards such movements and not sound.

The artificial learner requires preprocessed data in form of features to learn from. A feature is a measurable property fed to the learning algorithm. These are normally manually extracted relying on knowledge of a human expert. This expertise is domain- and/or sensor-specific and is required for each new dataset or sensor modality in order to engineer the suitable features for a specific application. Therefore, the use of manual feature extraction is very limited. Furthermore it cannot be generalised across different application domains.

As accelerometer and VEH sensors are of a non-acoustic nature, we do not benefit from a prior knowledge about useful measurements to apply in order to extract the acoustic information. To deal with this issue, we propose an unsupervised deep learning approach to automatically learn suitable features without relying on hand-crafted features. Here features are automatically extracted from data through layers were each successive layer acts as a feature

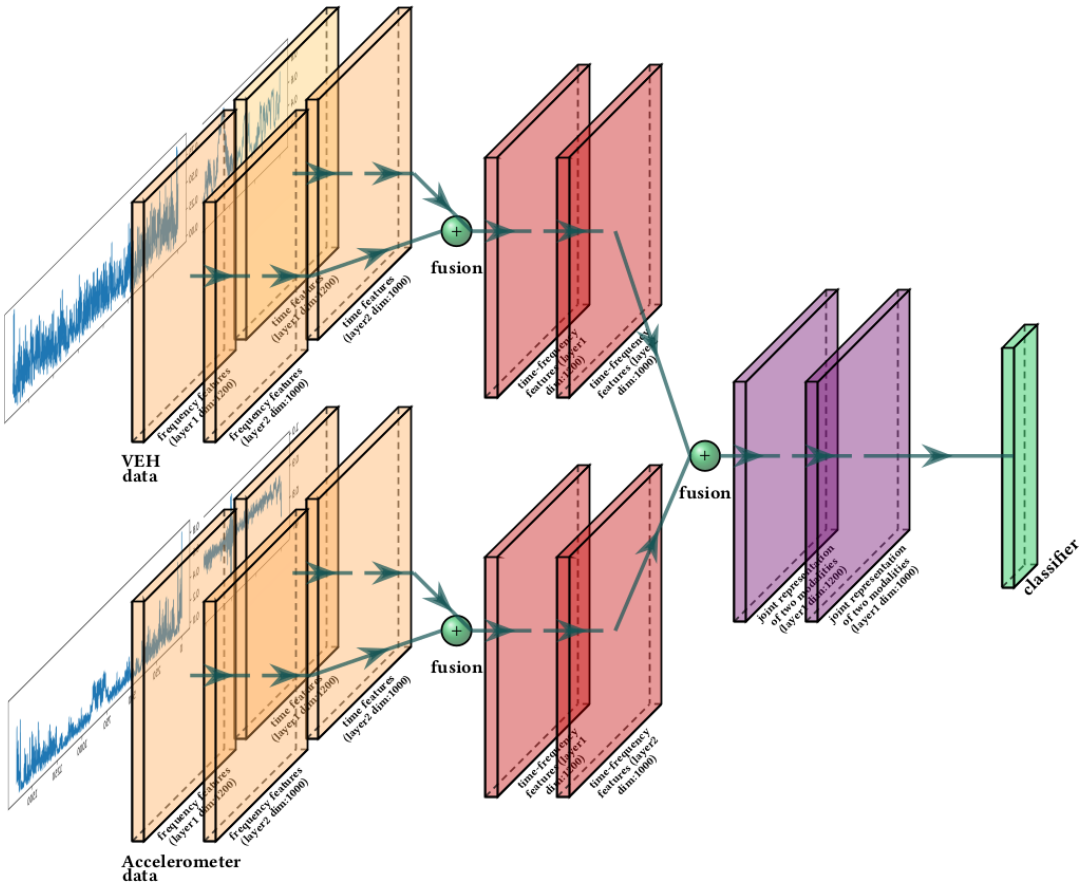


FIGURE 8.2: The figure shows the whole network architecture. The two inputs are the frequency and time representation of the VEH and the accelerometer data. Each layer is the hidden layer encoded using the autoencoder. The layers are stacked using layer wise training strategy. The last layer represents the classifcation step performed after the unsupervised features learning from previous layers.

extractor and is hypothesized to represent the data in a more abstract way. This process is unsupervised, which means that it is independent to a specific classification task. To this end we propose a Stacked Auto-Encoder (SAE) based model to discover relevant complex structures underlying speech and to learn a deep and high-level representation robust to intra-class variability including the sensor direction and the speaker speaking.

Abackground information about an Autoencoder neural network is presented in section 8.4.2. An architectural overview of our approach is illustrated in figure 8.2. It is divided into two main phases: unsupervised feature learning (section 8.4.3), where we added the combination of the data-sources as well (section 8.4.4), and supervised classification (section 8.4.5).

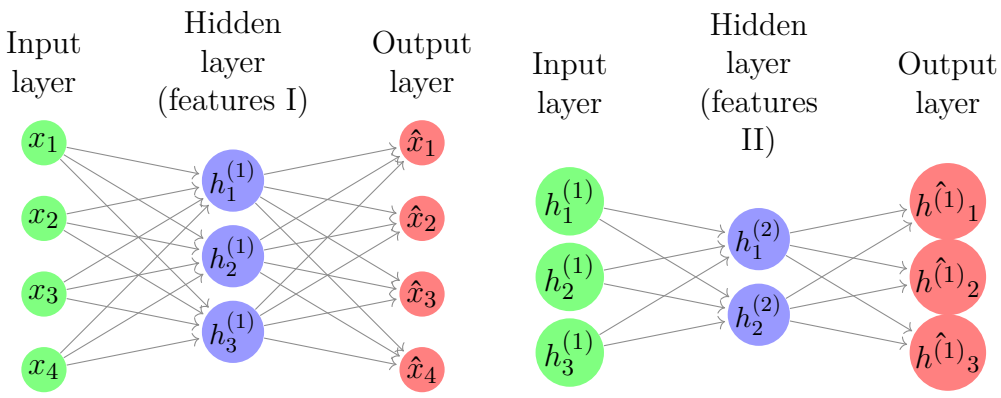


FIGURE 8.3: Typical architecture of autoencoder with one hidden layer; to obtain deep AE, the hidden layer of first (above) is used as an input to second (below) based on the Greedy layer-wise training

8.4.2 Stacked Auto-Encoders

An auto-encoder (AE) is a type of artificial neural network (ANN) for unsupervised learning that applies the back-propagation learning algorithm. Back-propagation is the standard learning algorithm for the training of feed-forward neural networks as used for supervised learning. Feed-forward neural networks are very simple neural networks with an input layer, an intermediate hidden layer, and an output layer. The neurons in the input layer forward some representation of the input data to the hidden layer of the neural network. The hidden layer transforms that input data for subsequent layers using some learned function. The output layer is the final layer with the scaled target values. The learning objective of the AE is to map the data of the input layer to the output layer in the way it is desired. Enforced limitations of the neural network structure complicate training for the learning algorithm. The result is an approximation of the so-called identity function, where the output is a representation of the input. The architecture of an AE divides the ANN into an encoder and a decoder. The encoder takes the data at the input neurons and creates a "restricted" representation of it at the hidden layer. Since the hidden layer is smaller than the input layer it learns only the most relevant aspects of the input. The decoder then tries to reconstruct the original input from the representation in the hidden layer. This produces a higher-level representation from the lower-level representation of the input [205].

A stacked auto-encoder (SAE) is an ANN consisting of multiple hidden layers creating a deep neural network architecture. SAEs overcome limitations preventing deep architectures of multi-layer feed-forward neural network trained with back-propagation. The SAE applies the so-called greedy layer-wise pre-training strategy (Figure 8.3) which addresses the error-causing vanishing gradient problem. In SAEs the input layer is the encoded layer trained on the

raw input. The output then is used as input to the next AE to obtain the next encoded layer and this process is repeated for subsequent layers. Stacking layers like this can then lead to deep stacked auto-encoders that carry some of the interesting properties of deep models [205].

8.4.3 Feature learning

In this step, we investigate time and frequency data separately to train a bimodal representation from each sensor. We perform Fourier transformations on the frequency data. Then we use the greedy layer-wise training for the SAE. Therein, the features learned in a hidden layer are used as input to the next AE in order to produce a new representation of the data. By representing the data through layers we enable learning of complex patterns across data variations. After extracting the features separately from each source, that is time and frequency, we combine them into a joined time-frequency representation. This joint representation leads to a shallow model, thereby making it difficult for a single hidden layer model to directly find correlations between representations that have been joined. We, therefore, again apply greedy layer-wise training to improve discovery of high-level correlations across the two representations.

8.4.4 Data From Multiple Sources

With the assumed availability of different sensors, we then have separate types of data sources about a given moment. The machine learning community assumes potential in improved learning algorithms to specifically exploit such multi-modal data to form a unified picture [206]. Modelling speech recognition from data of non-acoustic sensors is challenging. Additional problems to tackle are the limited sampling frequency and the interference from the device's original function (detecting movement, harvesting energy) with our intended function (detecting spoken language). Our study specifically aims to determine to what extent combining data provided by different sensors can provide improved results. To that end, our multi-layer approach combines separately trained models into a joint representation.

8.4.5 Supervised classification

We then use supervised classification on the extracted features. For this the fused representation functions as the input, thus providing features across the original data sources. We repeat the three classification tasks for the speaker's gender identification, the hot-word detection, and the recognition of the sample phrases.

8.5 Experimental study

In this section we are describing our experiments in more detail. We start by presenting the available data and then explain how we pre-processed it and performed the feature learning and classification.

8.5.1 Dataset

The dataset we used is described in more detail in [176]. It is the only work we are aware of that already studied the potential of detecting acoustic information from VEH data. It contains the data for both, a VEH and an accelerometer, while different persons performed identical tasks repeatedly. Involved were eight individuals, four being male and four female, and the experiments were performed with two different orientations of the devices (horizontal and vertical). The devices were positioned close to the persons (3 cm) and the experiments repeated 30 times for the hot-word "Ok Google" and at least ten times for the phrases "Good morning", "how are you", and "fine thank you". Overall the data-set contains 1155 samples. Figure 8.5, 8.4, represents the accelerometer and VEH sensor outputs while a person spoke the four phrases.

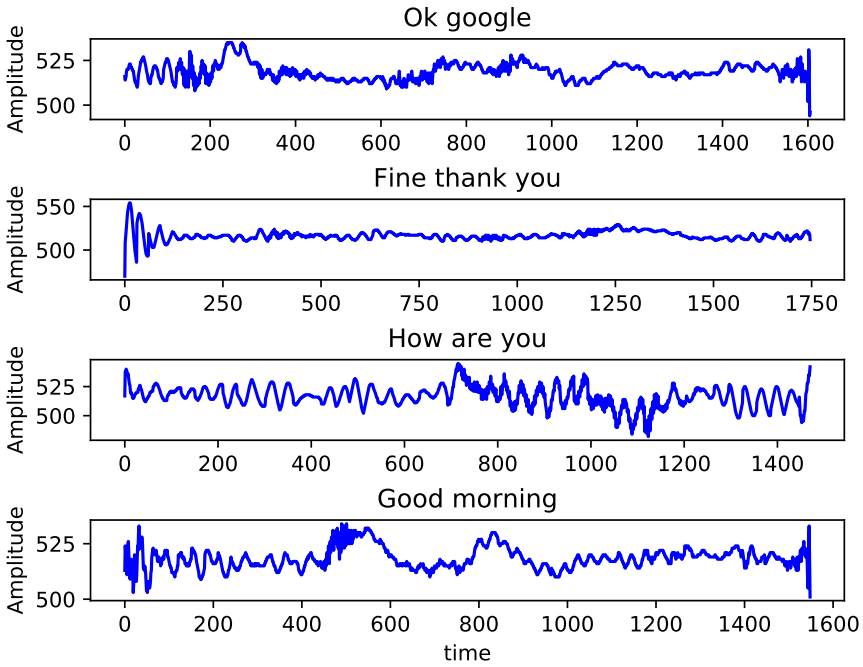


FIGURE 8.4: The VEH signal while the user is speaking the four phrases ("good morning", "okay google", "fine thank you" and "how are you")

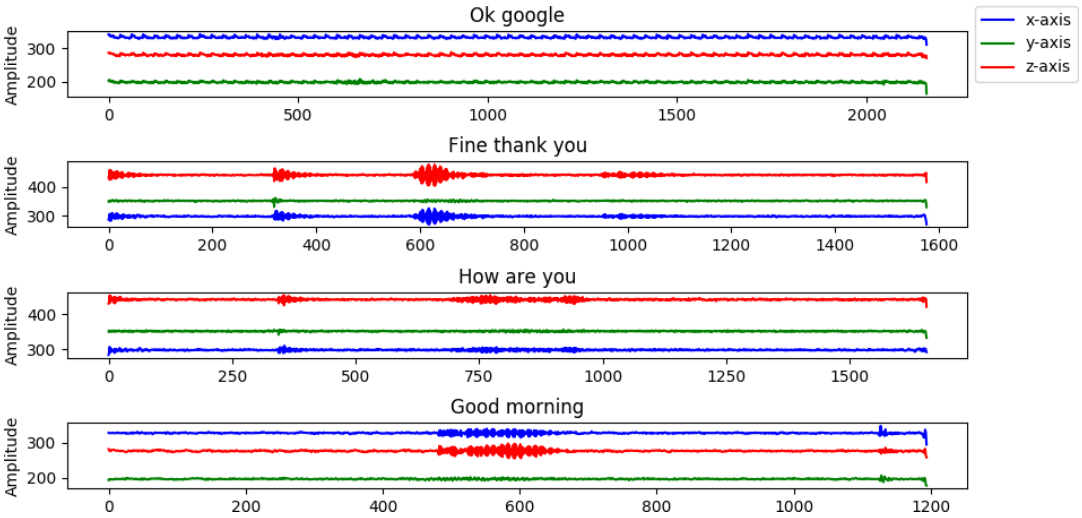


FIGURE 8.5: The accelerometer outputs (x axis,y-axis, z-axis on left) while the user is speaking the four phrases ("good morning", "okay google", "fine thank you" and "how are you")

8.5.2 Preprocessing

In the pre-processing we apply our domain knowledge to address the specifics of the different data sources.

VEH: As we face varying lengths of the samples, we started by interpolating all the samples in the data to the mean length. We separately handle the temporal and frequency representations of the VEH signal. To minimize the signal-to-noise ratio we normalise and filter the temporal representation. On the frequency representation we apply a fourier transformation to obtain the frequency values. Since they are complex valued, we calculate the real part which corresponds to the magnitude of the amplitudes, which we then also normalised.

Accelerometer: For the accelerometer data, we down-sample the signal to 200Hz. This is the limit on sampling frequency as posed by the mobile operating systems Android and iOS. We then interpolate the samples to their mean length and normalize the data. The three acceleration channels were combined as one using square summing to obtain the magnitude acceleration, which is orientation independent.

8.5.3 Feature Learning and Classification

The training procedure for time and frequency representations each is executed for 50 epochs, using a mini-batch size of 30 and learning rate of 0.001. The RMSprop variant of the stochastic gradient descent is used as the optimization algorithm. For the feature learning step, we used a hidden layer of size 1200. For the classification we used and compared three common learning algorithms, that

is Support Vector Machine (SVM), K-nearest neighbours (KNN) and Neural Network Classifier (NN), in order to study their performance.

8.5.4 Evaluation

In the evaluation we used a k-fold cross-validation with $k = 10$. For this we divided the data into k equal folds (portions). We then trained the model on the $k-1$ folds and test it against the remaining folds. That process was repeated $k = 10$ times. The final performance after that corresponds to the average of the obtained values. We used cross-validation analysis to ensure that all data was used for both, training and test. The classification of the proposed framework was performed using the four metrics accuracy, precision, recall, and F-measure.

8.6 Results and discussion

In this section we present the results from our experiments in detail.

8.6.1 Single modal performance

We first evaluate the results from using the data of the accelerometer or VEH on their own, each. Tables 8.1, 8.2 present the classification performance for each of our metrics, that is accuracy (acc), precision (prec), recall (rec) and F-measure (f-score), as determined for each of the classifications (gender identification, and hot-word detection, and phrase recognition) for each of the used learning algorithms KNN, SVM, and NN (alg) for each of the data representations time-only, frequency-only, and our model. These allow us to see how our model compares to using only the time- or only the frequency-data. On the hot-word classification the KNN and SVM algorithms with our model both achieved an accuracy of 75% when used on the VEH data and 76% when used on the accelerometer data, in all cases out-performing the use of only time- or frequency representations. For gender identification results, the best classification performance was achieved by our model in combination with SVM, having an accuracy of 86% using the accelerometer data and near 80% using the VEH data, again out-performing the use of only one data-representation. The accuracy of our model in recognizing the phrases was in the range of 64-65% for all combinations but using NN on the accelerometer data and once more out-performed the use of single data representations in all combinations. The F-score shows comparable values. Therefore, we can conclude that features learned from the joint representation of time and frequency information leads to a considerable improvement of the classifications. Both, frequency and time representations, contain important information that can be combined for better results here. We assume that more abstract features have been learned in the

process. We highlight that sufficient information about the original activity of shouting can already be extracted with relevant accuracy even when using only one of the data sources.

		Hot-word detection				**Gender identification**				**Sentences recognition**			
alg.	**rep**	**acc**	**prec**	**rec**	**f-score**	**acc**	**prec**	**rec**	**f-score**	**acc**	**prec**	**rec**	**f-score**
KNN	time	74	76	74	74	71	72	71	70	62	62	62	60
	freq	69	69	69	69	76	77	76	76	55	55	55	53
	our model	75	76	75	75	79	80	79	79	64	65	64	62
SVM	time	71	72	71	71	70	71	70	69	63	64	63	61
	freq	69	70	69	69	76	77	76	76	54	54	54	53
	our model	75	76	75	75	80	80	80	80	64	65	64	64
NN	time	70	72	70	70	62	64	62	61	61	62	61	60
	freq	65	68	65	63	72	74	72	72	54	54	54	51
	our model	75	76	75	75	77	79	77	77	65	65	65	64

TABLE 8.1: The obtained results (%) using the VEH data for the obtained time features, frequency features and the joint time-frequency representation (our model).

8.6.2 Bimodal performance

Next we examine if access to multiple data sources further increases the classifications. For this we repeated the training using both, the VEH and the accelermeter data, as described above. The results are — formatted as the previous tables — shown in table 8.3. Combining the data-sources has significantly increased the accuracy of the classifications across the board by around 10%. The highest F-scores of 91%, 85%, and 77% for gender identification, hot-words detection and recognition of phrases respectively, were achieved when using the SVM classifier with our model. The increase was higher for our model than if using only the time or the frequency representation. In each of its levels, the ANN must have learned additional correlations between the data variables across frequency and time representations. Overall, the joint representation has lead to remarkably improved accuracy. We then performed the k-fold cross-validation (figure 8.6), to ensure that our model was trained and tested on representative samples. We considered five k values ($k = 4, 5, 7, 9, 10$). The

		Hot-word detection				Gender identification				Sentences recognition			
alg.	**rep**	**acc**	**prec**	**rec**	**f-score**	**acc**	**prec**	**rec**	**f-score**	**acc**	**prec**	**rec**	**f-score**
KNN	time	71	72	71	71	83	83	83	83	58	59	58	56
	freq	55	55	55	55	64	65	64	63	42	38	42	37
	our model	77	77	77	77	83	83	83	83	65	65	65	63
SVM	time	58	60	58	54	80	80	80	80	48	31	48	33
	freq	58	58	58	58	71	72	71	71	41	39	41	39
	our model	76	76	76	76	86	86	86	86	65	65	65	64
NN	time	53	55	53	48	61	65	61	58	47	23	47	30
	freq	54	56	54	50	66	68	66	65	45	35	45	34
	our model	68	70	68	67	80	81	80	79	54	57	54	50

TABLE 8.2: The obtained results (%) using the accelerometer data for the obtained time features, frequency features and the joint time-frequency representation (our model).

results show, that the accuracy of our model does not decrease on repeated validation runs, and, therefore, was not just a lucky pick of matching data.

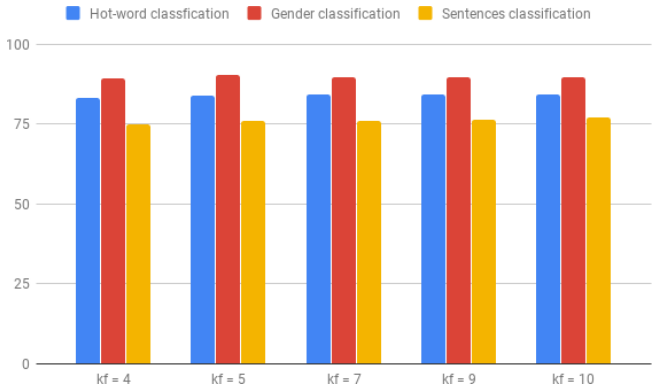


FIGURE 8.6: Obtained accuracies (%) combining VEH and accelerometer data with repeated k folds using the SVM classifier

In table 8.4 we compare our model with the results from the original work from [176] which only used the VEH data. The authors there compared results for different positions of the VEH. Recognising the importance of positioning, we specifically wanted the DNN to cope with this, as we hardly can influence the positioning in the scenario of spying. This way our results should be better

		Hot-word detection				Gender identification				Sentences recognition			
alg.	**rep**	**acc**	**prec**	**rec**	**f-score**	**acc**	**prec**	**rec**	**f-score**	**acc**	**prec**	**rec**	**f-score**
KNN	time	78	79	78	78	81	81	81	81	69	69	69	67
	freq	69	70	69	69	79	80	79	79	55	55	55	52
	our model	83	83	83	83	88	88	88	88	73	74	73	72
SVM	time	76	76	76	76	79	79	79	79	66	66	66	64
	freq	73	73	73	73	81	82	81	81	59	59	59	58
	our model	86	86	86	86	91	92	91	91	77	78	77	77
NN	time	68	69	68	67	73	74	73	73	61	64	61	59
	freq	71	72	71	71	79	80	79	79	59	59	59	57
	our model	81	82	81	80	88	89	88	88	74	75	74	74

TABLE 8.3: The obtained results (%) combining the VEH and the Accelerometer data for the each representation (time, frequency and their combination) acheiving the highest performance compared to one modality.

suited to assess the practicality of an according attack vector. Moreover, we applied our model to the gyroscope data used by [190] for isolated words recognition. We compare our results with those obtained by the authors for the user independent case. The results show that our model provided better accuracy than the state of the art works, that were based on manual features.

The results show that our deep auto-encoder approach can improve the recognition of acoustic patterns from non-acoustic sensors, here acceleration and voltage readings. We do not claim that the proposed approach represents a direct substantial risk to privacy, yet, as the data-set is small and was derived in a very specific setting. However, mobile sensors beyond the obvious microphone and camera could become targeted by attackers as they — as of today — are often less protected. Despite the fact that the data provided by such sensors is always cluttered due to their main purpose, it still is possible to draw conclusions on audio information from them by using two main approaches. The first is to combine data from multiple sources by considering a multimodal architecture. This will exploit the complementary between multiple modalities (information obtained from multiple sensors) and will lead to more accurate results. The second is to include a de-noising component in the autoencoder. In fact, de-noising data is one of the areas where auto-encoders have been most successful [207].

method	**accuracy**
Hot-words detection using VEH in a horizontal position [176]	73%
Hot-words detection using VEH in a vertical position [176]	63%
Hot-words detection using VEH invariant to sensor orientation (our model)	75%
Hot-words detection combining VEH with accelerometer (our model)	**86%**
Isolated words recognition (11 words) using gyroscope with SVM (Speaker-independent) [190]	10%
Isolated words recognition (11 words) using gyroscope (our model) Speaker-independent case	**25%**

TABLE 8.4: Comparison with state of the art methods

Considering that such sensors might generally not be considered as sensible and therefore be less protected, they could rise to become popular attack surfaces in the future. Based on our results an attacker with access to the readings of multiple sensors must be regarded dangerous, even if the primary function of the sensors seem harmless at first. With today's mobile devices many people already carry a multitude of sensors around and the trend seems to be for *even more*. Based on our experiment, it seems clear to us that uncontrolled access to these sensors imposes serious security and privacy risks to the users.

8.7 Disscussion

The findings compiled in this work show that motion data are a rich source of personal data. The misuse of such data using learning algorithms that can extract visible and invisible patterns and correlations from it can lead to leakage of sensitive information such as the individual's speech. Furthermore, by combining different data sources we can learn more than from one source independently. Thus, more accurate information can be inferred from a combined analysis (in the studied case the accuracy has increased by 10%), which increases the risk of the privacy breach.

Recognizing the speech does not give only information about what a speaker says, but also its attitude toward the listener and the topic under discussion, and the speaker own current state of mind as well. Many works have discussed the inferences that can be drawn from human speech extracted from audio data [208]. Such inferences include information about identity, body features, gender,

age, personality traits, mental and physical health condition, emotions, origin, and socio-economic status [209–214].

Therefore, the inferred speech information from motion sensors, in turn, can be used to deduce more non-transparent insights about individuals and manipulation about their private life [167, 169, 215]. Several examples of data breaches where personal information was exploited for political purposes [61], financial ones [62], and other purposes [216]. Therefore, the misuse of such data can seriously affect an individual's relationships, employability, or financial status, or lead to negative consequences for essential rights and social values such as freedom of expression, respect for private life. In addition, the violation of privacy can be extended from individual privacy risks to groups privacy violation. It can be easily applied to large numbers of individuals such as political collectives, ethnic and religious groupings, group in commercial companies, and governmental institutions [217].

The privacy threat of unexpected inferences from unprotected data sources are not limited to those discussed in this work. The problem of undesired inferences goes far beyond motion sensors and the deduced insights are related to the samples present in the used dataset. Thus, a larger database will contain a variety of characteristics from personal attributes in addition to the gender and identity or age. Such attributes may include, emotions, personality traits, sexual orientation, ethnicity, religious and political views to cite a few. This diversity of data will allow discovering other correlations, and obtaining more analysis. Therefore, given appropriate training data, more private insights may be derived mainly with the increasing evolution in machine learning algorithms, which offer complex mechanisms for predicting future values and even predicting causal correlations[218–220].

In sum, the aim of the work is mainly to raise awareness about the ethical and privacy implications of the advancement of learning algorithms coupled with the growing availability of data. This is achieved by demonstrating how machine learning can be used as a tool for privacy breach and manipulation. Advances in technology change how personal information is collected and analysed, and therefore create new privacy risks. Thus the continued debate is needed to guide the development not only of technology but also of the policies that enable its use. And governments need to be more serious about finding a solution to limit the power that larger companies have over citizens.

8.8 Conclusion

In this work, we investigate the technical feasibility of speech inference from motion data using advanced machine learning models. we explore how non-acoustic sensor readouts can be used in uni-/multi-modal attacks. We propose a multi-level time-frequency based deep neural network to extract acoustic patterns from an accelerometer sensor and an energy harvesting component. Our

model detects gender, single hot-words and spoken phrases with an accuracy of up to 91%, 85%, and 77% respectively. This findings show that motion sensors data are a rich source of personal data. They can be sufficient to obtain information about a device holder's speech especially when used data is combined from multiple sensors. By combining data sources we can learn more than would be the case from analysing single source independently, and more accurate predictions may be inferred which increases issues about privacy or decreases the privacy guarantee. An attacker with access to an accelerometer and some sensitive energy harvesting module is able to eavesdrop on human speech and draw conclusions about its content. Therefore, they could be considered private data in the same sense as audio data.

The privacy of mobile device and wearables users, is a concern of growing importance. The zero permission nature of embedded motion sensors make acquiring the data easier. The collection of such data combined with a misuse of machine learning algorithms (which learn from data) can lead to a serious privacy risks and leakage of sensitive inferences about the user including his speech. The problem of undesired inferences goes far motion sensors data and needs to be addressed for other data sources as well. Therefore, further research is required into the privacy implications of unprotected data collection taking into account the evolving state of the art in machine learning algorithms. Furthermore, a continued debate is needed not only about control over all sensor data, but also to guide the development of technology and of the policies that enable its use.

Chapter 9

Conclusion and perspectives

This chapter provides a summary and conclusion. Recommendations for future research are also given.

9.1 Summary

In this thesis, various experiments were carried out to investigate machine learning solutions for learning classification problems using the information provided by multiple sources. The goal is first to study how to exploit multiple types of information as an opportunity provided by the affordable sources of information, second to study the benefits versus the implications of such an opportunity. The thesis is divided into two main parts; each one corresponds to a specific application.

In part I, an ensemble learning-based model was proposed to exploit the potential of the RGB-D information provided by the Kinect sensor for solving face classification related tasks. Machine learning techniques were proposed to tackle pattern recognition related issues to the data and the model. Such issues include the noise, class imbalance, small sample size and high intra-class variance. Various experiments were performed to understand how different techniques can be used to solve such issues. For better evaluation, the proposed model used various performance metrics to evaluate the model performance such as the ROC curves, the balanced accuracy and F-measure. The results show the effectiveness of the proposed approach. They also show that the understanding of the relationship between data issues and learning model complexity is crucial to propose suitable solutions and to enhance the system performance. In addition, the right formulation and understanding of the problem are required to design accountable and transparent models. Also, to avoid ethical concerns related to the misinformation and discrimination that can result from a bad model design.

In part II, we investigated the privacy problems of mobile devices users raised by the misuse of machine learning and personal data. In particular, we proposed an attack scenario using motion sensors data to infer private user information such as his speech. We designed an unsupervised feature learning model that can learn useful features and reduce both overfitting and underfitting. We then

studied various classification problems and discussed the implications of the obtained results. The results show the efficiency of our model. They also show that the data provided by these sensors can be a rich source of private and sensitive information through unintended inferences, especially when multiple sources are combined. These inferences are not only limited to the information obtained from raw data, but also all types of derived information from them. Given these risks, calls for accountable and responsible use of data and AI technologies become with crucial importance. Collaborative work is required from both machine learning and cybersecurity communities. First, to control data sources and second to build accountable, robust and privacy-preserving machine learning models. Furthermore, additional law restrictions are needed on controlling not only personal data collection but also data processing and data subjects.

9.2 Future work

Research and experiments conducted in this study suggest that many more directions can be identified. Mainly exploring more challenges related to learning the complement information from multiple sources, and investigating probabilistic models to deal with uncertainty in data. Moreover, we consider the contents of the last work to be early work on this topic. An interesting next step would be to examine the attack vector under real-world conditions. For further experiments, a larger annotated dataset including more sensors as found in smartphones and a possibly large set of recorded situations would be needed. Only then would it be possible to realistically judge the threat that, for example, is posed by smartphones today, when installed applications are allowed access to sensors without care. Many factors usually do affect sensors and different sensors each have their specifics in how they are affected. This provides a large variety of possible experiments from recording and annotating data to performing analyses on that data then. Concerning the neural network, it might be beneficial to use recurrent neural networks with long-short-term-memory to capture the temporal relationships in the data. Also working on privacy-preserving machine learning will be an interesting direction to help to build robust and trusted machine learning models against privacy violations and data breaches.

Bibliography

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[2] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[4] Tom Mitchell. Introduction to machine learning. *Machine Learning*, 7: 2–5, 1997.

[5] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[6] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[7] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[8] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[9] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[10] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019.

[11] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[12] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[13] Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, 47(3):1–36, 2015.

[14] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

[15] Emilie Morvant, Amaury Habrard, and Stéphane Ayache. Majority vote of diverse classifiers for late fusion. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 153–162. Springer, 2014.

[16] Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing*, 3(2):211–223, 2011.

[17] Alex Pappachen James and Belur V Dasarathy. Medical image fusion: A survey of the state of the art. *Information fusion*, 19:4–19, 2014.

[18] Zhen-Zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G Hauptmann. Multimedia classification and event detection using double fusion. *Multimedia tools and applications*, 71(1):333–347, 2014.

[19] Arun A Ross and Rohin Govindarajan. Feature level fusion of hand and face biometrics. In *Biometric technology for human identification II*, volume 5779, pages 196–204. International Society for Optics and Photonics, 2005.

[20] Xiao-Na Xu, Zhi-Chun Mu, and Li Yuan. Feature-level fusion method based on kfda for multimodal recognition fusing ear and profile face. In *2007 International Conference on Wavelet Analysis and Pattern Recognition*, volume 3, pages 1306–1310. IEEE, 2007.

[21] Mohammad Haghighat, Mohamed Abdel-Mottaleb, and Wadee Alhalabi. Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition. *IEEE Transactions on Information Forensics and Security*, 11(9):1984–1996, 2016.

[22] Juan E Tapia and Claudio A Perez. Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of lbp, intensity, and shape. *IEEE transactions on information forensics and security*, 8(3):488–499, 2013.

[23] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, 2005.

[24] Ekaterina Shutova, Douwe Kiela, and Jean Maillard. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, 2016.

[25] Arun Ross and Anil Jain. Information fusion in biometrics. *Pattern recognition letters*, 24(13):2115–2125, 2003.

[26] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, 2013.

[27] Giridharan Iyengar, Harriet J Nock, and Chalapathy Neti. Audio-visual synchrony for detection of monologues in video archives. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 5, pages V–772. IEEE, 2003.

[28] Michael Glodek, Stephan Tschechne, Georg Layher, Martin Schels, Tobias Brosch, Stefan Scherer, Markus Kächele, Miriam Schmidt, Heiko Neumann, Günther Palm, et al. Multiple classifier systems for the classification of audio-visual emotional states. In *International Conference on Affective Computing and Intelligent Interaction*, pages 359–368. Springer, 2011.

[29] Geovany A Ramirez, Tadas Baltrušaitis, and Louis-Philippe Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *International Conference on Affective Computing and Intelligent Interaction*, pages 396–406. Springer, 2011.

[30] Pradeep K Atrey, Mohan S Kankanhalli, and John B Oommen. Goal-oriented optimal subset selection of correlated multimedia streams. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 3(1):2–es, 2007.

[31] David Lee Hall and Sonya AH McMullen. *Mathematical techniques in multisensor data fusion*. Artech House, 2004.

[32] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[33] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134. IEEE, 2015.

[34] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[35] Wanli Ouyang, Xiao Chu, and Xiaogang Wang. Multi-source deep learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2329–2336, 2014.

[36] Zuxuan Wu, Yu-Gang Jiang, Jun Wang, Jian Pu, and Xiangyang Xue. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 167–176, 2014.

[37] Desmond Elliott and Frank Keller. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, 2013.

[38] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[39] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

[40] Yu-Gang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shih-Fu Chang. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(2):352–364, 2017.

[41] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[42] Nitish Srivastava and Ruslan Salakhutdinov. Learning representations for multimodal data with deep belief nets. In *International conference on machine learning workshop*, volume 79, 2012.

[43] Yelin Kim, Honglak Lee, and Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 3687–3691. IEEE, 2013.

[44] Di Wu and Ling Shao. Multimodal dynamic networks for gesture recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 945–948, 2014.

[45] Wanli Ouyang, Xiao Chu, and Xiaogang Wang. Multi-source deep learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2329–2336, 2014.

[46] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[47] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255, 2013.

[48] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

[49] Yuxin Peng, Jinwei Qi, and Yuxin Yuan. Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing*, 27(11):5585–5599, 2018.

[50] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.

[51] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[52] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

[53] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[54] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*, 2014.

[55] Malcolm Slaney and Michele Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In *Advances in Neural Information Processing Systems*, pages 814–820, 2001.

[56] Mehmet Emre Sargin, Yücel Yemez, Engin Erzin, and A Murat Tekalp. Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Transactions on Multimedia*, 9(7):1396–1403, 2007.

[57] Pei Ling Lai and Colin Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(05):365–377, 2000.

[58] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255, 2013.

[59] Sofia C Olhede and Patrick J Wolfe. The growing ubiquity of algorithms in society: implications, impacts and innovations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170364, 2018.

[60] David Leslie. Understanding artificial intelligence ethics and safety. *arXiv preprint arXiv:1906.05684*, 2019.

[61] The Cambridge Analytica scandal Britain moves to rein in dataanalytics. The Economist.

[62] Equifax finds more victims of 2017 breach.

[63] Tom Chan, Concetta Tania Di Iorio, Craig Kuziemsky, Siaw-Teng Liaw, Simon De Lusignan, and D Lo Russo. The uk national data guardian for health and care's review of data security, consent and opt-outs: leadership in balancing public health with rights to privacy? *BMJ Health and Care Informatics*, 23(3), 2016.

[64] Porject maven.

[65] Kate Crawford, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, et al. Ai now 2019 report. *New York, NY: AI Now Institute*, 2019.

[66] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Mysers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. *AI now report 2018*. AI Now Institute at New York University New York, 2018.

[67] FAT/ML organization.

[68] Meredith Ringel Morris. Ai and accessibility: A discussion of ethical considerations. *arXiv preprint arXiv:1908.08939*, 2019.

[69] Ginger Zhe Jin. Artificial intelligence and consumer privacy. Technical report, National Bureau of Economic Research, 2018.

[70] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, Kanta Dihal, and Stephen Cave. Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. *London: Nuffield Foundation*, 2019.

[71] Vincent C Müller. Ethics of artificial intelligence and robotics. 2020.

[72] Colin Allen, Iva Smit, and Wendell Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology*, 7(3):149–155, 2005.

[73] Michael Anderson and Susan Leigh Anderson. *Machine ethics*. Cambridge University Press, 2011.

[74] Peter M Asaro. Ai ethics in predictive policing: From models of threat to an ethics of care. *IEEE Technology and Society Magazine*, 38(2):40–53, 2019.

[75] Virginia Dignum. Ethics in artificial intelligence: introduction to the special issue, 2018.

[76] Vincent C Müller and Nick Bostrom. Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental issues of artificial intelligence*, pages 555–572. Springer, 2016.

[77] Alexis Bogroff and Dominique Guegan. Artificial intelligence, data, ethics an holistic approach for risks and regulation. *University Ca'Foscari of Venice, Dept. of Economics Research Paper Series*, (19), 2019.

[78] Conseil de l'Europe et al. *Convention for the protection of individuals with regard to automatic processing of personal data*, volume 108. Council of Europe, 1981.

[79] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

[80] Will Knight. Biased algorithms are everywhere, and no one seems to care. *Technology Review*, 2017.

[81] Adrienne Yapo and Joseph Weiss. Ethical implications of bias in machine learning. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[82] Kristian Hammond. unexpected sources of bias in artificial intelligence. *online:< https://techcrunch. com/2016/12/10/5-unexpected-sources-of-bias-inartificial-intelligence*, 5.

[83] Brent Daniel Mittelstadt and Luciano Floridi. The ethics of big data: current and foreseeable issues in biomedical contexts. *Science and engineering ethics*, 22(2):303–341, 2016.

[84] Linnet Taylor and Nadezhda Purtova. What is responsible and sustainable data science? *Big Data & Society*, 6(2):2053951719858114, 2019.

[85] Bernd Carsten Stahl and David Wright. Ethics and privacy in ai and big data: Implementing responsible research and innovation. *IEEE Security & Privacy*, 16(3):26–33, 2018.

[86] John M Abowd. How will statistical agencies operate when all data are private? 2016.

[87] Bernd Carsten Stahl and David Wright. Ethics and privacy in ai and big data: Implementing responsible research and innovation. *IEEE Security & Privacy*, 16(3):26–33, 2018.

[88] Sandra Wachter and Brent Mittelstadt. A right to reasonable inferences: Re-thinking data protection law in the age of big data and ai. *Colum. Bus. L. Rev.*, page 494, 2019.

[89] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on privacy enhancing technologies*, 2015(1):92–112, 2015.

[90] Yuri Gurevich, Efim Hudis, and Jeannette M Wing. Inverse privacy. *Communications of the ACM*, 59(7):38–42, 2016.

[91] European Group on Ethics in Science, New Technologies, et al. Statement on artificial intelligence, robotics and 'autonomous' systems. *Retrieved September*, 18:2018, 2018.

[92] Michael Anderson and Susan Leigh Anderson. Machine ethics: Creating an ethical intelligent agent. *Ai Magazine*, 28(4):15–15, 2007.

[93] Raul Hakli and Pekka Mäkelä. Moral responsibility of robots and hybrid agents. *The Monist*, 102(2):259–275, 2019.

[94] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[95] Patrick Lin. Why ethics matters for autonomous cars. In *Autonomous driving*, pages 69–85. Springer, Berlin, Heidelberg, 2016.

[96] Sven Nyholm. The ethics of crashes with self-driving cars: a roadmap, ii. *Philosophy Compass*, 13(7):e12506, 2018.

[97] Geoff Keeling. Why trolley problems matter for the ethics of automated vehicles. *Science and engineering ethics*, 26(1):293–307, 2020.

[98] Daniele Amoroso and Guglielmo Tamburrini. The ethical and legal case against autonomy in weapons systems. *Global Jurist*, 18(1), 2017.

[99] Amanda Sharkey. Autonomous weapons systems, killer robots and human dignity. *Ethics and Information Technology*, 21(2):75–87, 2019.

[100] Patrick Lin, George Bekey, and Keith Abney. Autonomous military robotics: Risk, ethics, and design. Technical report, California Polytechnic State Univ San Luis Obispo, 2008.

[101] John Danaher and Neil McArthur. *Robot sex: Social and ethical implications*. MIT Press, 2017.

[102] Anne Gerdes. The issue of moral consideration in robot ethics. *Acm Sigcas Computers and Society*, 45(3):274–279, 2016.

[103] Vincent C Müller and Thomas W Simpson. Autonomous killer robots are probably good news. *Drones and responsibility: Legal, philosophical and socio-technical perspectives on the use of remotely controlled weapons*, pages 67–81, 2016.

[104] General Data Protection Regulation. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46. *Official Journal of the European Union (OJ)*, 59(1-88):294, 2016.

[105] Elisa Bertino, Ahish Kundu, and Zehra Sura. Data transparency with blockchain and ai ethics. *Journal of Data and Information Quality (JDIQ)*, 11(4):1–8, 2019.

[106] Chenghua Xu, Yunhong Wang, Tieniu Tan, and Long Quan. A new attempt to face recognition using 3d eigenfaces. In *Proceedings of the Asian Conference on Computer Vision*, volume 2, pages 884–889. Citeseer, 2004.

[107] R Sala Llonch, Effrosini Kokiopoulou, I Tošić, and Pascal Frossard. 3d face recognition with sparse spherical representations. *Pattern Recognition*, 43 (3):824–834, 2010.

[108] Stefano Berretti, Alberto Del Bimbo, and Pietro Pala. Distinguishing facial features for ethnicity-based 3d face recognition. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):45, 2012.

[109] Lahoucine Ballihi, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava, and Driss Aboutajdine. Boosting 3-d-geometric features for efficient face recognition and gender classification. *IEEE Transactions on Information Forensics and Security*, 7(6):1766–1779, 2012.

[110] Kevin W Bowyer, Kyong Chang, and Patrick Flynn. A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition. *Computer vision and image understanding*, 101(1):1–15, 2006.

[111] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[112] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[113] Peter N. Belhumeur, João P Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7): 711–720, 1997.

[114] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[115] Marian Stewart Bartlett, Javier R Movellan, and Terrence J Sejnowski. Face recognition by independent component analysis. *IEEE Transactions on neural networks*, 13(6):1450–1464, 2002.

[116] Itzhak Fogel and Dov Sagi. Gabor filters as texture discriminator. *Biological cybernetics*, 61(2):103–113, 1989.

[117] Matti Pietikäinen. Local binary patterns. *Scholarpedia*, 5(3):9775, 2010.

[118] Matti Pietikäinen, Abdenour Hadid, Guoying Zhao, and Timo Ahonen. Local binary patterns for still images. In *Computer vision using local binary patterns*, pages 13–47. Springer, 2011.

[119] Phillip Ian Wilson and John Fernandez. Facial feature detection using haar classifiers. *Journal of Computing Sciences in Colleges*, 21(4):127–133, 2006.

[120] Ville Ojansivu and Janne Heikkilä. Blur insensitive texture classification using local phase quantization. In *International conference on image and signal processing*, pages 236–243. Springer, 2008.

[121] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[122] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. 2005.

[123] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 471–478. IEEE, 2018.

[124] Guosheng Hu, Yongxin Yang, Dong Yi, Josef Kittler, William Christmas, Stan Z Li, and Timothy Hospedales. When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 142–150, 2015.

[125] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[126] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[127] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[128] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[129] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[130] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[131] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1891–1898, 2014.

[132] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014.

[133] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008.

[134] Rui Min, Neslihan Kose, and Jean-Luc Dugelay. Kinectfacedb: A kinect database for face recognition. *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, 44(11):1534–1548, 2014.

[135] Billy YL Li, Ajmal Mian, Wanquan Liu, and Aneesh Krishna. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 186–192. IEEE, 2013.

[136] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013.

[137] Andrea F Abate, Michele Nappi, Daniel Riccio, and Gabriele Sabatino. 2d and 3d face recognition: A survey. *Pattern Recognition Letters*, 28(14): 1885–1906, 2007.

[138] RI Hg, Petr Jasek, Clement Rofidal, Kamal Nasrollahi, Thomas B Moeslund, and Gabrielle Tranchet. An rgb-d database using microsoft's kinect for windows for face detection. In *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*, pages 42–46. IEEE, 2012.

[139] Rui Min, Jongmoo Choi, Gérard Medioni, and Jean-Luc Dugelay. Real-time 3d face identification from a depth camera. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1739–1742. IEEE, 2012.

[140] Gaurav Goswami, Samarth Bharadwaj, Mayank Vatsa, and Rajdeep Singh. On rgb-d face recognition using kinect. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pages 1–6. IEEE, 2013.

[141] Tri Huynh, Rui Min, and Jean-Luc Dugelay. An efficient lbp-based descriptor for facial depth images applied to gender recognition using rgb-d face data. In *Computer Vision-ACCV 2012 Workshops*, pages 133–145. Springer, 2013.

[142] Mauricio Pamplona Segundo, Santonu Sarkar, Dmitry Goldgof, Leandro Silva, and Olga Bellon. Continuous 3d face authentication using rgb-d cameras. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 64–69. IEEE, 2013.

[143] Rahul Ajmera, Aditya Nigam, and Phalguni Gupta. 3d face recognition using kinect. In *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*, page 76. ACM, 2014.

[144] Stepán Mrácek, Martin Drahanskỳ, Radim Dvorák, Ivo Provaznik, and Jan Vána. 3d face recognition on low-cost depth sensors. In *BIOSIG*, pages 195–202, 2014.

[145] Gee-Sern Jison Hsu, Yu-Lun Liu, Hsiao-Chia Peng, and Po-Xun Wu. Rgb-d-based face reconstruction and recognition. *IEEE Transactions on Information Forensics and Security*, 9(12):2110–2118, 2014.

[146] Elhocine Boutellaa, Abdenour Hadid, Messaoud Bengherabi, and Samy Ait-Aoudia. On the use of kinect depth data for identity, gender and ethnicity classification from facial images. *Pattern Recognition Letters*, 2015.

[147] João Baptista Cardia Neto and Aparecido Nilceu Marana. 3dlbp and haog fusion for face recognition utilizing kinect as a 3d scanner. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pages 66–73. ACM, 2015.

[148] Yonggang Huang, Yunhong Wang, and Tieniu Tan. Combining statistics of geometrical and correlative features for 3d face recognition. In *BMVC*, pages 879–888. Edinburgh, 2006.

[149] Hamed Kiani Galoogahi and Terence Sim. Inter-modality face sketch recognition. In *2012 IEEE International Conference on Multimedia and Expo*, pages 224–229. IEEE, 2012.

[150] Poornima Krishnan and S Naveen. Rgb-d face recognition system verification using kinect and frav3d databases. *Procedia Computer Science*, 46:1653–1660, 2015.

[151] Munawar Hayat, Mohammed Bennamoun, and Amar A El-Sallam. An rgb–d based image set classification for robust face recognition from kinect data. *Neurocomputing*, 171:889–900, 2016.

[152] Xu Dai, Shouyi Yin, Peng Ouyang, Leibo Liu, and Shaojun Wei. A multi-modal 2d+ 3d face recognition method with a novel local feature descriptor. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 657–662. IEEE, 2015.

[153] Qiu Jin, Jieyu Zhao, and Yuanyuan Zhang. Facial feature extraction with a depth aam algorithm. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*, pages 1792–1796. IEEE, 2012.

[154] Anil K Jain and Stan Z Li. *Handbook of face recognition*, volume 1. Springer, 2005.

[155] Stephen Milborrow. Stasm 4 user manual. *h ttp://www. milbo. org. stasm-files/stasm4. pdf*, 2013.

[156] S. Milborrow and F. Nicolls. Active Shape Models with SIFT Descriptors and MARS. *VISAPP*, 2014. `http://www.milbo.users.sonic.net/stasm`.

[157] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037–2041, 2006.

[158] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[159] Manuele Bicego, Andrea Lagorio, Enrico Grosso, and Massimo Tistarelli. On the use of sift features for face authentication. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pages 35–35. IEEE, 2006.

[160] Piyanuch Silapachote, Deepak R Karuppiah, and Allen R Hanson. Feature selection using adaboost for face expression recognition. Technical report, MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER SCIENCE, 2005.

[161] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *ICML*, volume 96, pages 148–156, 1996.

[162] Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *Journal of machine learning research*, 5(Jan):101–141, 2004.

[163] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[164] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[165] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[166] Kobbi Nissim and Alexandra Wood. Is privacy privacy? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170358, 2018.

[167] SC Olhede and PJ Wolfe. The growing ubiquity of algorithms in society: implications, impacts and innovations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376 (2128):20170364, 2018.

[168] C-A Azencott. Machine learning and genomics: precision medicine versus patient privacy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170350, 2018.

[169] Bettina Berendt. Privacy beyond confidentiality, data science beyond spying: From movement data and data privacy towards a wider fundamental rights discourse. In *Annual Privacy Forum*, pages 59–71. Springer, 2019.

[170] SB Choi, MS Seong, and KS Kim. Vibration control of an electrorheological fluid-based suspension system with an energy regenerative mechanism. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 223(4):459–469, 2009.

[171] J Maxwell Donelan, Qinggua Li, Veronica Naing, JA Hoffer, DJ Weber, and Arthur D Kuo. Biomechanical energy harvesting: generating electricity during walking with minimal user effort. *Science*, 319(5864):807–810, 2008.

[172] Lawrence C Rome, Louis Flynn, Evan M Goldman, and Taeseung D Yoo. Generating electricity while walking with loads. *Science*, 309(5741):1725–1728, 2005.

[173] Yuan Rao, Shuo Cheng, and David P Arnold. An energy harvesting system for passively generating power from human activities. *Journal of Micromechanics and Microengineering*, 23(11):114012, 2013.

[174] Guohao Lan, Weitao Xu, Sara Khalifa, Mahbub Hassan, and Wen Hu. Veh-com: Demodulating vibration energy harvesting for short range communication. In *Pervasive Computing and Communications (PerCom), 2017 IEEE International Conference on*, pages 170–179. IEEE, 2017.

[175] RJM Vullers, Rob van Schaijk, Inge Doms, Chris Van Hoof, and R Mertens. Micropower energy harvesting. *Solid-State Electronics*, 53 (7):684–693, 2009.

[176] Sara Khalifa, Mahbub Hassan, and Aruna Seneviratne. Feasibility and accuracy of hotword detection using vibration energy harvester. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2016 IEEE 17th International Symposium on A*, pages 1–9. IEEE, 2016.

[177] Li Zhang, Parth H Pathak, Muchen Wu, Yixin Zhao, and Prasant Mohapatra. Accelword: Energy efficient hotword detection through accelerometer. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, pages 301–315. ACM, 2015.

[178] Nikolay Matyunin, Jakub Szefer, and Stefan Katzenbeisser. Zero-permission acoustic cross-device tracking. In *2018 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, pages 25–32. IEEE, 2018.

[179] Anupam Das, Nikita Borisov, and Matthew Caesar. Tracking mobile web users through motion sensors: Attacks and defenses. In *NDSS*, 2016.

[180] Jun Han, Emmanuel Owusu, Le T Nguyen, Adrian Perrig, and Joy Zhang. Accomplice: Location inference using accelerometers on smartphones. In *2012 Fourth International Conference on Communication Systems and Networks (COMSNETS 2012)*, pages 1–9. IEEE, 2012.

[181] Adam J Aviv, Benjamin Sapp, Matt Blaze, and Jonathan M Smith. Practicality of accelerometer side channels on smartphones. In *Proceedings of the 28th Annual Computer Security Applications Conference*, pages 41–50. ACM, 2012.

[182] Chen Song, Feng Lin, Zhongjie Ba, Kui Ren, Chi Zhou, and Wenyao Xu. My smartphone knows what you print: Exploring smartphone-based side-channel attacks against 3d printers. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 895–907. ACM, 2016.

[183] Emmanuel Owusu, Jun Han, Sauvik Das, Adrian Perrig, and Joy Zhang. Accessory: password inference using accelerometers on smartphones. In *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications*, page 9. ACM, 2012.

[184] Emiliano Miluzzo, Alexander Varshavsky, Suhrid Balakrishnan, and Romit Roy Choudhury. Tapprints: your finger taps have fingerprints. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, pages 323–336. ACm, 2012.

[185] Laurent Simon and Ross Anderson. Pin skimmer: Inferring pins through the camera and microphone. In *Proceedings of the Third ACM workshop on Security and privacy in smartphones & mobile devices*, pages 67–78. ACM, 2013.

[186] Yang Zhang, Peng Xia, Junzhou Luo, Zhen Ling, Benyuan Liu, and Xinwen Fu. Fingerprint attack against touch-enabled devices. In *Proceedings of the second ACM workshop on Security and privacy in smartphones and mobile devices*, pages 57–68. ACM, 2012.

[187] Liang Cai and Hao Chen. Touchlogger: Inferring keystrokes on touch screen from smartphone motion. *HotSec*, 11:9–9, 2011.

[188] Sanorita Dey, Nirupam Roy, Wenyuan Xu, Romit Roy Choudhury, and Srihari Nelakuditi. Accelprint: Imperfections of accelerometers make smartphones trackable. In *NDSS*, 2014.

[189] Tom Van Goethem, Wout Scheepers, Davy Preuveneers, and Wouter Joosen. Accelerometer-based device fingerprinting for multi-factor mobile authentication. In *International Symposium on Engineering Secure Software and Systems*, pages 106–121. Springer, 2016.

[190] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. Gyrophone: Recognizing speech from gyroscope signals. In *USENIX Security*, pages 1053–1067, 2014.

[191] Simon Castro, Robert Dean, Grant Roth, George T Flowers, and Brian Grantham. Influence of acoustic noise on the dynamic performance of mems gyroscopes. In *ASME 2007 International Mechanical Engineering Congress and Exposition*, pages 1825–1831. American Society of Mechanical Engineers, 2007.

[192] Robert N Dean, George T Flowers, A Scotte Hodel, Grant Roth, Simon Castro, Ran Zhou, Alfonso Moreira, Anwar Ahmed, Rifki Rifki, Brian E Grantham, et al. On the degradation of mems gyroscope performance in the presence of high power acoustic noise. In *2007 IEEE International Symposium on Industrial Electronics*, pages 1435–1440. IEEE, 2007.

[193] Robert Neal Dean, Simon Thomas Castro, George T Flowers, Grant Roth, Anwar Ahmed, Alan Scottedward Hodel, Brian Eugene Grantham, David Allen Bittle, and James P Brunsch. A characterization of the performance of a mems gyroscope in acoustically harsh environments. *IEEE Transactions on Industrial Electronics*, 58(7):2591–2596, 2010.

[194] Jun Han, Albert Jin Chung, and Patrick Tague. Pitchin: Eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion. 2017.

[195] S Abhishek Anand and Nitesh Saxena. Speechless: Analyzing the threat to speech privacy from smartphone motion sensors. In *2018 IEEE Symposium on Security and Privacy (SP). Vol. 00*, pages 116–133, 2018.

[196] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2012.

[197] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[198] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[199] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[200] Daixin Wang, Peng Cui, Mingdong Ou, and Wenwu Zhu. Deep multimodal hashing with orthogonal regularization. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[201] Carina Silberer and Mirella Lapata. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, 2014.

[202] Houssem Maghrebi, Thibault Portigliatti, and Emmanuel Prouff. Breaking cryptographic implementations using deep learning techniques. In *International Conference on Security, Privacy, and Applied Cryptography Engineering*, pages 3–26. Springer, 2016.

[203] Eleonora Cagli, Cécile Dumas, and Emmanuel Prouff. Convolutional neural networks with data augmentation against jitter-based countermeasures. In *International Conference on Cryptographic Hardware and Embedded Systems*, pages 45–68. Springer, 2017.

[204] Jun Han, Albert Jin Chung, and Patrick Tague. Pitchln: eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion. In *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pages 181–192. ACM, 2017.

[205] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49, 2012.

[206] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[207] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408, 2010.

[208] Jacob Leon Kröger, Otto Hans-Martin Lutz, and Philip Raschke. Privacy implications of voice and speech analysis–information disclosure by inference. In *IFIP International Summer School on Privacy and Identity Management*, pages 242–258. Springer, 2019.

[209] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, 2015.

[210] Sei Jin Ko, Melody S Sadler, and Adam D Galinsky. The sound of power: Conveying and detecting hierarchical rank through voice. *Psychological Science*, 26(1):3–14, 2015.

[211] Guozhen An and Rivka Levitan. Lexical and acoustic deep learning model for personality recognition. In *Interspeech*, pages 1761–1765, 2018.

[212] Mahesh Kumar Nandwana, Julien van Hout, Mitchell McLaren, Allen R Stauffer, Colleen Richey, Aaron Lawson, and Martin Graciarena. Robust speaker recognition from distant speech under real reverberant environments using speaker embeddings. In *Interspeech*, pages 1106–1110, 2018.

[213] Iosif Mporas and Todor Ganchev. Estimation of unknown speaker's height from speech. *International Journal of Speech Technology*, 12(4):149–160, 2009.

[214] Harishchandra Dubey, Matthias R Mehl, and Kunal Mankodiya. Bigear: Inferring the ambient and emotional correlates from smartphone-based acoustic big data. In *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 78–83. IEEE, 2016.

[215] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805, 2013.

[216] The 14 biggest data breaches of the 21st century .

[217] Linnet Taylor, Luciano Floridi, and Bart Van der Sloot. *Group privacy: New challenges of data technologies*, volume 126. Springer, 2016.

[218] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[219] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

[220] Martin Längkvist, Lars Karlsson, and Amy Loutfi. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42:11–24, 2014.