UCLouvain

École polytechnique de Louvain

Tourists' Emotion and Satisfaction Estimation: Mitigating Burden, Cost, and Privacy Concerns

Author: **Lucas Maris**
Supervisors: **Ramin Sadre, Keiichi Yasumoto, Yuki Matsuda**
Readers: **Jean Vanderdonckt, Benoît Duhoux**
Academic year 2021–2022
Master [120] in Computer Science

Abstract

The ever-growing amount of sensors and smart devices has lead to a growing interest in context-aware systems, to leverage the data collection possibilities these ubiquitous sensors can provide. Smart tourism is one of these research directions, where the aim is to recognise the state of tourists to provide them with a personalised sightseeing experience. With wearable sensing devices, it becomes possible to monitor tourist's behavioural cues and facial expressions, amongst other signals. By applying artificial intelligence and machine learning techniques, this data can prove useful to recognise their emotional status and predict their satisfaction level at given points in time. Such data is expected to offer beneficial insights regarding an individual's interests, and can serve to improve their overall sightseeing experience.

This master thesis, realised in collaboration with the Ubiquitous Computing Systems Laboratory at the Nara Institute of Science and Technology, focuses on their previously devised EmoTour system, aiming to make it better-suited for real-world applications. EmoTour is a context-aware system, capable of estimating a tourist's emotions and satisfaction throughout sightseeing by combining multimodal sensing with predictive models. Two main directions are explored, the first of which centres on reducing the financial cost and physical burden of the system. The second direction aims to answer the privacy concerns such a system raises, and is examined through a series of different approaches. New experimental data was gathered to evaluate the applicability of the proposed changes to the system, and to further investigate differences between user subgroups.

Acknowledgements

I would like to express my gratitude to everyone who has contributed to this thesis or supported me at any point during my past five years at UCLouvain.

This includes Professor Ramin Sadre, for taking it upon him to supervise me and for bearing with me all throughout; Professors Keiichi Yasumoto and Yuki Matsuda, who did their best to provide me with guidance despite the rather unusual setting; and Ninouchka Malan, for her help and patience with organising the exchange programme that could never be.

My deepest thanks to my mother, my father and my beloved sister, for their undying love and support; Guillaume Navette, for his life-long friendship; Corentin Petre, for this memorable final year; Christophe Antoni and Mathilde Perruchoud, for the lovely breaks they've made possible; Sanae Harika, for her kindness; my friends from le U, for their lasting camaraderie; my buddies from the matchbox, for the comfort they've brought me; and last but not least, my roommates at Oenokot, for the memories I will treasure.

Contents

Chapter 1

Introduction

The Internet of Things (IoT) is a technology that enables physical objects to communicate, perform sensing, process data and be controlled through a communication network (which, interestingly, does not necessarily have to be the Internet). It is a broadly used term, often used to characterise any system that collects data through a variety of devices (i.e., sensors) in a given environment and uses said data to understand or have an effect on the environment. Somewhere in between fields as varied as Ubiquitous Computing, Wireless Sensor Networks, Automation and Machine Learning, IoT's applications range from Smart Homes, Cities and Healthcare to Smart Manufacturing, Agriculture and Energy Management.

The number of Internet of Things devices worldwide has been steadily growing over the years and is forecasted to amount to over 25 billion by 2030 [1]. In the meantime, machine learning has become the dominant trend within artificial intelligence, almost as a response to the ever-increasing volume of data that these (and numerous other) devices generate. One of the many sectors that can profit from the combination of these two fields is tourism, where the application of such technologies is believed to enable the development of innovative tools that improve tourism and tourist's experiences, and various research projects centre around this topic.

Smart Tourism is generally defined as a component of a Smart City [2]. It aims to improve the attractiveness of touristic experiences, to achieve long-term growth in the tourism sector, one of the world's largest sectors [3]. Understanding tourist's desires and mood is capital to offer tourists with a satisfying and personal experience, and this is something emotion recognition technologies based on wearable devices and IoT can help realise.

One research project focused on such matters is EmoTour [4, 5], which was initiated by a collaboration between the Ubiquitous Computing Systems Laboratory at the Nara Institute of Science and Technology, in Japan, and the Institute of Communications Engineering at Ulm University, in Germany. The aim of this project is to estimate the emotional status and satisfaction level of tourists during sightseeing, by multimodally sensing their unconscious or natural behaviour. Previous experiments conducted in Ulm (Germany), Nara (Japan) and Kyoto (Japan) had participants wear various wearable devices throughout predefined touristic routes.

While this information could prove useful for improving tourist experience, wearable sensing systems face many challenges surrounding user experience and privacy. For Smart Tourism to be implemented in real-world applications, it must therefore focus on providing useful applications, being easy and convenient to use, as well as preserving the privacy of its users. The objective of this research is thus two-fold: to lessen the financial cost and user burden of the EmoTour system and to address (some of the) privacy concerns it raises.

The first of these objectives is explored here through a comparative analysis of machine learning models trained on different feature sets, some of them leaving out the features obtained through costly and/or burdensome sensors. This allows to characterise modality relevance, and gives insights on the components that have the largest impact on the system's prediction results. On top of this, Principal Component Analysis is applied as a dimensionality reduction technique, as a way to streamline the high-dimensional dataset and to improve the accuracy of predictions.

The second objective is considered through a small variety of methods, evaluated on intentional and adverse tasks. These approaches include an importance-based feature selection, multi-task adversarial learning, and differential privacy. To confirm their effectiveness, a new data collection experiment was conducted at UCLouvain's Musée L, in Louvain-la-Neuve. A subgroup analysis was also conducted on the newly extended EmoTour dataset, to evaluate the possibility to tailor models to user's cultural backgrounds, gender identities or age.

The rest of this work is organised as follows; Chapter 2 goes over the fundamental concepts and tools this work relies on; Chapter 3 gives a complete overview of the EmoTour system, as proposed in [4, 5]; Chapter 4 explores the relative importance of costly and/or burdensome features; Chapter 5 summarises the result of different privacy approaches tried on EmoTour; Chapter 6 presents the experiment conducted at Musée L; finally, Chapter 7 concludes this work.

The data, code and results presented in this work have been uploaded to the repository available at `https://github.com/LucasMaris/LEPL2990`.

Chapter 2

Fundamentals

This chapter introduces the theoretical concepts used over the course of this thesis. These include machine learning and neural networks at large, central to EmoTour's prediction model, principal component analysis, for use throughout the next chapters, and permutation importance, minimax filters, and differential privacy, which ground the privacy aspect explored here. EmoTour's feature extraction tools, openSMILE, OpenFace, and Pupil, are also introduced at the end of this chapter.

2.1 Machine Learning

2.1.1 Neural Networks

An Artifical Neural Network (ANN), more commonly referred to as a Neural Network, is a computation model loosely inspired by a biological brain and aiming to emulate its natural perceptive skills. It is a collection of connected units, often referred to as (artificial) neurons, each of which will receive, process and output a signal. ANNs are often used in the context of classification tasks and regression analysis; in other words, they are often trained to approximate a target function, which either classifies data points into classes or estimates the relationship between inputs and outputs.

Artificial Neuron

The first artificial neuron proposed was the perceptron [6], i.e., a simple function that maps its input $\mathbf{x}$ to a binary output $f(\mathbf{x})$, of which the equation is given below. More broadly, artificial neurons are functions that proceed in two steps: they first compute a weighted sum on their inputs, with bias, and they then apply some

activation function to the resulting value. The perceptron defines this activation function as the unit step function, but this need not be the case.

$$z = \mathbf{w}\mathbf{x} + b$$
$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{otherwise} \end{cases}$$

Multilayer Perceptron

The most "vanilla" form of an ANN is the Multilayer Perceptron (MLP), which consists of at least 3 fully connected layers of feed-forward neurons: an input layer, one or more hidden layers, and an output layer. The input layer passes on input data to the hidden layer, which performs a weighted sum and applies a bias to this data, and then passes the result through a nonlinear activation function. The outcome is then fed to the next hidden layer, which repeats the process, and so on until the output layer is reached. This last layer repeats the process one final time, thereby determining the output of the entire network.

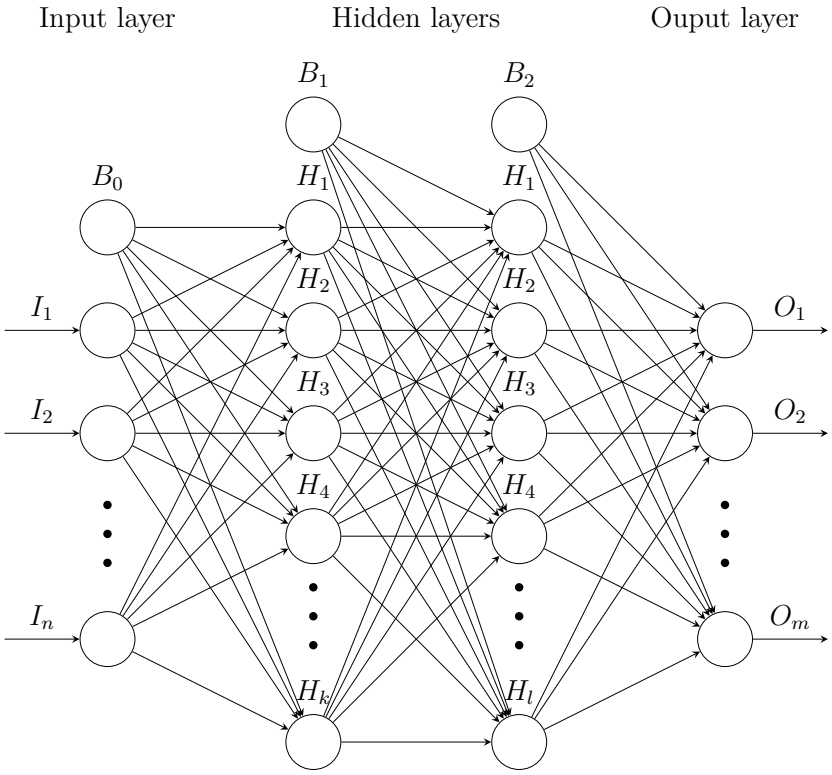

Figure 2.1: Simple Multilayer Perceptron with two hidden layers.

Model selection

An example MLP is shown in Figure 2.1. This example network has n inputs, m outputs, and two hidden layers composed of k and l neurons, respectively. These values, as well as many other parameters, such as the chosen activation function or the number of training epochs, collectively referred to as hyperparameters, impact the behaviour and performance of the model in a way that is difficult to quantify in a theoretical manner. It is therefore common to try out many different possible models with varying hyperparameters on the same task and dataset, in order to discover an architecture that works for the task at hand. This process is called model selection.

2.1.2 Learning

Learning is the process of fitting a model to a chosen task. In a Multilayer Perceptron, it happens iteratively through a supervised learning process: this means the model is fed labelled data, i.e., data points with their expected outputs, and gradually adjusts its parameters (weights and biases) in an attempt to minimise the error of the model on the training data.

Data splitting

Training a model on some dataset always requires some form of data separation, to evaluate the ability of the trained model to generalise to new, unseen data samples. As models are explicitly trained to minimise their error on training samples, it is necessary to keep aside a portion of the dataset to act as unseen, "new" data samples. This partition of the dataset is usually refered to as the test(ing) set, while the partition of the dataset used for training is aptly called the train(ing) set.

The actual learning process is only carried out on the training set. After model training has been completed, the model is fed the testing set, and the difference between the network's output and the expected output is used as a means to quantify the generalisation ability of the trained model. This difference is computed through a performance metric that can be freely defined.

In this work, Unweighted Average Recall (UAR) is used as the performance metric for classification tasks, and is to be maximised. It is defined as the mean recall value, i.e., the average fraction of samples that are correctly labelled, meaning the predictions y correspond to the expected outputs x, for every class c:

$$UAR(x, y) = \frac{1}{c} \sum_{i=1}^{c} \frac{TP(x, y, c)}{TP(x, y, c) + FN(x, y, c)}$$

$$TP(x, y, c) = |\{i \in \{1, 2, ..., n\} : x_i = y_i = c\}|$$
$$FN(x, y, c) = |\{i \in \{1, 2, ..., n\} : x_i \neq y_i = c\}|$$

For regression tasks, the performance metric used in this work is the Mean Absolute Error (MAE), which is to be minimised. It is simply defined as the absolute difference between predictions y and expected outputs x:

$$MAE(x, y) = \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i|$$

Backpropagation

Supervised learning within artificial neural networks hinges on the backpropagation algorithm, which uses stochastic gradient descent to update the model's parameters. Each iteration of this algorithm will compute the prediction error on a (set of) training sample(s) put through the network. The error or loss function will quantify the difference between the output of the network and the expected output, for one or more data samples.

The prediction error computed at the output layer is partially derived with respect to every single weight and bias in the model in a backwards fashion, from the last to the first layer of the network. Using this partial derivative, the value of every parameter in the model is slightly nudged towards a value that reduces the computed prediction error. This process is repeated either for a fixed number of steps, called epochs, or until stabilisation of the prediction error.

2.2 Principal Component Analysis

2.2.1 Overview

Principal Component Analysis (PCA) [7, 8] is one of the oldest and most widely used data analysis techniques, capable of drastically reducing the dimensionality of large datasets while preserving as much information as possible contained in the data it is applied to. Its application to a given dataset creates new, uncorrelated variables, called Principal Components (PC), which are simple linear combinations of the existing variables in such a way that they successively maximise the data's variability. PCA can help increase the interpretability of large datasets and decrease model complexity, which can reduce overfitting, improve the quality of predictions and reduce their computational cost.

2.2.2 Decorrelation process

To apply PCA to a given $n \times m$ data matrix $\mathbf{X}$, where n is the number of data samples and m is the number of features (dimensions), the following steps are applied:

- **Standardisation:** For each feature, the mean $\overline{x}$ and standard deviation s is computed. For each value x of that feature in matrix $\mathbf{X}$, the mean $\overline{x}$ is then subtracted, and the result is divided by the standard deviation s. Mathematically, $\forall i \in [1, n], \forall j \in [1, m]$:

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{s_j}$$

- **Covariance matrix:** To characterise the linear relationships between features z_j, the $m \times m$ covariance matrix $\mathbf{\Sigma}$ is computed on the standardised data matrix $\mathbf{Z}$.

$$\mathbf{\Sigma} = \begin{bmatrix} cov(z_1, z_1) & cov(z_1, z_2) & \cdots & (z_1, z_m) \\ cov(z_2, z_1) & cov(z_2, z_2) & \cdots & (z_2, z_m) \\ \vdots & \vdots & \ddots & \vdots \\ cov(z_m, z_1) & cov(z_m, z_2) & \cdots & (z_m, z_m) \end{bmatrix} = \frac{1}{m-1}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}$$

- **Singular value decomposition:** The covariance matrix $\mathbf{\Sigma}$ is symmetric, and can thus be diagonalised as $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, where each column in the $m \times m$ matrix $\mathbf{V}$ is an eigenvector and $\mathbf{\Lambda}$ is a $m \times m$ diagonal matrix with eigenvalues λ_i in decreasing order on the diagonal. These eigenvectors correspond to the principal directions of the data, and their corrresponding eigenvalue is a measure of its significance.
- **Deriving the decorrelated data:** The new, decorrelated, $n \times m$ data matrix $\mathbf{X}'$ is then simply $\mathbf{X}' = \mathbf{X}\mathbf{V}$. Each column of this data matrix $\mathbf{X}'$ is called a principal component.

2.2.3 Dimensionality reduction

Once the decorrelated data $\mathbf{X}'$ has been obtained, PCA dimensionality reduction is as simple as dropping a given number of principal components. Keeping only the p first columns of $\mathbf{X}'$, i.e., the p most significant principal components of the dataset, is guaranteed, by construction, to maximise the variance preserved from the original data $\mathbf{X}$.

2.3 Permutation importance

Permutation importance [9] is a method to compute the importance of features, by measuring how much the model's score decreases when a feature is made unavailable. Intuitively, to evaluate the impact of the absence of a given feature, a simple solution appears to simply retrain a model on the same dataset, minus the chosen feature. However, such an approach would require retraining a model for (the absence of) each feature, which can be computationally expensive.

To avoid this complication, the permutation importance method will simply evaluate the model's performance on a dataset where the chosen feature is replaced with random noise, which is in theory equivalent to it being absent. The name "permutation importance" comes from the fact that said noise is obtained by simply shuffling around the values for a feature, i.e., every data sample will use another data sample's value for the feature.

2.4 Minimax Filters

2.4.1 Overview

When an individual agrees to the usage of their data for some specified purpose, e.g., the recognition of their emotional status during sightseeing, they do not want that very same data to be revealing of other sensitive information, e.g., their identity or gender. Simply stripping the collected data of these sensitive attributes alone is not enough to guarantee this cannot happen, as the data itself is likely to correlate to some extent to said attributes, and can be used to infer them in many cases. Minimax filters [10, 11] aim to answer this concern, as a task-specific approach capable of reducing potential inference attacks on publicly shared data.

The idea is to model the above-stated privacy problem as a three-party learning game: data contributors agree to the use of their data for intentional tasks, but not adversary tasks; a data aggregator collects their data with permission, and uses it to infer interesting information modelled through intentional tasks; adversaries can potentially gain access to that same data and use it to infer sensitive information modelled as adversary tasks. Minimax filters are then a transformation that can be applied by data contributors on their data to minimise the inference ability of attackers without affecting the inference ability of the data aggregator too much.

This effectively translates to a multi-objective optimisation problem searching for the optimal utility-privacy trade-off, where one wants to transform the input data using some filter such that it has maximum utility for intentional tasks but minimal utility for adverse tasks, i.e., maximum privacy.

2.4.2 Definitions

Utility measure f_{util}

Let the utility measure f_{util} be defined as below, i.e., as the expected loss l_u measured on the intentional predictions $\hat{z}$, when compared with the expected labels z for the data x. This utility measure f_{util} is dependent on model parameters θ_f, which define the filter g's transformation on x, and on model parameters θ_u, which define the intentional prediction models.

$$f_{util}(\theta_f, \theta_u) = E[l_u(\hat{z}(g(x; \theta_f), \theta_u), z)]$$

Privacy measure f_{priv}

Analogously, let the privacy measure f_{priv} be defined as the expected loss l_p measured on adversarial predictions $\hat{y}$, when compared with the sensitive labels y for the data x, as below. This privacy measure f_{priv} is also dependent on the model parameters θ_f that define the filter g's transformation on x, and on model parameters θ_p, which define the adverse prediction models.

$$f_{priv}(\theta_f, \theta_p) = E[l_p(\hat{y}(g(x; \theta_f), \theta_p), y)]$$

Optimisation objectives

Going back to the situation used to visualise the idea behind minimax filters, both the data aggregator and the potential adversary will evidently be minimising their respective measures f_{util} and f_{priv}, by tuning their respective model parameters θ_u and θ_p, in order to produce the best predictions on their respective tasks. In this situation, the data contributors however have their own model parameters θ_f to play with, in order to transform the data such that it is of maximum utility to the data aggregator but of minimum utility to the potential adversary. This translates to the two following objectives:

$$\begin{aligned}
\min_{\theta_f} \min_{\theta_u} f_{util}(\theta_f, \theta_u) &\iff \min_{\theta_f} - \max_{\theta_u} - f_{util}(\theta_f, \theta_u) \\
\max_{\theta_f} \min_{\theta_p} f_{priv}(\theta_f, \theta_p) &\iff \min_{\theta_f} \max_{\theta_p} - f_{priv}(\theta_f, \theta_p)
\end{aligned}$$

Achieving these two conflicting goals then corresponds to the following joint optimisation problem, where the constant λ defines the utility-privacy trade-off. The higher this constant, the closer the problem to a utility-only task.

$$\min_{\theta_f} \left[-\lambda \max_{\theta_u} - f_{util}(\theta_f, \theta_u) + \max_{\theta_p} - f_{priv}(\theta_f, \theta_p) \right]$$

2.5 Differential Privacy

2.5.1 Overview

Differential Privacy (DP) [12, 13] is a mechanism for publicly sharing information about a dataset without compromising the privacy of the individuals that provided its data samples. It aims to give each individual whose data is found in the database roughly the same privacy that would result from having their data removed, i.e., to make the output of the dataset's application not overly dependent on the data of any one individual. As it aims to address the paradoxical problem of learning useful information about a population without learning anything about a given individual, it defines a tradeoff between data privacy and utility.

2.5.2 Definitions

ϵ-Differential Privacy

A randomised algorithm M is said to be ϵ-differentially private (ϵ-DP) if, for any two datasets x and y differing in only one data sample, and for any output S the algorithm M could yield, it holds that:

$$P[M(x) \in S] \leq e^{\epsilon} P[M(y) \in S]$$

$$\Leftrightarrow \frac{P[M(x) \in S]}{P[M(y) \in S]} \leq e^{\epsilon}$$

The smaller ϵ, the more likely M is to yield the same output for datasets differing only in one data sample, thus the more individuals' privacy is protected. This introduces plausible deniability, for every individual, as to whether or not that individual's data is present in the dataset.

(ϵ,δ)-Differential Privacy

The definition of differential privacy is sometimes extended with an extra parameter δ. As such, a randomised algorithm M is said to be (ϵ,δ)-differentially private ((ϵ,δ)-DP) if, for any two datasets x and y differing in only one data sample, and for any output S the algorithm M could yield, it holds that:

$$P[M(x) \in S] \leq e^{\epsilon} P[M(y) \in S] + \delta$$

This additional δ parameter relaxes DP slightly, by accounting for the fact that sometimes, it can be reasonable for the outputs of the randomised algorithm M on x and y to differ more than slightly, as in some cases, it can happen that a given

individual's data has a larger effect on the algorithm's output. In other words, δ is the probability that the privacy guarantee does not hold.

2.5.3 Laplacian noising mechanism

The most common way to achieve ϵ-DP is by using a Laplacian noise mechanism. It defines the randomised algorithm $M(x)$ as the application of a function $f(x)$ and the addition of noise n drawn from a Laplace distribution.

$$M(x) = f(x) + n, \text{ where } n \sim Laplacian\left(0, \frac{S(f)}{\epsilon}\right)$$

As the addition of noise n aims to hide the contribution of a single individual's data to the output of the randomised algorithm $M(x)$, it must be calibrated to the magnitude by which a single individual's data can change the output of function $f(x)$ in the worst case, i.e., for all datasets x, y that differ only in a single data sample. This quantity is defined as the $L1$-sensitivity $S(f)$:

$$S(f) = max\|f(x) - f(y)\|_1 = max(|f(x) - f(y)|)$$

2.5.4 Robustness to post-processing

A very useful property of differential privacy mechanisms is their immunity to post-processing, i.e., the impossibility to make the output of a differential privacy mechanism less differentially private by some sort of post-processing. This means one can perform arbitrary computations on the output of a differential privacy mechanism, with no danger of reversing the privacy protection it provides.

Formally, if M is an (ϵ,δ)-differentially private randomised algorithm with image R, and $f : R \to R'$ is an arbitrary deterministic mapping, then for any two datasets x and y differing in only one data sample, if $S \subseteq R'$ and $T = \{r \in R : f(r) \in S\}$, it holds that:

$$\begin{aligned} P[f(M(x)) \in S] &= P[M(x) \in T] \\ &\leq e^{\epsilon} P[M(y) \in T] + \delta = e^{\epsilon} P[f(M(y)) \in S] + \delta \end{aligned}$$

This property also holds for any randomised mapping $f : R \to R'$ [13].

2.6 Tools

2.6.1 openSMILE

openSMILE (open-source Speech and Music Interpretation by Large-space Extraction) [14] is a an open-source[1] audio feature extraction toolkit, implemented in C++. Its development first started in 2008 at the Technische Universität München in Munich, Germany, and it is now owned and maintained by *audEERING GmbH*[2]. Since its release in 2010, it has been widely used in speech recognition conferences and emotion recognition challenges such as INTERSPEECH (International Speech Communication Association) [15] and Emotiw (Emotion Recognition in the Wild Challenge) [16].

In the context of this work, openSMILE is used to extract a set of low-level descriptors from audio, namely a subset of the extended Geneva Minimalistic Accoustic Parameter Set (eGeMAPS) which is made up of the following 88 parameters, as described by Eyben et al. in [17]:

Frequency related parameters

- **Pitch**, logarithmic F_0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0).
- **Jitter**, deviations in individual consecutive F_0 period lengths.
- **Formant 1, 2, and 3 frequency**, centre frequency of first, second, and third formant.
- **Formant 1**, bandwidth of first formant.
- **Formant 2-3 bandwidth**, added for completeness of Formant 1-3 parameters.

Energy/Amplitude related parameters

- **Shimmer**, difference of the peak amplitudes of consecutive F_0 periods.
- **Loudness**, estimate of perceived signal intensity from an auditory spectrum.
- **Harmonics-to-noise ratio (HNR)**, relation of energy in harmonic components to energy in noise-like components.

[1] `https://github.com/audeering/opensmile`
[2] `https://audeering.com`

Spectral (balance/shape/dynamics) parameters

- **Alpha Ratio**, ratio of the summed energy from 50-1000 Hz and 1-5 kHz.
- **Hammarberg Index**, ratio of the strongest energy peak in the 0-2 kHz region to the strongest peak in the 2–5 kHz region.
- **Spectral Slope 0-500 Hz and 500-1500 Hz**, linear regression slope of the logarithmic power spectrum within the two given bands.
- **Formant 1, 2, and 3 relative energy**, as well as the ratio of the energy of the spectral harmonic peak at the first, second, third formant's centre frequency to the energy of the spectral peak at F_0.
- **Harmonic difference H1-H2**, ratio of energy of the first F_0 harmonic (H1) to the energy of the secondF_0 harmonic (H2).
- **Harmonic difference H1-A3**, ratio of energy of the first F_0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3).
- **MFCC 1-4**, Mel-Frequency Cepstral Coefficients 1-4.
- **Spectral flux**, difference of the spectra of two consecutive frames.

2.6.2 OpenFace

OpenFace [18] is an open-source[3] computer vision toolkit, implemented in Matlab, C++ and Python. It was first proposed in 2016 [19], then later updated to its current version, version 2.0, in 2018. Its features include facial landmark detection, head pose estimation, facial action unit recognition and eye-gaze estimation, with state-of-the-art performances for all of these tasks. Since its introduction, it has become a go-to tool for emotion recognition research and challenges, such as [20] and AVEC (Audio/Visual Emotion Challenge and Workshop) [21].

For this work, OpenFace serves as a way to characterise videos in terms of the Action Units portrayed by the people on video [22]. Action Units (AU) describe specific movements of facial muscles, as defined by the Facial Action Coding System (FACS) [23]. For each frame in a given video, OpenFace will detect facial landmarks, and use those to estimate AUs expressed on the frame. In particular, it will output a value $\in [0, 5]$ as a measure of the confidence with which it detects a given AU, for each of the following AUs:

[3] `https://github.com/TadasBaltrusaitis/OpenFace`

- AU01 – Inner Brow Raiser
- AU02 – Outer Brow Raiser
- AU04 – Brow Lowerer
- AU05 – Upper Lid Raiser
- AU06 – Cheek Raiser
- AU07 – Lid Tightener
- AU09 – Nose Wrinkler
- AU10 – Upper Lip Raiser
- AU12 – Lip Corner Puller
- AU14 – Dimpler
- AU15 – Lip Corner Depressor
- AU17 – Chin Raiser
- AU20 – Lip Stretcher
- AU23 – Lip Tightener
- AU25 – Lips Part
- AU26 – Jaw Drop
- AU45 – Blink

2.6.3 Pupil

Pupil [24] is an open-source[4] eye-tracking platform developed by Pupil Labs, to be used together with their Pupil Core hardware, a wearable eye-tracking headset. It is designed to detect pupil positions from raw eye videos captured by the headset and to map the user's gaze onto a video of the real world, which the headset also captures. Pupil has previously been used in emotion recognition research, e.g., for content recommendation based on reactions to videos [25] or to determine the emotional impact of haute cuisine for use in smart tourism [26].

Within this work, Pupil is used as a feature extraction tool for Pupil Core videos, specifically to extract spherical coordinates ϕ and θ, which represent pupil positions on a 3D circle. These values are useful to characterise the intensity of eye movement, as well as to compute statistics on the eye movement.

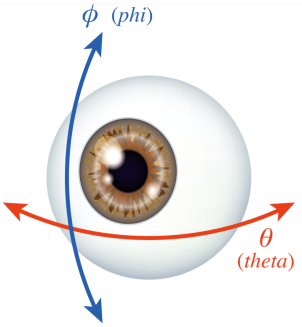

Figure 2.2: Spherical coordinates of the pupil's position[5].

[4] `https://github.com/pupil-labs/pupil`
[5] Figure reprinted from [4].

Chapter 3

EmoTour

3.1 Motivation

Traditionally, collecting the emotional status and satisfaction level of individuals has been done through questionnaires [27, 28] or online user reviews [29, 30]. While de facto standards, these methods are quite burdensome for users, requiring them to plough through long lists of questions or to write out a few paragraphs about their experience. Moreover, incentives must be devised for all sorts of participants to respond, to avoid any selection bias and subsequent biased distribution of reviews [31].

To address these issues, context-aware sensing has been investigated in many different studies, in order to get real-time information about users' emotional status and satisfaction level. Previous unimodal systems have used audio sensors [32, 33], cameras [34] or accelerometer data [35] for this purpose. Multimodal systems such as [36, 37], combining different feature types, have also been proposed, and these usually achieve higher accuracies.

EmoTour [4, 5] is one of those multimodal systems, aiming to extensively sense tourists during sightseeing using various wearable devices to measure their eye, head and body movement, as well as their facial and vocal expressions and physiological data. During experiments, participants were asked to report their emotional status and satisfaction level at different times of the trip as a way to ground-truth the collected data. The data was used to train a neural network in a supervised manner to estimate subjects' emotional status and satisfaction level for each portion of the sightseeing route.

This chapter summarises the EmoTour system devised by the Ubiquitous Computing Systems Laboratory, which serves as the starting point of this thesis.

3.2 Experiment setup and locations

The data for the EmoTour dataset was collected from voluntary participants along predefined touristic routes. Each of these routes was divided into 8 to 11 sessions, usually including at least one sight each, during which the data from the participants' wearable devices was continuously recorded, and after each of which participants were asked to record a selfie video to briefly comment on their enjoyment of this session. Table 3.1 gives an overview of the previously collected dataset. Each session was labelled in terms of emotional status and satisfaction level by the participants themselves, during their sightseeing.

Place	**#Participants**	**#Sessions**	**Experiment dates**
Ulm, Germany	18	149	Dec. 2017 - Aug. 2018
Nara, Japan	5	40	Jan. 2018
Kyoto, Japan	24	263	Mar. 2019
Total	47	452	Dec. 2017 - Mar. 2019

Table 3.1: Overview of the EmoTour dataset.

3.3 Devices and modalities

This section goes over the different wearable devices used in EmoTour, as well as the features extracted for each modality.

3.3.1 Pupil Core eye tracker

The Pupil Core eye tracker [24] was used to record infrared videos of the participant's eyes during sightseeing, as well as to record a video of the participants surroundings from their point of view. The recorded world video was used to extract information about the objects surrounding participants in [4, 5], but these *Object Detection* features were left out in this work, because of their poor performance.

Features were extracted from the eye videos using the Pupil software, introduced in Subsection 2.6.3. Specifically, the spherical pupil coordinates ϕ and θ extracted from every video frame were summarised in two ways: (1) by computing the observed mean and standard deviation for these values, per participant, their field of view was divided into 8 portions, and the percentage of time outside each threshold was computed for every session; (2) by averaging, for every session, the mean and standard deviation of these values, computed for different time windows $\in \{1, 5, 10, 20, 60, 120, 180, 240\}$ seconds with a window offset of $\frac{1}{3}$. In

total, this process results in 80 features characterising eye movement intensity and statistics, which are referred to as *Pupil* features for the remainder of this work.

3.3.2 SenStick sensor board

A SenStick sensor board [38] was mounted on one side of the eye-tracking device, and concentrates a number of different sensors, including an accelerometer, a gyroscope, magnetic, illumination, ultraviolet, temperature, humidity and pressure sensors.

Data from the gyroscope was used to characterise head movements during sessions. By computing the mean and standard deviation of the recorded gyroscope values, participant-specific thresholds for the gyroscope values were defined, and used to count the amount of head tilts in different directions as well as the average time interval and standard deviation between head tilts, for each session. By using the approach defined by [39], data from the accelerometer was used to characterise body movement, namely the amount of footsteps taken by participants during a session, and the average time interval and standard deviation between footsteps.

In total, 28 features were extracted from this sensor board, and referred to as *SenStick* features.

3.3.3 Smartphone

Participants used a smartphone to record brief selfie videos after each session to express their current mood. Both audio and video features were extracted from these videos, and both of these were declined in two variants. The ones referred to as *avg(Selfie)* features include vocal expressions (i.e., low-level audio descriptors) and facial expression (i.e., Action Units) averaged over recordings, for a total of 80 features. These were extracted using the openSMILE [14] and the OpenFace [18] toolkits, described in Subsection 2.6.1 and Subsection 2.6.2, respectively.

Another 72 new, high-level features were derived from the avg(Selfie) features by feeding them into a recurrent neural network with long-short term memory (RNN-LSTM). This network was trained to annotate audiovisual data with arousal and valence scores in a time-continuous manner, as a way to get a "real-time" understanding of the emotional status of participants. This time-continuous labelling involves a certain amount of context, to avoid affective labels to suddenly shift from one opposite to another, which would be unrealistic. The optimal amount of context to consider was studied [40] and fixed to 7.6 s. The RNN-LSTM was pretrained on existing emotion recognition corpora: RECOLA [41], SEMAINE [42], CreativeIT [43] and RAMAS [44]. The trained model was used to annotate the

recordings collected during EmoTour experiments, and these annotations were then summarised through simple functionals on arousal and valence scores, per recording. These derived features are referred to as *high(Selfie)* features.

3.3.4 Empatica E4 wristband

Participants wore an Empatica E4 wristband [45], which was used to measure participants' physiological signals, i.e., their heart rate, electrodermal activity, and body skin temperature. Different functionals were applied to the time series collected for each session to produce a total of 24 *Empatica* features. These functionals include the mean, standard deviation, minimum, maximum, range, as well as the ratio to a participant-specific reference value.

3.4 Labels and ground truths

The collected data was labelled manually by participants, which were asked to input their emotional status and satisfaction level after each session through a simple smartphone application.

Emotional states are defined using Russell's circumplex model of affect [46], which uses a two-dimensional space to arrange emotional states along its two Valence and Arousal axes, as illustrated in Figure 3.1. Participants were asked to pick one of the nine emotions defined through this map.

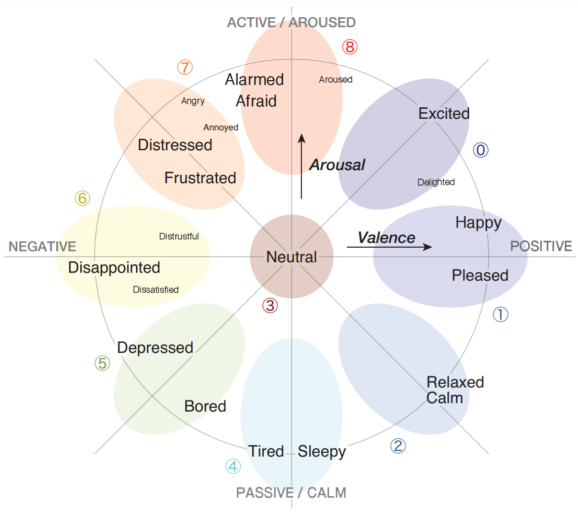

Figure 3.1: Russel's circumplex model of affect[1].

[1]Figure reprinted from [4].

Due to an imbalance in the reported emotions, which renders it hardly possible to perform nine class emotion classification, these emotional labels were streamlined into three more general emotion levels: Positive (includes Excited, Happy/Pleased, Calm/Relaxed), Neutral and Negative (includes Sleepy/Tired, Bored/Depressed, Disappointed, Distressed/Frustrated, Afraid/Alarmed).

Satisfaction levels are defined on a seven-point Likert scale, ranging from 0 (completely unsatisfied) to 6 (completely satisfied), with 3 corresponding to a neutral satisfaction level.

3.5 Model

The EmoTour dataset is used to build two machine learning models, one to estimate tourists' emotions from data through a 3-class classification task, the other one to predict tourists' satisfaction levels as a regression task along a 7-point Likert scale. These models are defined as simple multilayer perceptrons, with an input layer of dimensionality equal to the number of features that are being used, a hidden layer of 30 neurons, and an output layer of 3 neurons for the classification task or 1 neuron for the regression task. Feature-level fusion is applied to the features obtained through the different considered modalities before they are fed to the model.

The performance of these models is evaluated using Unweighted Average Recall (UAR) for emotion estimation and using Mean Absolute Error (MAE) for satisfaction estimation. The aim is to maximise UAR and minimise MAE. Models are trained using 10-fold cross-validation, while a Leave-One-Out methodology is used for testing models, by building a model that is trained on all data but the data relative to a specific participant, and tested on this left-out data, for every participant. Results reported in this work correspond to the average evaluation metric value of all models evaluated on their respective test sets. The baseline performance of models is reported in further chapters, to allow for comparisons with new results.

3.6 Overview

Figure 3.2 gives an overview of the data collection workflow, as detailed in Section 3.3 and Section 3.4. Figure 3.3 summarises the way the features are used to train prediction models, as gone over in Section 3.5.

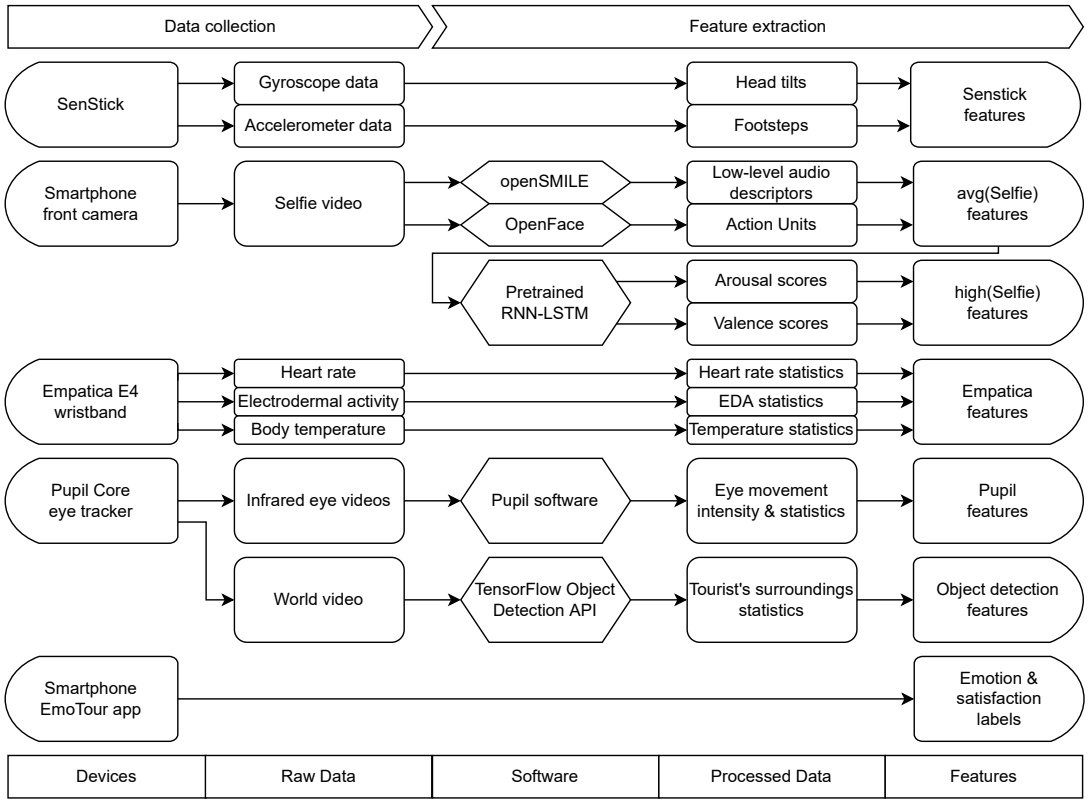

Figure 3.2: Overview of the data collection workflow within the EmoTour system.

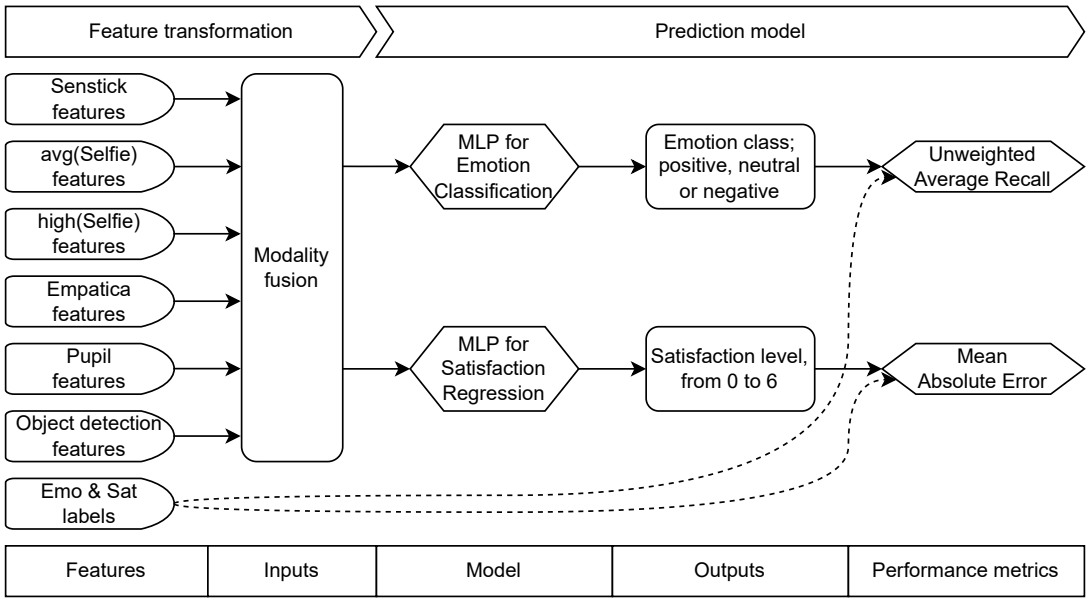

Figure 3.3: Overview of the prediction workflow within the EmoTour system.

Chapter 4

Reducing EmoTour's burden and cost

4.1 Introduction

While multimodal emotion recognition approaches appear intuitively more desirable, as they often yield better results, they don't come without drawbacks. On one hand, they increase the number of sensors needed for data collection, in turn increasing the overall cost of the system. On the other hand, multimodal systems also worsen the burden on the user, who is expected to carry and manage more devices. As the purpose of EmoTour is specifically to lessen the burden on users by passively collecting data, it appears important to assess the relevance of used modalities.

In order to do so, this chapter explores how leaving out the most burdensome or expensive sensors from the original system affects prediction performances. This is expected to give insights on which of its components has the largest impact on prediction results, ultimately allowing to decide whether the load or cost of a particular sensor is justified by the improvements in results it leads to. By considering this, the aim is to improve the usability of the system for real-world applications.

Moreover, Principal Component Analysis (PCA) is applied to the data, both to mitigate the possible drop in performance from leaving out modalities and to get an overview of the features that offer the highest variability. It is used here as a dimensionality reduction method, in order to compare the quality of predictions when varying the data's dimensionality.

4.2 Comparison of modalities

Figure 4.1 shows a participant with the devices used in previous experiments. From this, the Pupil Core eye tracker appears rather burdensome to wear, on top of requiring extensive calibration [5] and coming at the steep price point of 3440€[1]. While the Empatica E4 wristband is hardly considerable as burdensome on users, it is still a research device priced at 1690$[2] (∼1600€). The two other devices, namely the SenStick sensor board and the smartphone are hereafter considered "cheap" sensors. The first device was priced 29,800¥[3] (∼220€) when commercialised, and could likely be replaced with a similar environmental sensor of lower cost. As for the second device, it can reasonably be assumed that most users own one already.

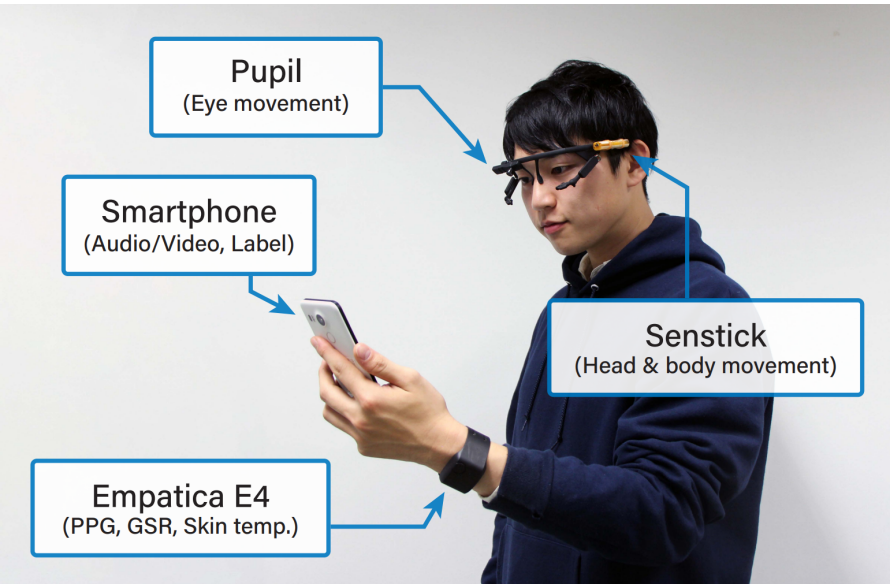

Figure 4.1: Device setup for EmoTour[4].

As such, different feature sets leaving out the data from the Pupil Core eye tracker and the Empatica E4 wristband are considered. In total, 4 feature sets are defined: one using all available features, one using all but the Pupil features, one using all but the Empatica features, and one using only the SenStick & Selfie features, i.e., the cheapest sensors. Each of these sets is declined into 2 variants, depending on whether they use the avg(Selfie) features or the high(Selfie) features.

[1] `https://pupil-labs.com/products/core`
[2] `https://store.empatica.com/products/e4-wristband`
[3] `http://senstick.com`
[4] Figure reprinted from [5].

4.3 PCA dimensionality reduction

To both improve the accuracy of predictions and streamline the high-dimensional dataset, every feature set is standardised and PCA is applied to it before feeding it through the prediction models. To decide an appropriate number of principal components to be considered in the following section, and thus of how much dimensionality reduction to apply, the amount of variance principal components capture is looked at, as illustrated in Figure 4.2.

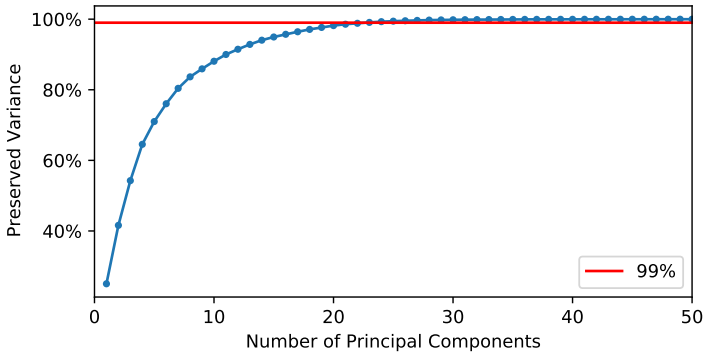

Figure 4.2: Data variance captured by the x first principal components.

From this, it appears that very few principal components suffice to capture most of the variance in the EmoTour data: 16 PC suffice to capture 95% of the variance, 23 PC suffice to capture 99% of the variance, and 34 PC suffice to capture 99.9% of the variance in the data. In the following section, results are hence considered when keeping a number of principal components $\in \{5, 10, 15, 20, 25, 30, 40, 50, 60, 70\}$.

4.4 Results

4.4.1 Performance baseline

Results from previous works are used as a performance baseline and referred to as EmoTour2018 [4] and Fedotov2020 [5]. These results were not obtained on the exact same data/feature sets but were chosen for being the available reported results with datasets that most closely match those considered here. The results reported as EmoTour2018 [4] used only the data collected in the cities of Ulm and Nara, and all feature types but the avg(Selfie) and Empatica features. On the other hand, the reported results for Fedotov2020 [5] use all available feature types (meaning both the avg(Selfie) and high(Selfie) features, described in Subsection 3.3.3), as well as extra Object Detection features, which were dropped here because of their poor performances.

4.4.2 Emotional status predictions

Figure 4.3 reports all the UAR scores for models trained on each of the considered feature sets, when training models that use anywhere between 5 and 70 principal components. Table 4.1 streamlines this information to only show the best UAR scores achieved on each feature set. It appears quite clearly that the use of PCA improves previously achieved emotion prediction performances for all of the considered feature sets. As reported in [5], avg(Selfie) features tend to perform better than their high-level counterparts, making it safe to focus on these more easily obtainable features, i.e., on Figure 4.3(a) and the left column in Table 4.1. It is however interesting to note that feature sets that include high(Selfie) instead of avg(Selfie) appear to benefit most from a drastic dimensionality reduction.

With the intent of reducing the cost of the sensors used in the system, results seem to suggest that leaving out the Pupil features (orange line) actually has a positive impact on predictions, especially when considering 30 (or less) principal components. Indeed, the highest prediction accuracy is obtained when using SenStick, avg(Selfie) and Empatica features and keeping 30 PCs, with an UAR score of 0.602. The cheapest feature set (red line), consisting just of SenStick and avg(Selfie) features, allows for predictions nearly as good when using 60 PC specifically, with an UAR score of 0.600. Figure 4.3(a) does in fact suggest a reduced contribution from Empatica features when 40 or more PCs are kept (comparing the orange with the red line).

Capital modalities for emotional status predictions thus comprise SenStick and Selfie features, with Empatica features enabling a slight performance bonus. Keeping around 60 PCs tends to lead to maximal performances overall.

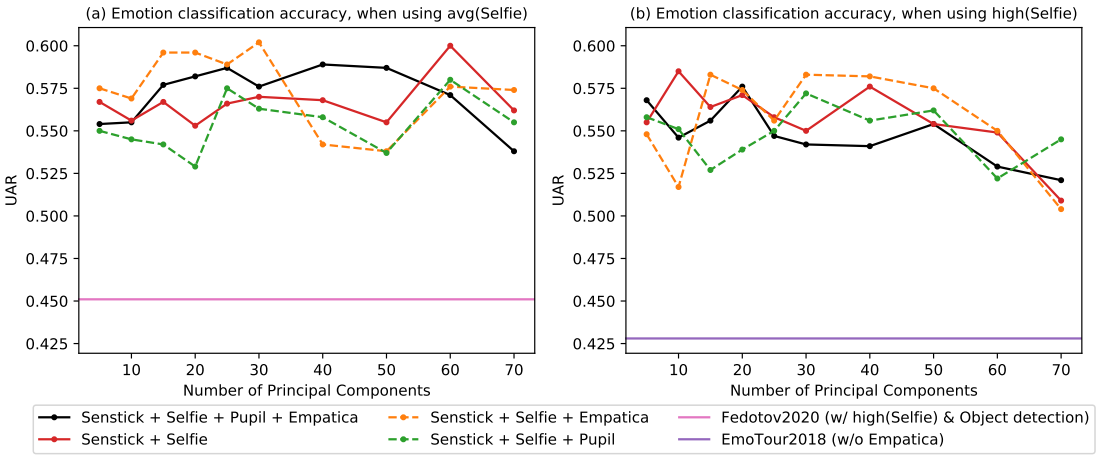

Figure 4.3: Emotional status predictions per feature set and #PCs.

	avg(Selfie)		high(Selfie)	
	#PC	*UAR*	*#PC*	*UAR*
SenStick + Selfie + Pupil + Empatica	40	0.589	20	0.576
SenStick + Selfie + Empatica	30	**0.602**	15	0.583
SenStick + Selfie + Pupil	60	0.580	30	0.572
SenStick + Selfie	60	0.600	10	**0.585**
Previous works*	-	0.451	-	0.428

* Scores reported in Fedotov2020 and EmoTour2018, respectively.

Table 4.1: Best emotional status predictions per feature set, w/ respective #PCs.

4.4.3 Satisfaction level predictions

Figure 4.4 reports all the MAE losses for models trained on each of the feature sets with a number of principal components varying from 5 to 70. Table 4.2 highlights the best MAE loss for every feature set. From these figures, it is less immediately apparent that the use of PCA improves previously achieved satisfaction prediction performances, as all obtained results are systematically worse than those published in [5], but better than those from [4]. Nonetheless, a clear upwards trend is visible in both graphs, suggesting that reducing the number of kept PCs helps to improve performances. Again, avg(Selfie) features consistently perform better than their high-level counterparts, so we focus on those for the rest of our analysis, i.e., on Figure 4.4(a) and the left column in Table 4.2.

Best prediction accuracies are obtained when using all feature types (black line) and keeping 15 to 20 PCs, the latter achieving a MAE of 1.100. Keeping only the cheapest features (red line), i.e., SenStick and avg(Selfie) features, barely worsens results when keeping 15 PCs or less, with 15 PCs scoring a MAE of 1.116. Unlike emotion estimation, it appears that Pupil features (green line) are consistently more important than Empatica features (orange line), with especially good performances at 10 kept PCs, for a MAE of 1.105. When keeping a number of PCs around the threshold of 50, using only SenStick and avg(Selfie) features (red line) consistently and significantly outperforms other feature combinations.

While satisfaction level predictions thus benefit from all modalities, using only the SenStick and Selfie features allows for nearly the same performances. Throwing Pupil features into the mix is slightly helpful but not unmissable. A low number of PCs generally suffices for maximal performances, but keeping as many as 50 works well when using “cheap” features only.

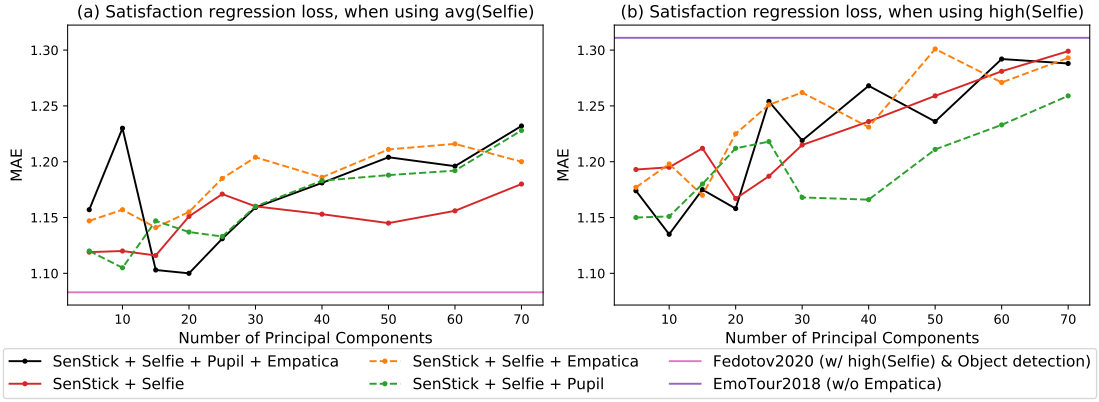

Figure 4.4: Satisfaction level predictions per feature set and #PCs.

	avg(Selfie)		**high(Selfie)**	
	#PC	*MAE*	*#PC*	*MAE*
SenStick + Selfie + Pupil + Empatica	20	**1.100**	10	**1.135**
SenStick + Selfie + Empatica	15	1.141	15	1.170
SenStick + Selfie + Pupil	10	1.105	5	1.150
SenStick + Selfie	15	1.116	20	1.167
Previous works*	-	1.083	-	1.311

* Scores reported in Fedotov2020 and EmoTour2018, respectively.

Table 4.2: Best satisfaction level predictions per feature set, w/ respective #PCs.

4.5 Composition of Principal Components

To get a better understanding of the features that are considered most prominent by PCA, i.e., of the features with highest variability, the composition of the 20 first principal components is illustrated in Figure 4.5. This table shows the 10 most contributing features to each of these PCs, ranked by contribution importance. It appears both Pupil and Empatica features are prominently used by the PCA transformation, with PC1, PC3 and PC16 being majoritarily defined in terms of Pupil features, and Empatica features contributing heavily to many PCs as well. Taken alone, this view would suggest these features provide valuable, highly varied data, whose inclusion is likely to positively affect prediction results on any model using this data.

However, as previously noted in Section 4.4, their omission does not necessarily worsen prediction accuracies, and can sometimes even improve them. This can be attributed in part to the lack of consideration of PCA for the sample's labels, as well as to the biased view offered by Figure 4.5. Indeed, as there are far more

Pupil features than there are, say, SenStick features (80 versus 28, respectively), it is quite natural for them appear more in this table.

Figure 4.6 aims to balance this observation, by plotting the cumulative importance of feature types, i.e., the cumulative sum of the magnitude of their contributions to PCs, normalised by the number of features provided by that feature type, for any number of kept PCs. On this graph, it can be seen Pupil features quickly lose their prominent influence on PCs as their number grows, while Empatica features lose importance at a much steadier rate. Selfie features see their influence vary, while SenStick features consistently augment their influence on PCs.

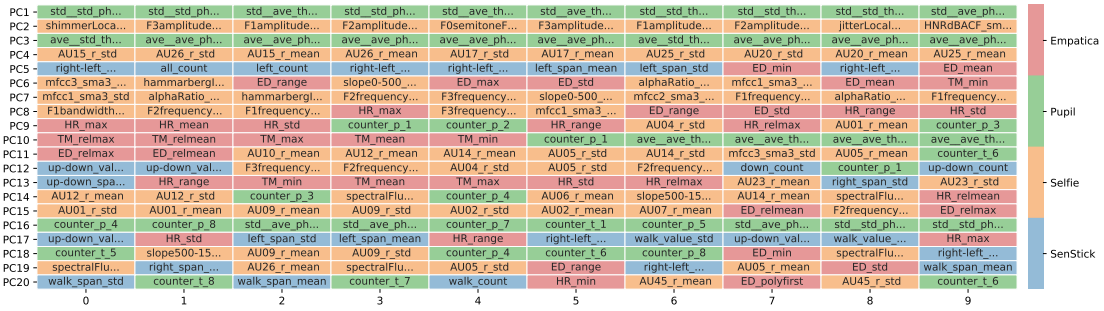

Figure 4.5: Composition of the 20 most important PCs. The table shows the 10 most contributing features to each PC, ranked by the contribution importance.

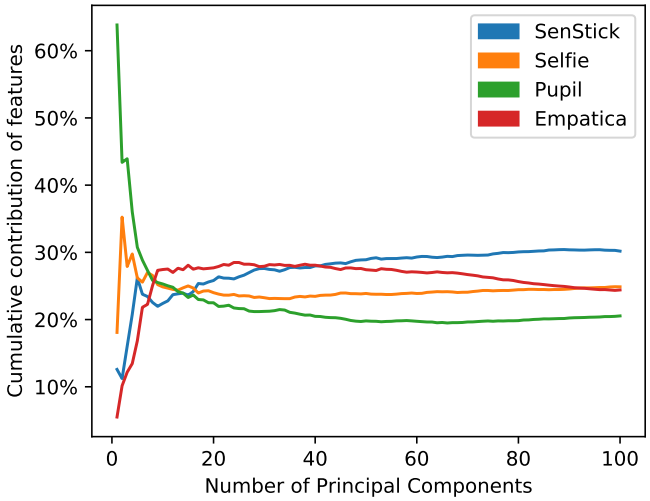

Figure 4.6: Cumulative contribution of each feature type, by number of kept PCs.

4.6 Discussion

In order to reduce the cost and user burden of the EmoTour system, this chapter analysed the effect of leaving out features relative to certain sensors on estimation performances. Expensive sensors such as the Pupil Core eye tracker and the Empatica E4 wristband can be avoided for certain tasks: emotion classification does not necessarily benefit from eye-tracking data, while satisfaction estimation can prove just as good without physiological data. Both tasks can realistically be carried out without either of the feature sets stemming from the more expensive sensors, with predictions that use just the SenStick and Selfie features scoring admirably well.

With regards to the difference between the avg(Selfie) features and the high(Selfie) features, the first type allows for consistently better predictions. This suggests there is no need to compute the high(Selfie) features, which are significantly less straightforward to calculate, as they stem for another machine learning model pretrained on a variety of emotion recognition corpora.

The use of Principal Component Analysis was also investigated, and it appears PCA dimensionality reduction can vastly improve the results of the prediction models used in EmoTour. The optimal level of dimensionality reduction to apply varies between tasks and feature sets; very drastic reductions to $\sim$20 PCs do not necessarily worsen performances much, but keeping 50 to 60 PCs can yield slightly better results.

Overall, the EmoTour workflow can thus be simplified to use less sensors and less features with little to no impact on its ability to classify emotions or predict satisfaction. This allows to discard the use of the expensive Pupil Core and Empatica E4 sensors, the first of which can be reasonably qualified as unpractical to wear.

Chapter 5

Adding privacy to EmoTour

5.1 Introduction

By its ability to understand their emotions and satisfaction, through recognition technologies based on wearable devices, EmoTour could be used with IoT systems to offer tourists a personal experience tailored to their mood. However, to be of practical use in real-world applications, it must be designed in such a way that it can provide precise predictions without leaking any user-specific information. In this chapter, different privacy approaches are explored in an attempt to mitigate the inference capabilities of adversaries on user-collected data.

Specifically, the aim of this chapter is to minimise the ability for attackers to infer sensitive labels such as gender, nationality and age, or to identify a given user from their data. As a real-world system would have the collected data transit through a network, from the end-user's devices to EmoTour's servers, this would minimise the extent of potential privacy breaches.

The prediction tasks considered in this chapter can be classified in two categories: emotion classification and satisfaction regression are intentional tasks, while gender classification, nationality classification, age regression and participant ID classification are adverse tasks. Table 5.1 characterises the predictive ability of models for every task/feature set pair considered here, and serves as a performance baseline for the rest of this chapter. As before, the Cheap features include both the SenStick and avg(Selfie) features, while All features include, as one would expect, all SenStick, avg(Selfie), Empatica and Pupil features.

The most vulnerable feature set clearly appears to be the Selfie features, which achieve the best prediction metrics for all adverse tasks; limiting the sensitive

attribute leakage of these features is therefore a priority of this chapter. It is also very apparent that the multimodal approaches (Cheap and All) allow for high ID classification accuracy, which is something a privacy-aware system would want to limit.

	Intentional		**Adverse**			
	Emotion (UAR)	**Satisfaction (MAE)**	**Gender (UAR)**	**Nationality (UAR)**	**Age (MAE)**	**ID (UAR)**
SenStick	0.546	1.165	0.683	0.162	4.558	0.004
avg(Selfie)	0.552	1.217	0.801	0.470	3.914	0.099
Empatica	0.522	1.346	0.621	0.166	7.096	0.038
Pupil	0.528	1.343	0.509	0.248	4.806	0.053
Cheap	0.547	1.163	0.790	0.409	6.320	0.115
All	0.578	1.234	0.655	0.359	10.490	0.247

Table 5.1: Performance baseline of different feature sets on diverse prediction tasks.

5.2 Permutation importance feature selection

One intuitive idea to reduce the personal data leakage within the system is to identify exactly which features are revealing of sensitive information, and to simply leave them out of the system entirely. A straightforward way to characterise the impact of features on given tasks is to compute their permutation importance, as discussed in Section 2.3, on every task of interest, be it intentional or adversarial. This information can then be used to define some sort of equilibrium between data utility and sensitive information leakage.

While computing the performance metrics reported in Table 5.1, permutation importance coefficients were computed for every feature in every considered model. As the scale of permutation importance coefficients can and will vary between tasks, the coefficients were scaled, by task, to a $[-1, 1]$ range. This scaling was performed by dividing all positive values by the maximum coefficient, and all negative values by the minimum coefficient, in order to preserve the sign of coefficients and thus avoid tampering with the set of features deemed important.

Figure 5.1 illustrates how different prediction tasks on the same feature set, in this case the SenStick features, will benefit differently from each individual feature. The *up-down _span_mean* and *down_span_std* features, for instance, are deemed important for emotion recognition and satisfaction regression, respectively, but do not appear very useful for adverse tasks. On the other hand, the *up-down_val_std* feature only really benefits adverse tasks. Figure 5.2 streamlines this information to simply oppose the importance of features for intentional versus adverse tasks.

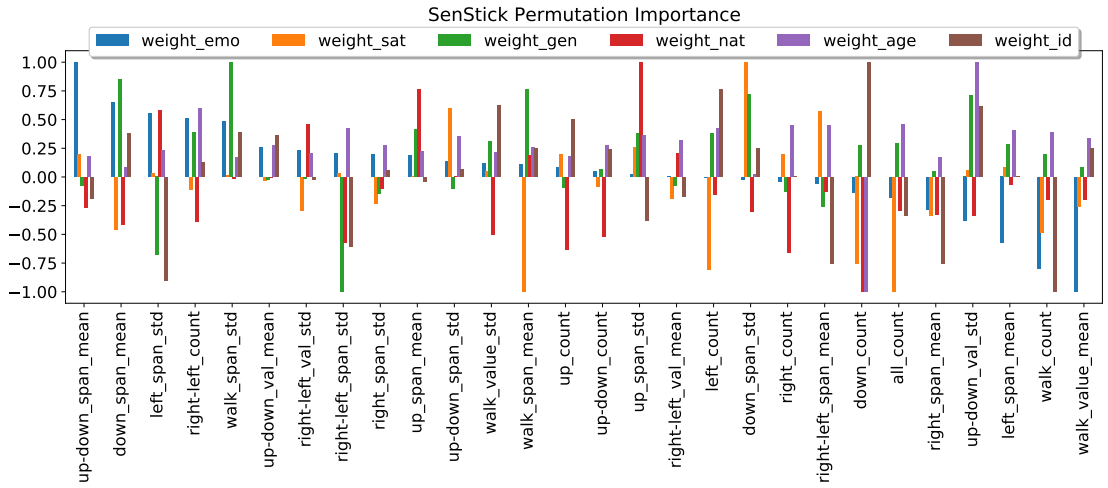

Figure 5.1: Permutation importance coefficients per task for SenStick features.

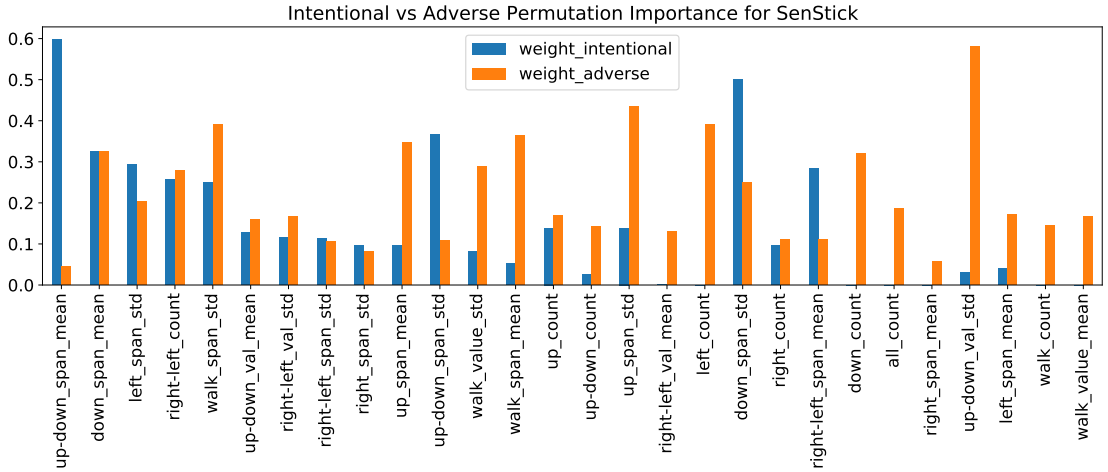

Figure 5.2: Summed SenStick feature permutation importance, per task type.

A feature selection of sorts can be realised by balancing these scaled permutation weights, in order to keep features that appear both important for intentional tasks and not all that important for adverse tasks. If E_k, S_k, G_k, N_k, A_k, I_k are the scaled permutation importance coefficients of feature k for the emotion, satisfaction, gender, nationality, age and identity tasks, respectively, the filtering approach used here is to compute importance_k, as follows, and keep all features for which $\text{importance}_k > 0$. Graphically, the features kept are all those where weight_intentional is larger than weight_adverse, as shown on Figure 5.2. Note that negative permutation importance coefficients are clipped to 0, as a negative

permutation importance coefficient simply indicates a given feature does not contribute to predictions for the task at hand and is not a good indicator of how detrimental a feature is to the task.

$$\begin{aligned} \text{intentional}_k &= 0.5 \cdot max(0, E_k) + 0.5 \cdot max(0, S_k) \\ \text{adverse}_k &= 0.25 \cdot max(0, G_k) + 0.25 \cdot max(0, N_k) \\ &\quad + 0.25 \cdot max(0, A_k) + 0.25 \cdot max(0, I_k) \\ \text{importance}_k &= \text{intentional}_k - \text{adverse}_k \end{aligned}$$

This rather simple approach works decently well, with models trained on reduced feature sets keeping or improving on the intentional performance of models trained on unfiltered feature sets, with slightly reduced adverse performances in some cases. Table 5.2 reports the performance metrics obtained using this approach.

It appears the predictions for intentional tasks barely suffer from this feature selection, and even improve most of the time, while prediction metrics for adverse tasks slightly dip, for some feature sets. The most impressive feat of this approach is to massively reduce the ability for an adversary to identify the participant from which the data is collected, with ID classification's UAR at least halved for every feature set considered. Other adverse tasks do not sport such a convincing performance reduction, perhaps because these 4 adverse tasks kind of cancel each other out within the way features are selected. The avg(Selfie) features, initially the feature set that leaks the most personal information about participants, does however see its adverse performances reduced on almost all fronts, which is encouraging. This likely impacts the observed performances for multimodal feature sets too.

	Intentional		**Adverse**			
	Emotion (UAR)	**Satisfaction (MAE)**	**Gender (UAR)**	**Nationality (UAR)**	**Age (MAE)**	**ID (UAR)**
SenStick	0.509	**1.163**	0.759	0.189	**4.561**	**0.001**
avg(Selfie)	**0.575**	**1.203**	**0.673**	**0.235**	3.400	**0.021**
Empatica	**0.574**	**1.287**	0.741	0.179	6.670	**0.009**
Pupil	**0.537**	**1.289**	0.619	0.287	4.524	**0.022**
Cheap	**0.606**	**1.159**	0.830	**0.375**	5.225	**0.061**
All	0.571	**1.124**	**0.633**	**0.339**	9.559	**0.146**

Table 5.2: Performance after permutation importance feature selection on diverse prediction tasks. Improvements over the baseline performance (i.e., better predictions for intentional tasks and worse predictions for adverse tasks) reported in Table 5.1 are marked in bold.

5.3 Multi-task learning with minimax filters

Another privacy-improving idea explored here is switching the model from a single-task to a multi-task prediction model, where the interests of the data contributor, the data aggregator and potential attackers are considered jointly. The idea is to allow data contributors to distort their data through a minimax filter, as defined in Section 2.4, and have them only upload this distorted data, which is such that it is maximally useful for the intentional tasks the data aggregator is interested in, while minimally useful for the adverse tasks of potential attackers.

As set out at the beginning of this chapter, the prediction tasks considered here are emotion classification and satisfaction regression for intentional tasks, and gender classification, nationality classification, age regression and participant ID classification for adverse tasks. Optimising all 6 of these objectives jointly with minimax filters, by aiming to maximise intentional performance and to minimise adverse performance, the problem becomes the following:

$$\begin{aligned}\min_{\theta_f}[-\lambda(&\rho_1 \max_{\theta_{u_1}}[-f_{emo}(\theta_f, \theta_{u_1})] + \rho_2 \max_{\theta_{u_2}}[-f_{sat}(\theta_f, \theta_{u_2})]) \\ &+(\kappa_1 \max_{\theta_{p_1}}[-f_{gen}(\theta_f, \theta_{p_1})] + \kappa_2 \max_{\theta_{p_2}}[-f_{nat}(\theta_f, \theta_{p_2})] \\ &+\kappa_3 \max_{\theta_{p_3}}[-f_{age}(\theta_f, \theta_{p_3})] + \kappa_4 \max_{\theta_{p_4}}[-f_{id}(\theta_f, \theta_{p_4})])]\end{aligned}$$

In this objective function, f_x is the utility/privacy measure for task x, θ_f represent the data contributor's minimax filter parameters, θ_{u_i} represent the model parameters for intentional task i, θ_{p_i} represent the model parameters for adverse task i, λ is the utility-privacy tradeoff, ρ_i is the relative importance of intentional task i and κ_i is the relative importance of adverse task i.

Hyperparameter	**Values considered**
Apply PCA?	true, false
λ	1, 5, 10, 20, 50
$[\rho_1, \rho_2]^*$	[[1, 1], [1, 2], [2, 1]]
$[\kappa_1, \kappa_2, \kappa_3, \kappa_4]^*$	[[1, 1, 1, 1], [2, 1, 0.2, 2], [3, 1, 0.2, 2], [2, 1, 0.1, 3]]

* These vectors are normalised to sum to 1.

Table 5.3: Hyperparameters considered for the grid search.

Due to the high number of hyperparameters in this learning problem, a grid search was performed as a way to select a model most suited to the problem at hand. Table 5.3 shows the candidate hyperparameters considered in this grid search. A total of 120 models were trained, one for each potential hyperparameter

combination. Models were trained exclusively on avg(Selfie) features, as these are the features leaking the largest amount of sensitive information.

With 6 tasks of interest, it is difficult to precisely determine the overall best-performing model. Table 5.4 lists the 10 "best" hyperparameter combinations found during grid search, where models were ranked relatively to all other models for each specific task ($R(\text{Task})$), from best to worst for intentional tasks and from worst to best for adverse tasks, and where models were ordered according to:

$$\text{Score} = \frac{R(\text{Emo}) + R(\text{Sat})}{2} + \frac{R(\text{Gen}) + R(\text{Nat}) + R(\text{Age}) + R(\text{ID})}{4}$$

From Table 5.4, it immediately appears that these minimax models perform a bit worse than their single-task equivalents in terms of satisfaction prediction (average variation of -0.115), but slightly better with regards to emotion prediction (average variation of $+0.014$). This is rather surprising, as this multi-task model forces all six predictions to stem from a shared internal representation; intuitively, one would think this would have for effect to weaken the prediction power for all tasks of interest.

The results also show worsened prediction metrics for all four adverse tasks. While this dip in performance for gender prediction is rather limited (average variation of -0.024), and it remains fairly easy to infer one's gender from distorted data, nationality and age prediction are rendered much harder by the minimax filter (average variations of -0.233 and $+16.941$, respectively). Most notably, minimax filters render it seemingly impossible to infer the ID of a participant from distorted data, with prediction performance dropping all the way to chance level or lower (average variation of -0.098).

Regarding the choice of hyperparameters, it appears applying PCA to the data prior to feeding it to the multi-objective model has limited utility, or even slightly worsens intentional performances. The utility-privacy tradeoff λ does not need to be very high for a minimax model to acutely reduce adverse performance, with λ values as low as 1 and 5 sufficing in most cases. As for the relative importance coefficients of intentional and adverse tasks, it is difficult to uniquely determine a set better than the others. It is however interesting to note that a very low relative importance for age prediction suffices to completely destroy the ability of attackers to infer age.

	Hyperparameters				Metrics (UAR/MAE)					
	PCA	λ	$[\rho_1, \rho_2]$	$[\kappa_1, \kappa_2, \kappa_3, \kappa_4]$	**Emo**	**Sat**	**Gen**	**Nat**	**Age**	**ID**
1	False	5	[2, 1]	[3, 1, 0.2, 2]	**0.575**	1.325	**0.769**	**0.237**	**20.497**	**0.000**
2	False	10	[1, 1]	[1, 1, 1, 1]	**0.566**	1.339	**0.772**	**0.259**	**21.082**	**0.000**
3	False	5	[1, 1]	[1, 1, 1, 1]	**0.571**	1.343	**0.783**	**0.201**	**21.447**	**0.000**
4	False	5	[1, 2]	[2, 1, 0.1, 3]	**0.577**	1.295	**0.783**	**0.266**	**20.507**	**0.003**
5	False	50	[1, 2]	[3, 1, 0.2, 2]	**0.583**	1.368	**0.758**	**0.257**	**20.076**	**0.000**
6	True	1	[1, 2]	[1, 1, 1, 1]	0.550	1.328	**0.769**	**0.201**	**22.210**	**0.008**
7	False	10	[2, 1]	[1, 1, 1, 1]	**0.555**	1.346	**0.783**	**0.204**	**20.761**	**0.000**
8	False	1	[1, 1]	[3, 1, 0.2, 2]	**0.566**	1.296	**0.783**	**0.263**	**20.526**	**0.003**
9	True	20	[1, 1]	[1, 1, 1, 1]	**0.559**	1.341	**0.783**	**0.275**	**21.382**	**0.000**
10	True	1	[1, 2]	[3, 1, 0.2, 2]	**0.558**	1.336	**0.784**	**0.206**	**20.066**	**0.000**

Table 5.4: Performance of the top-10 multi-objective minimax filter models on diverse prediction tasks, when trained on avg(Selfie) features alone. Improvements over Table 5.1 (i.e., better predictions for intentional tasks and worse predictions for adverse tasks) are marked in bold.

5.4 Noising data for differential privacy

The last approach considered here is the use of differential privacy, described in Section 2.5, as a means to provide a formal privacy guarantee for hypothetical users of the system. This is a more general approach, as it can obfuscate sensitive data beyond predefined "sensitive" attributes.

Differential privacy can be implemented quite easily into the existing system by applying the Laplacian noise mechanism to features prior to feeding them to the prediction model. Using the post-processing property, stating the immunity of differential privacy guarantees to transformations of the data, this is sufficient to guarantee differential privacy over the model's outputs. Precisely, noise n drawn from a Laplacian distribution is added during standardisation $f(x)$ of the input features x, with sensitivity $S(f) = 10$, a reasonable upper bound on the domain size of standard normally distributed data.

Figure 5.3 shows the performance of ϵ-differentially private prediction models trained on avg(Selfie) features for all tasks of interest and ϵ values $\in \{0.01, 0.1, 0.5, 1, 5, 10, 15, 20, 25, 30, 40, 50, 75, 100, 150\}$. It appears differential privacy can indeed decrease adverse task prediction performance for ϵ values as high as 50, as 75 is the first value for which one of the adverse tasks reaches performances on par with those observed without any privacy mechanism. As expected, differential privacy also decreases prediction performance for intentional tasks, but only until ϵ values between 25 and 30. This indicates applying ϵ-differential privacy with $\epsilon \in [25, 50]$ can slightly decrease adverse performance with little to no impact on

intentional performance. ID classification is rendered practically impossible for all 15 levels of differential privacy tried here. Note that both emotion and gender classification display an odd drop in classification accuracy when $\epsilon = 1$, for which I could not identify a reasonable cause despite further investigation of the data and code.

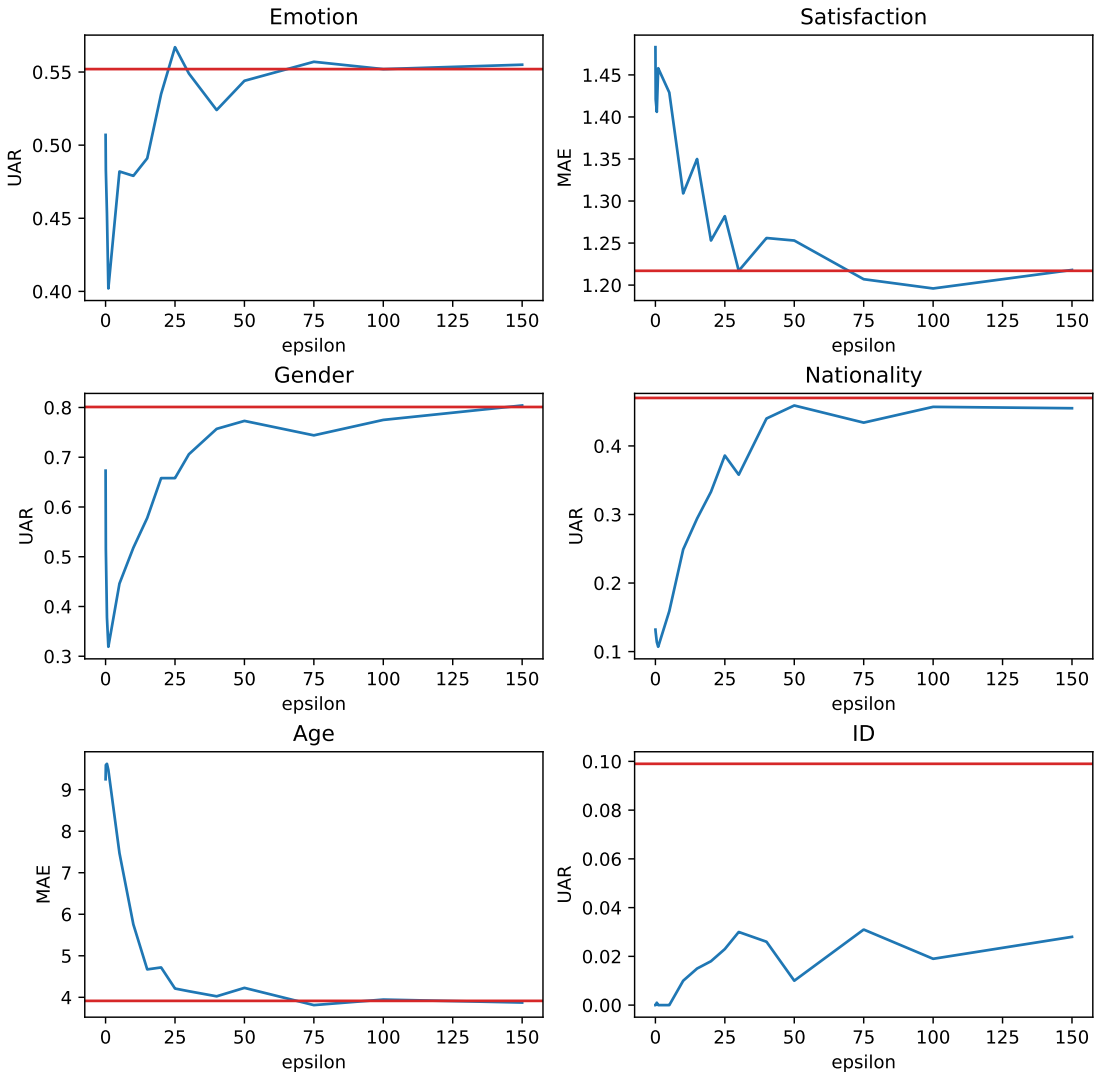

Figure 5.3: Performance comparison for models trained with ϵ-Differential Privacy on avg(Selfie) features for varying ϵ values. Red lines indicate the baseline performance reported in Table 5.1.

5.5 Discussion

Through different privacy approaches, this chapter has confirmed the possibility to limit the leakage of sensitive attributes in the data collected from tourists in the EmoTour dataset, especially the selfie data. With the right parameters, it is possible to use these approaches to notably decrease the ability of attackers to predict sensitive attributes from the data with little to no impact on the prediction performances of intentional tasks.

These three privacy possibilities for EmoTour offer different types of privacy. Feature selection with permutation importance slightly reduces the ability to predict gender, and renders nationality and participant ID classification impossible. Multi-task learning with minimax filters additionally renders age prediction close to impossible. Noising data for differential privacy is the only possibility offering a formal privacy guarante and capable of reducing gender classification to chance-level. It also renders nationality and participant ID predictions unusable, but it fails at noticeably obfuscating age and has a more important impact on intentional prediction performances.

Chapter 6

Experiment at Musée L

6.1 Introduction

This final chapter details the premise, setup and results of a new data collection experiment realised at Louvain-la-Neuve's Musée L during Spring 2022. My work being based on a ready-to-use dataset so far, this experiment had for objectives both to confirm my proposed changes to the system with a new dataset, and to familiarise myself with the data collection and feature extraction aspect of the EmoTour system.

Musée L is the UCLouvain's university museum, "where scientific collections and works of art are combined in dialogue, for unprecedented emotions"[1]. Its inventory counts 20000 items, which include natural history specimens, ethnographic objects, scientific inventions and works of art, part of them displayed on over 2500m^2 in one of the city's most emblematic buildings.

This location was picked to contrast with previous locations; as the setting is no longer outdoors sightseeing but an indoors environment with limited space, this allows to explore how specific model performance is to the type of location. Knowing how well a system developed for emotion sensing during sightseeing can carry over to other smart city applications would help to assess the emotion recognition ability of EmoTour in more diverse applications. Musée L is also a location many students have never visited, making it a better candidate location than Louvain-la-Neuve itself, which coincidentally lacks sights of interest to those living there.

[1]Direct quote from `https://museel.be/en/about-museum`

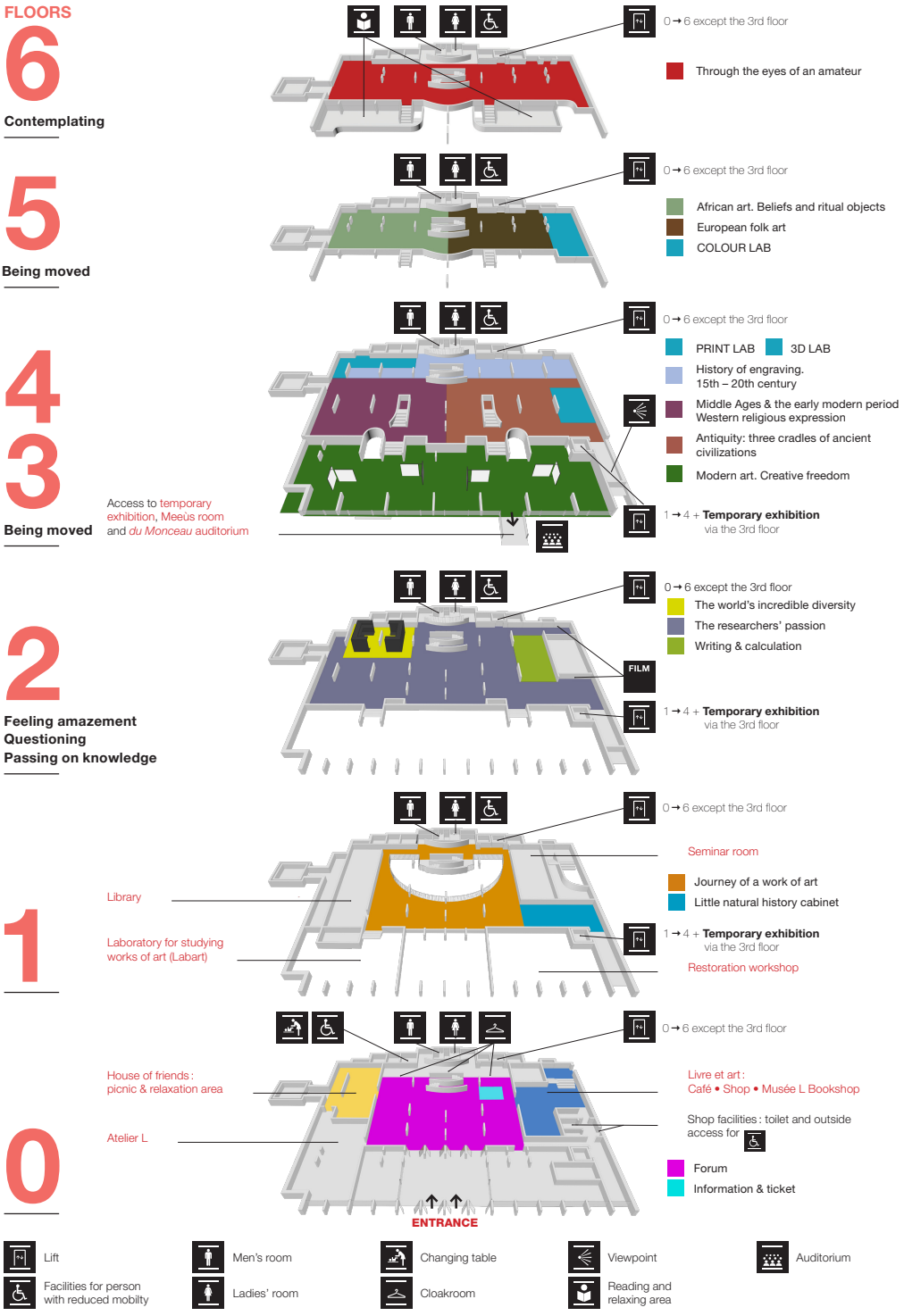

Figure 6.1: Floor plan of the museum, courtesy of Musée L.

6.2 Experimental setup

6.2.1 Sessions

As with previous EmoTour experiments, the experiment location was split into different sessions for participants to review individually. These sessions more or less match the structure of the building, with each floor amounting to one session. The exceptions to this are the very short first floor, which was grouped with the second floor, and the fifth floor, which was split in two different sessions, as it features two radically different exhibitions. Figure 6.1 shows Musée L's floor plan, as handed out to visitors of the museum, while Table 6.1 summarises the chosen sessions and their contents.

Session	**Exhibitions**
Floors 1 & 2	Natural history cabinet, cabinet of curiosities, UCLouvain researchers, writing and calculation
Floor 3	Modern art
Floor 4	Antique and medieval art, history of engraving
Floor 5 (a)	African art
Floor 5 (b)	European folk art
Floor 6	Charles Delsemme collection

Table 6.1: Sessions delimited within Musée L.

6.2.2 Devices

Taking into account the results from Chapter 4, two data collection devices were used in this experiment: a smartphone (iPhone X or Xs) and an environmental sensor (2JCIE-BL01 Omron sensor [47]). The smartphone was used to collect selfie videos, podometer data and labels from participants, while the environmental sensor replaces the SenStick sensor used in previous experiments, and includes temperature, humidity, light, UVI, pressure, noise and acceleration sensors.

Participants were asked to wear a belt bag equipped with these devices; the environmental sensor was attached on the outside of this belt bag, while the inside was used to carry the smartphone. The smartphone came with the EmoTour+ app, a simple data-collecting application developed by the Ubiquitous Computing Systems Laboratory. After every session, participants used the application to retrieve the data from the environmental sensor, to record themselves on video, to answer a small questionnaire and to upload all this data to a Firebase.

The questionnaire was used as a way to ground-truth the data, and included the same multiple-choice questions as those used in previous EmoTour experiments, as presented in Figure 6.2:

1. **How do you feel after this session?**: Multiple-choice question with 9 possible answers, 3 of which positive, 1 neutral, and 5 of which negative.
2. **What is your satisfaction level after this session?**: Likert-scale from 0 to 6, where 0 is fully unsatisfied, 3 is neutral, and 6 is fully satisfied.
3. **How would you rate your Touristic Experience Quality [5]?**: Two-dimensional evaluation space plotting "interest" on one axis and "meeting expectations" on the other. The x-axis ranges from "not interesting / nothing to see" (-3) to "very interesting" (3), while the y-axis ranges from "did not get what expected" (-3) to "got what expected" (3).

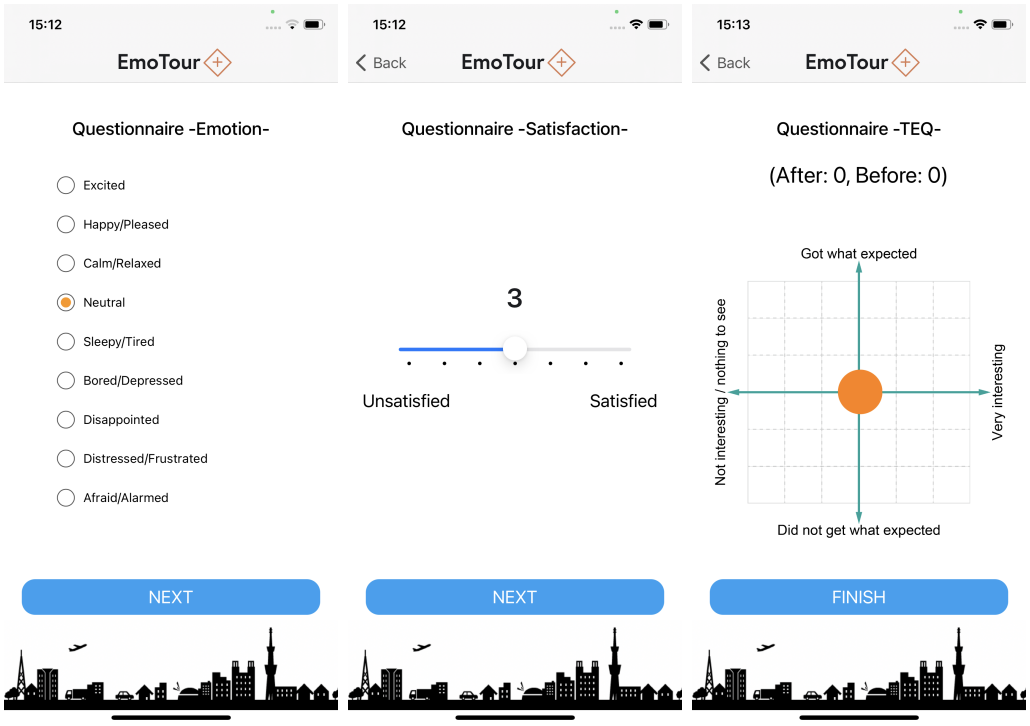

Figure 6.2: EmoTour+ questionnaire, filled in by participants after every session.

6.2.3 Participants & missing data

A total of 25 participants visited Musée L in the context of this experiment, the vast majority of them for the very first time. The distribution of participants was

as follows: 18 men, 7 women; 19 students in their twenties, 6 adults aged in their forties/fifties; 21 Belgian, 4 other nationalities. Together, their data amounts to 150 sessions of data, some of which were unfortunately lost due to technical issues I was not able to identify.

While the selfie videos made it for most sessions, i.e., 138 out of 150 sessions, the collected dataset presents huge gaps in smartphone data, which was only acquired for 93 sessions, and environmental data, which is only available for 51 sessions. Questionnaire answers, i.e., labels, only made it for 94 sessions; the remaining 44 labels were recovered by manual labelling of the videos, either by the participants themselves or by myself.

While the lack of consistent availability of smartphone and environmental data seriously affects the scale of this experiment, the labelled selfie videos alone have previously proved to be one of the modalities most central to the system. It is this data that is hence discussed for the rest of this chapter.

6.2.4 Feature extraction

As with previous experiments, features were extracted from the raw selfie videos using the openSMILE [14] and OpenFace [18] softwares. openSMILE 3.0 was used with its `eGeMAPSv01b.conf` configuration file and time windows of 60ms, each of them offset by 10ms with respect to the previous one, as described in Subsection 2.6.1 and the original EmoTour paper [4]. The same feature subset as in previous experiments was kept, and values were averaged per recording. OpenFace 2.2.0 was used to extract the magnitude with which it detects one of the 17 Action Units listed in Subsection 2.6.2 in each of the selfie videos' frames. Frames without any detected Action Unit were discarded, and the rest of the values were averaged per recording.

6.3 Prediction performances

Training the different prediction models touched upon throughout this work on this new dataset yields slightly underwhelming results. These results are summarised in Table 6.2, in contrast with those achieved on the previous EmoTour dataset.

The most jarring difference between the two datasets is their ability to correctly classify emotions. Where the previous dataset easily achieves an UAR of around 0.55 on emotion classification tasks, the Musée L datasets hovers at an unimpressively low UAR of 0.4. Separating the selfie data into its audio and video components, it appears it is mostly the audio component of the data that appears unhelpful for

classification, while the video component alone scores slightly better than both components together.

Satisfaction predictions are also down a bit, but remain rather similar, as do gender predictions. Nationality classification performances are much higher, which is to be expected, seeing how the 3-class task was simplified to a 2-class task. Age predictions with the Musée L are completely unusable, to levels which were only observed for the minimax model on the previous dataset. Finally, ID predictions are definitely worse too, especially since Musée L only has a 25-class ID classification task, against a 47-class task for the previous dataset.

As for the differences between models, when compared to the baseline model (model A), it appears PCA dimensonality reduction (model B) does slightly improve Musée L predictions, exactly as with the previous dataset. Using permutation importance feature selection (model C) hardly increases privacy, unlike the previous dataset which benefited more from this approach. The performances of the minimax model (model D) are comparable to those observed previously, at least on the emotion, satisfaction, age and ID fronts. The use of differential privacy (model E) is the only privacy approach consistently offering some level of attribute obfuscation, albeit to a relatively mild degree.

		Musée L						**Ulm + Nara + Kyoto**					
Model	**Input**	**Emo**	**Sat**	**Gen**	**Nat***	**Age**	**ID†**	**Emo**	**Sat**	**Gen**	**Nat***	**Age**	**ID†**
A	V+A	0.357	1.302	0.762	0.763	22.107	0.036	0.552	1.217	0.801	0.470	3.914	0.099
	Video	0.405	1.293	0.471	0.840	21.476	0.017	0.541	1.256	0.625	0.287	3.640	0.019
	Audio	0.370	1.298	0.877	0.787	21.915	0.009	0.559	1.252	0.835	0.581	3.757	0.010
B	V+A	0.387	1.274	0.732	0.767	21.693	0.026	0.564	1.258	0.706	0.485	3.610	0.020
	Video	0.411	1.273	0.518	0.817	21.390	0.024	0.548	1.254	0.610	0.285	3.457	0.023
	Audio	0.391	1.287	0.747	0.797	21.263	0.018	0.562	1.262	0.780	0.505	3.764	0.011
C	V+A	0.373	1.267	0.796	0.807	21.408	0.003	**0.575**	**1.203**	0.673	**0.235**	3.400	0.021
	Video	0.420	1.248	0.571	0.817	21.338	0.003	0.560	1.268	0.633	0.259	3.451	**0.000**
	Audio	0.395	**1.236**	0.773	0.840	21.261	0.010	0.555	1.282	0.729	0.417	3.514	0.004
D	V+A	0.387	1.429	0.643	0.840	**25.701**	0.009	**0.575**	1.325	0.769	0.237	20.497	**0.000**
	Video	**0.425**	1.412	0.673	0.840	25.531	**0.002**	0.541	1.356	0.783	0.242	20.894	0.003
	Audio	0.420	1.352	0.629	0.840	25.539	0.004	0.547	1.374	0.783	0.346	**21.092**	**0.000**
E	V+A	0.317	1.365	0.669	**0.740**	21.550	0.025	0.567	1.282	0.658	0.386	4.211	0.023
	Video	0.362	1.339	**0.468**	0.817	21.468	0.006	0.533	1.276	**0.579**	0.250	3.640	0.013
	Audio	0.356	1.277	0.755	0.757	22.013	0.009	0.522	1.270	0.693	0.359	4.085	0.010

* This is a 2-class task for the Musée L dataset (Belgian or other), and a 3-class task for the previous dataset (Japanese, Russian, or other).

† This is a 25-class task for the Musée L dataset, and a 47-class task for the previous dataset.

A: MLP with 30-neuron hidden layer (original EmoTour model, as described in Chapter 3).

B: MLP with 30-neuron hidden layer and PCA dimensionality reduction to 15 PCs (as laid out in Chapter 4).

C: MLP with 30-neuron hidden layer and permutation importance feature selection (as in Section 5.2).

D: Multi-task MLP with minimax filter and 30-neuron hidden layer (with hyperparameters of model #1 in Table 5.4).

E: MLP with 30-neuron hidden layer and ϵ-differential privacy with $\epsilon = 25$ (as described in Section 5.4).

Table 6.2: Performance of the Musée L dataset on various prediction tasks, compared with that of the previous dataset. Column-wise best (or worst) metrics for intentional (respectively, adverse) tasks are marked in bold.

6.4 Differences with previous data

Observing these results, one might ask themselves why these results differ so much between datasets and precisely how the new data differs from previous data. This section compares the dataset collected in Musée L with the dataset collected during experiments in Ulm, Nara and Kyoto.

6.4.1 Label distribution

Figure 6.3 and Figure 6.4 oppose the distribution of emotion and satisfaction labels in both experiments. A contrast can be seen in the amount of neutral/negative labels, both in terms of emotion and satisfaction, which are slightly more numerous in the new experiment. Overall, the label distributions remain similar, with both experiments presenting an imbalance towards positive labels.

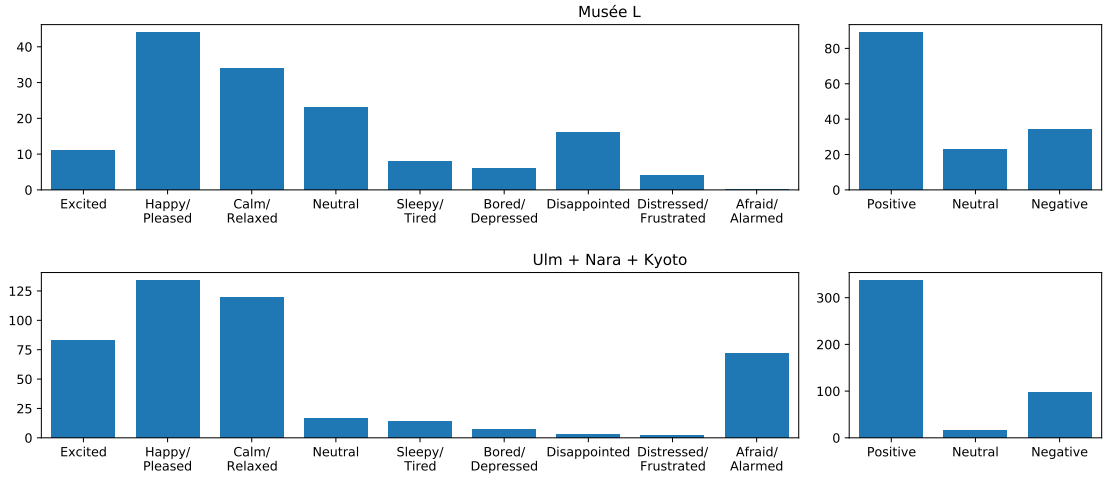

Figure 6.3: Comparison of emotional label distributions between experiments, in terms of the 9 in-app answers, as well as in terms of the 3 general emotion classes used within prediction models.

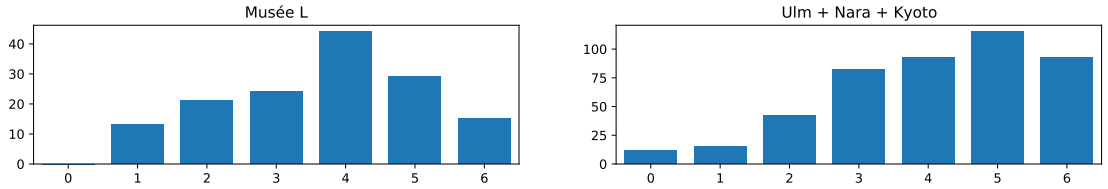

Figure 6.4: Comparison of satisfaction label distributions between experiments.

6.4.2 Data distribution

Figure 6.5 shows a visualisation of the data in all 4 experiments using the Uniform Manifold Approximation and Projection (UMAP) technique for dimension reduction [48]. What is striking from this data representation is how vastly different the Musée L audio data samples are from previous audio data samples. While the Musée L video data samples appear largely similar to previous video data samples, the Musée L audio data samples are severe outliers to previous audio data samples. This could explain the poor prediction performances observed on Musée L data samples, which are indeed worse than those observed on previous data samples, especially when using the audio features.

While the audio setting is indeed very different within a museum, especially when compared to the audio one would get when recording outside, surrounded by other people, it is still surprising how much the audio samples differ from each other. Perhaps this is due to the tendency of people to lower their voice in museums, which might cause people to restrain themselves when speaking, and thus have their speech convey less emotion than if they felt free to raise their voice. To my knowledge, the software and settings used to extract audio features from the selfie videos are the exact same as in previous experiments, but the extent of these differences may suggest some discrepancy in the extraction method.

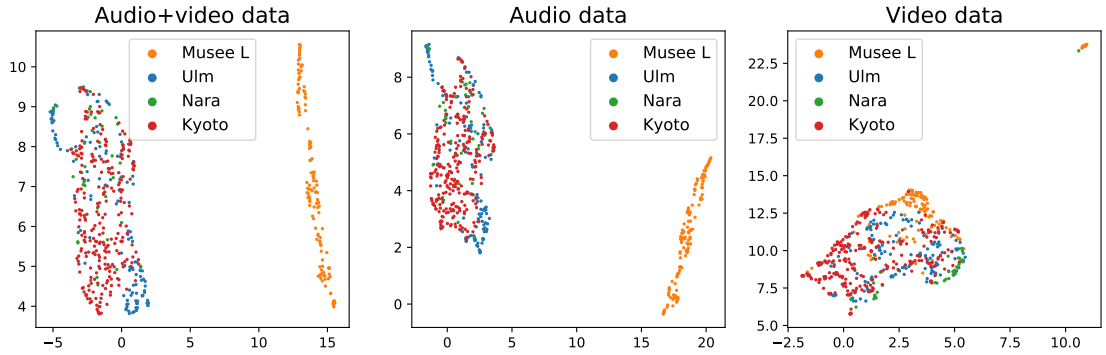

Figure 6.5: UMAP visualisation of EmoTour data, colour-coded by experiment.

6.5 Subgroup analysis

Subgroup analysis [49] is a widely used technique in the medical field aiming to explore how a shared characteristic between some participants in a study can affect their reaction to a given treatment. Beyond the medical field, it allows to explore how considering different subgroups within the participant pool might lead to

different results when applying given data analysis methods to them, independently of the other subgroups.

The original EmoTour paper [4] noted that emotion and satisfaction estimation accuracy can vary according to the cultural backgrounds of participants. In order to build onto that observation, this section extends this subgroup analysis to all plausible subgroups that can be delimited within the combined Ulm + Nara + Kyoto + Musée L dataset. These subgroups are defined with respect to the nationality, gender and age of participants, as well as relatively to the experiment location.

6.5.1 Subgroups

The subsets considered in this section are as follows:

- **Gender:** Male (464 samples) and Female (134 samples).
- **Nationality:** Japanese (187 samples), Russian (106 samples), Belgian (122 samples) and Other (183 samples). The last subgroup includes data samples from 13 different, mostly Asian, nationalities.
- **Age:** 20-23 years old (227 samples), 24-27 years old (205 samples), 28+ years old (166 samples). These subgroups were chosen mostly for their approximately even distribution of data samples.
- **Location:** Ulm (149 samples), Nara (40 samples), Kyoto (263 samples) and Musée L (138 samples).

6.5.2 Results

Each subgroup was used to train 6 different models, half of them tasked with estimating emotional status, the other half tasked with estimating satisfaction level. Models use only the avg(Selfie) features and all apply some level of PCA dimensionality reduction. Table 6.3 reports the prediction scores obtained on these subgroups, in order to get a feel for how different subgroups may express emotions differently.

Results by gender

As one might expect, the smallest of the two groups, containing only Female samples, leads to noticeably worse model performances in all cases. Reasonable emotion classification on Female samples however remains possible, which can hardly be said for satisfaction prediction. Both subgroups perform best with a drastic dimensionality reduction.

	15 PCs		25 PCs		40 PCs	
	Emo (UAR)	**Sat (MAE)**	**Emo (UAR)**	**Sat (MAE)**	**Emo (UAR)**	**Sat (MAE)**
All	0.508	1.270	0.490	1.249	0.484	1.276
Male	**0.524**	1.260	0.482	**1.247**	0.484	1.298
Female	0.476	1.449	0.452	1.476	0.447	1.680
Japanese	0.513	1.238	0.475	1.312	0.457	1.247
Russian	0.502	1.620	0.464	1.741	0.475	1.644
Belgian	0.370	1.318	0.331	1.287	0.346	1.317
Other	0.543	1.263	0.545	**1.139**	**0.589**	1.160
20-23	0.455	**1.229**	0.425	1.256	0.389	1.243
24-27	0.548	1.262	**0.558**	1.257	0.541	1.347
28+	0.454	1.407	0.463	1.359	0.432	1.421
Ulm	0.579	1.284	0.546	1.381	**0.584**	1.387
Nara	0.540	1.629	0.490	1.935	0.487	2.348
Kyoto	0.571	1.233	0.579	1.242	**0.584**	**1.219**
Musée L	0.410	1.305	0.374	1.297	0.338	1.302

Table 6.3: Prediction performances of different subgroups on various models. The best metrics for each subgroup category, per task, are marked in bold.

Results by nationality

Surprisingly, it is the subgroup with no particular dominant nationality that performs best for both tasks, by a noticeable margin. Prediction metrics on the Japanese and Russian subgroups are still good in general, with the exception of satisfaction prediction on Russian samples, which yields excessively bad results. As Belgian samples are entirely the result of the experiment at Musée L, with the rather disappointing performances reported in Section 6.3, it was to be expected that this subgroup would perform worse than the others, but this subpar performance is apparently limited to the emotion classification task. An important level of PCA dimensionality reduction is beneficial for pretty much all subgroups, except the one composed of other nationalities.

Results by age

Age-wise, it is the 24-27 years old subgroup that outperforms the others in emotion classification, while its satisfaction prediction metrics remain close to those of the 20-23 years old subgroup. The 28+ subgroup performs more poorly than the others on satisfaction estimation, perhaps because of its much less precise age range. PCA dimensionality reduction has varying impact on the predictions realised on these subgroups, with no clear tendency.

Results by location

Again, Musée L samples perform noticeably worse than the other sample subgroups, especially regarding emotion classification, for reasons that remain hard to identify. The Nara samples alone are too few to accurately train models, as could be expected, and systematically perform worse than the Ulm and Kyoto subgroups. Kyoto and Ulm samples are almost on the same performance level, with the Kyoto samples slightly outperforming the Ulm samples. This could be explained by the fact that these were collected on three consecutive days, as opposed to the other subgroups [5]. The subgroups with worst performances, i.e., Nara & Musée L, are those that benefit most from a high level of PCA dimensionality reduction, while this has a limited impact on the Ulm and Kyoto sample subgroups.

6.6 Discussion

The Musée L experiment has allowed the extension of the EmoTour dataset to an entirely new setting, that of an indoors museum, and with a new group of Belgian participants. Serious gaps in the collected data have made it necessary to restrain the scope of the experiment, by limiting the resulting dataset to include only labelled videos. This was shown to lead to somewhat disappointing results, with the new data samples always underperforming with respect to previously collected samples. This could be explained in part by the jarring difference in the values of audio features between both datasets, or might indicate that the workflow developed for EmoTour is simply less suited for emotion recognition in indoors museum conditions.

The new dataset however succeeds at underlining the improvements in performances that come from using PCA dimensionality reduction, which perhaps remain of limited magnitude but show nonetheless that PCA is always beneficial to predictions. The data also allows to confirm the proposed privacy approaches, which offer different sorts of privacy covers at a minimal cost in intentional performances.

A subgroup analysis using all available labels was conducted on the newly extended EmoTour dataset, but the insights that can be taken away from it remain limited, as subgroups remain largely unbalanced, and Musée L samples might artificially worsen the performance of certain subgroups. Further experiments focusing on collecting more female samples or samples that do not fit this simplistic binary opposition, or with larger age gaps between participants, would be beneficial to deepen this aspect of the analysis.

Chapter 7

Conclusion

With the aim of improving the EmoTour system, which multimodally senses tourists in order to estimate their emotional status and satisfaction level for context-aware smart tourism applications, this work has focused on a couple of questions expected to improve the applicability of the system to a non-research context. Specifically, the aim has been to simplify the project's data collection flow, to reduce the burden on users and the overall financial cost of the system, as well as to increase the privacy users can expect from such a system. On top of this, a new data collection experiment was conducted within the UCLouvain's Musée L in Louvain-la-Neuve, for evaluation of the proposed changes and comparison with previous experiments.

The burden and cost reduction aspect was evaluated by comparing prediction performances with and without EmoTour's most expensive sensors, namely the Pupil Core eye tracker and the Empatica E4 wristband. Results suggest their impact on performances is limited, with their inclusion only minimally boosting prediction metrics, especially when applying PCA dimensionality reduction on input data. It thus appears EmoTour can work using only audiovisual (Selfie) and environmental (SenStick) data from users, which allows for a system with reduced user burden and cost.

The privacy aspect was evaluated through 3 different approaches, each of them using slightly different methods and goals. The first of those aimed to identify features that leak the most information on users' sensitive attributes, and to decide of an appropriate feature selection trade-off between data usefulness for intentional tasks and sensitive attribute leakage. This simple approach based on permutation importance has been shown to minimally affect or even positively improve performance on intentional tasks, but its ability to hide sensitive attributes remains limited.

The second privacy approach redefined the system as a multi-objective optimisation problem, where users are allowed to distort their data through a minimax filter, such that it is maximally useful to emotion and satisfaction predictions, while minimally useful to potential adversaries attempting to uncover sensitive attributes or perform data reidentification. This was perhaps the most successful privacy technique used here, vastly reducing the risk for adversaries to recover sensitive attributes from data with very limited impact on intentional prediction performances.

The last privacy technique explored was the insertion of a Laplacian noising mechanism during feature preprocessing, therefore providing users with a formal privacy guarantee. This provides a broader form of privacy, as it does not rely on predefined sensitive attributes, but the resulting utility-privacy tradeoff ends up rather linear: an important increase in privacy is only possible at the cost of an import drop in utility. Nevertheless, slight privacy improvements can be achieved with low levels of noise, so without much impact on intentional task predictions.

Finally, a fresh, local dataset was collected within Louvain-la-Neuve, the first EmoTour dataset to feature extensive European data, and in a new setting, that of a museum. A total of 25 participants agreed to visit Musée L, good for a $\sim$28% increase of the size of the EmoTour dataset. A preliminary evaluation of the distribution and performance of this data was conducted for the various models explored in this work, as well as a comparison with previous data. A subgroup analysis was also conducted on the entirety of the newly extended EmoTour dataset.

The results presented here suggest that collecting audio recordings of similar quality across different sites and settings is challenging; perhaps this issue could be alleviated through some form of audio preprocessing, be it by normalising the audio recordings or by applying data augmentation techniques to increase the amount of samples across different settings. Investigating the use of Convolutional Neural Networks, which have proved well-suited to audio classification tasks [50, 51], might also increase the quality of predictions.

Differences observed between tourists from heterogeneous backgrounds would require further, more systematic and extensive data collection to be leveraged in a significant manner, in the form of background-specific prediction models. The current dataset remains too small in scope and too biased towards certain backgrounds to offer such possibilities.

Implementing EmoTour into a publicly available and user-friendly application, for use in a concrete smart tourism or smart city setting, could allow to touch a wider range of people. Such an application would need to implement a robust privacy mechanism, which could build onto those explored here, and could offer users recommendations on particular sights and locations they'd be expected to enjoy.

Bibliography

[1] Statista, "Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2030." `https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/`. (accessed March 7th 2022).

[2] P. Lee, W. C. Hunter, and N. Chung, "Smart Tourism City: Developments and Transformations," *Sustainability*, vol. 12, no. 10, 2020.

[3] "International Tourism Highlights, 2020 Edition," 2021.

[4] Y. Matsuda, D. Fedotov, Y. Takahashi, Y. Arakawa, K. Yasumoto, and W. Minker, "EmoTour: Estimating Emotion and Satisfaction of Users Based on Behavioral Cues and Audiovisual Data," *Sensors*, vol. 18, no. 11, 2018.

[5] D. Fedotov, *Contextual Time-Continuous Emotion Recognition Based on Multimodal Data*. PhD thesis, Ulm University, 2020.

[6] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain.," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[7] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[8] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of Educational Psychology*, vol. 24, pp. 498–520, 1933.

[9] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

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

[11] M. Jaiswal and E. Mower Provost, "Privacy Enhanced Multimodal Neural Representations for Emotion Recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 7985–7993, Apr. 2020.

[12] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," vol. 3876, pp. 265–284, Jan. 2006.

[13] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, pp. 211–407, Aug. 2014.

[14] F. Eyben, M. Wöllmer, and B. W. Schuller, "openSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pp. 1459–1462, ACM, 2010.

[15] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. R. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *INTERSPEECH*, 2013.

[16] A. Dhall, G. Sharma, R. Goecke, and T. Gedeon, "EmotiW 2020: Driver Gaze, Group Emotion, Student Engagement and Physiological Signal based Challenges," *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020.

[17] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[18] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *13th IEEE International Conference on Automatic Face and Gesture Recognition*, FG 2018, pp. 59–66, 2018.

[19] T. Baltrušaitis, P. Robinson, and L. P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision*, WACV 2016, pp. 1–10, Mar. 2016.

[20] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," *Procedia Computer Science*, vol. 108, pp. 1175–1184, Jun. 2017. International Conference on Computational Science.

[21] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC 2019, pp. 3–12, Association for Computing Machinery, 2019.

[22] T. Baltrusaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic Action Unit detection," pp. 1–6, May 2015.

[23] P. Ekman and W. V. Friesen, "Facial Action Coding System: A Technique for the Measurement of Facial Movement," *Consulting Psychologists Press*, 1978.

[24] M. Kassner, W. Patera, and A. Bulling, "Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction," in *Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp 2014 Adjunct, pp. 1151–1160, ACM, 2014.

[25] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, "CorrNet: Fine-Grained Emotion Recognition for Video Watching Using Wearable Physiological Sensors," *Sensors*, vol. 21, no. 1, 2021.

[26] A. Mengual-Recuerda, V. Tur-Viñes, D. Juárez-Varón, and F. Alarcón-Valero, "Emotional Impact of Dishes versus Wines on Restaurant Diners: From Haute Cuisine Open Innovation," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 7, no. 1, 2021.

[27] J. Alegre and J. Garau, "Tourist Satisfaction and Dissatisfaction," *Annals of Tourism Research*, vol. 37, no. 1, pp. 52–73, 2010.

[28] C. F. Chen and F. S. Chen, "Experience quality, perceived value, satisfaction and behavioral intentions for heritage tourists," *Tourism Management*, vol. 31, no. 1, pp. 29–35, 2010.

[29] TripAdvisor. `http://www.tripadvisor.com/`. (accessed June 1st 2022).

[30] Google Maps. `https://www.google.com/maps/`. (accessed June 1st 2022).

[31] I. Marinescu, N. Klein, A. Chamberlain, and M. Smart, "Incentives Can Reduce Bias in Online Reviews," *Economics of Networks eJournal*, 2018.

[32] H. Kaya, A. A. Karpov, and A. A. Salah, "Robust Acoustic Emotion Recognition Based on Cascaded Normalization and Extreme Learning Machines," in *Advances in Neural Networks*, ISNN 2016, pp. 115–123, 2016.

[33] W. Y. Quck, D.-Y. Huang, W. Lin, H. Li, and M. Dong, "Mobile acoustic emotion recognition," in *IEEE Region 10 Conference*, TENCON 2016, pp. 170–174, 2016.

[34] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," *Procedia Computer Science*, vol. 108, pp. 1175–1184, 2017.

[35] Z. Zhang, Y. Song, L. Cui, X. Liu, and T. Zhu, "Emotion Recognition based on Customized Smart Bracelet with Built-in Accelerometer," *PeerJ:e2258*, vol. 4, 2016.

[36] B. Resch, A. Summa, G. Sagl, P. Zeile, and J.-P. Exner, "Urban Emotions – Geo-Semantic Emotion Extraction from Technical Sensors, Human Sensors and Crowdsourced Data," in *Progress in Location-Based Services 2014*, pp. 199–212, Nov. 2014.

[37] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[38] Y. Nakamura, Y. Arakawa, T. Kanehira, M. Fujiwara, and K. Yasumoto, "SenStick: Comprehensive Sensing Platform with an Ultra Tiny All-In-One Sensor Board for IoT Research," *Journal of Sensors*, 2017.

[39] H. Ying, C. Silex, A. Schnitzer, S. Leonhardt, and M. Schiek, "Automatic Step Detection in the Accelerometer Signal," in *Proceedings of the 4th International Workshop on Wearable and Implantable Body Sensor Networks*, BSN 2007, Mar. 2007.

[40] D. Fedotov, D. Ivanko, M. Sidorov, and W. Minker, "Contextual Dependencies in Time-Continuous Multidimensional Affect Recognition," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, (Miyazaki, Japan), European Language Resources Association (ELRA), May 2018.

[41] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, 2013.

[42] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored

Conversations between a Person and a Limited Agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[43] A. Metallinou, Z. Yang, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations," *Language Resources and Evaluation*, vol. 50, Apr. 2015.

[44] O. Perepelkina, E. Kazimirova, and M. Konstantinova, "RAMAS: Russian Multimodal Corpus of Dyadic Interaction for Affective Computing," in *Proceedings of the International Conference on Speech and Computer*, SPECOM 2018, pp. 501–510, Sep. 2018.

[45] Empatica Inc., "Empatica E4." `https://www.empatica.com/research/e4/`. (accessed June 1st 2022).

[46] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[47] Omron Corporation, "Omron 2JCIE-BL01." `https://omronfs.omron.com/en_US/ecb/products/pdf/CDSC-010B.pdf`. (accessed June 1st 2022).

[48] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," 2018.

[49] D. I. Cook, V. J. Gebski, and A. C. Keech, "Subgroup analysis in clinical trials," *Medical Journal of Australia*, vol. 180, pp. 289–291, Mar. 2004.

[50] B. Zhang, C. Quan, and F. Ren, "Study on CNN in the recognition of emotion in audio and images," in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science*, ICIS 2016, pp. 1–5, 2016.

[51] Mustaqeem and S. Kwon, "A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition," *Sensors*, vol. 20, no. 1, 2020.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**