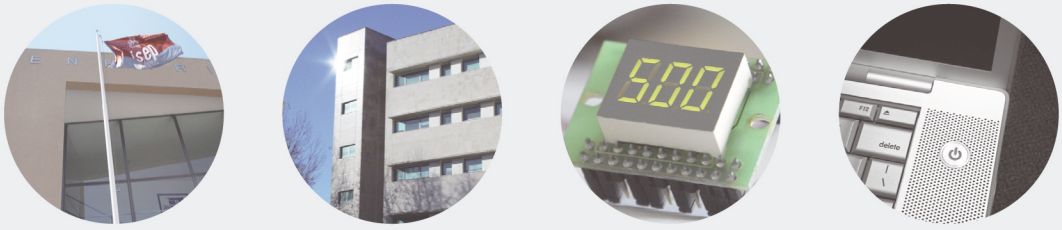
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA ELECTROTÉCNICA E DE COMPUTADORES



Towards a Scalable Dataset Construction for Facial Recognition: A guided data selection approach for diversity stimulation

LUÍS MIGUEL SALGADO NUNES VILAÇA
novembro de 2020

POLITÉCNICO
DO PORTO

Luís Miguel Salgado Nunes Vilaça

Towards a Scalable Dataset Construction for Facial Recognition

A guided data selection approach for diversity stimulation

Supervisor: Paula Maria Marques de Moura Gomes Viana
Co-supervisor: Pedro Miguel Machado Soares Carvalho

A dissertation submitted in partial fulfilment of the requirements for the degree of Master in Science in Electrical and Computer Engineering

School of Engineering, Polytechnic of Porto
2020

Abstract

Facial recognition is one of the most studied challenges in computer vision, proving to be a complex problem. This is mainly due to the variation of image capturing conditions, like object-camera relative motion or bad lightning, and the great diversity of faces in the world. For classification purposes using data-based techniques, the training dataset should reflect the diversity of characteristics of every target class (persons). Therefore, in an ideal scenario, a given classification algorithm should be able to distinguish correctly between those classes, thus maximising its performance due to the fairness of representations in the dataset.

Most approaches applied to Facial Recognition use large amounts of data to develop models for extracting facial features, making them not feasible for several application scenarios. For this reason, ensuring the variability of the representations for each person is an important requirement. Achieving this goal could also contribute to eliminate redundant and non-relevant information, reducing then the number of images used for training and consequently contributing to reduce the computational requirements.

The work developed in this dissertation aims at investigating the impact of selecting a reduced number of images in a Facial Recognition problem, when using a Deep Learning approach. The driving force behind this idea is to enable coping with scenarios where data is scarce or, although of large size, of poor quality. The main questions to answer are: How many training samples do we need to select? How long will it take to train with those training samples? How to select the best samples for the training dataset?

The solution proposed uses a feature engineering pipeline to discriminate the diversity of faces by increasing the amount of information. One of our contributions is the identification of a subgroup of metrics capable of representing diversity. As a further step, we also propose two methods that use these metrics to guarantee an increase in the amount of information. A cluster-based approach, that tries to maximise the distance between each selected item, thus maximising diversity, and an approach using Determinantal Point Process, a statistical modelling method that assigns higher probabilities to more diverse subsets using the dot product between its feature vectors are proposed. The experimental tests confirm the gain of the proposed methodology when compared with a standard random selection approach, proving to be effective in reducing the size of the dataset while maintaining a similar performance as the one obtained with the full dataset.

Resumo

O reconhecimento facial é um dos desafios mais estudados em visão computacional, revelando ser um problema complexo. Isto deve-se, principalmente, à volatilidade das condições de captura de imagem - movimentos entre objetos e a câmara, ou más condições de iluminação - e à grande diversidade de rostos no mundo. Em tarefas de classificação, usando técnicas baseadas em dados, os elementos de treino devem refletir a diversidade de características de cada classe (pessoas). Num cenário ideal, o algoritmo de classificação deverá ser capaz de distinguir corretamente essas classes, de forma a maximizar o seu desempenho. Para tal, é necessário que as representações de cada classe, observadas durante o processo de treino, sejam representativas da realidade.

A maioria das abordagens de Reconhecimento Facial utiliza uma grande quantidade de dados para desenvolver modelos que conseguem extrair características faciais separáveis e generalizáveis, sendo inviáveis em diversos cenários aplicacionais. É por isso, importante garantir a variabilidade de representações de cada pessoa. Garantindo esta diversidade, é também possível remover informação redundante e não relevante, diminuir o número de imagens utilizadas no treino destes modelos o que, consequentemente, contribui para a redução dos requisitos computacionais associados a este processo.

O trabalho desenvolvido nesta dissertação tem como objetivo analisar o impacto que uma seleção mais reduzida de imagens, focada em diversidade, exerce num problema de Reconhecimento Facial, utilizando uma abordagem de aprendizagem profunda ("Deep Learning"). O motivo principal é permitir a utilização destas abordagens em cenários onde os dados são escassos, ou, então, em grande quantidade, mas de fraca qualidade. As questões principais a responder são: Quantas imagens precisamos de utilizar para treinar o modelo? Quanto tempo leva o processo de treino com esta quantidade de dados? Como selecionar as melhores amostras a adicionar no conjunto de dados de treino de forma a maximizar o seu desempenho?

A solução proposta utiliza um processo de "Feature Engineering" para selecionar e extrair as características que melhor discriminam a diversidade de faces, o que levará a um aumento da quantidade de informação. Uma das contribuições deste trabalho é a identificação de um subgrupo de métricas capazes de representar a diversidade. Num passo seguinte, propomos também dois métodos que as utilizam para garantir um aumento da quantidade de informação num conjunto de dados: uma abordagem baseada em algoritmos de agrupamento, tentando maximizar a distância entre cada elemento selecionado, maximizando assim a diversidade, e uma abordagem, utilizando algoritmos baseados em "Determinantal Point Processes" (um método de modelagem estatística que atribui probabilidades elevadas a subconjuntos diversos usando o produto interno entre os seus vetores de características). As experiências realizadas demonstram a vantagem da utilização da heurística proposta, quando comparada com uma seleção aleatória de amostras, mostrando ser eficaz na redução do tamanho do conjunto de dados de treino, enquanto, em paralelo, mantém um desempenho similar ao obtido com o conjunto completo.

Acknowledgements

This stage had outstanding support and incentives without which it would not have been possible to achieve it and to which I am forever grateful.

I want to thank my advisor, Prof. Paula Viana, for the opportunity I was given two years ago, for your tireless pursuit and sterling work during this time steering me into the right direction. Your intervention definitely and actively contributed to the success of our work at INESC and my personal development in pursuing the academical path that I yearn to take. A sincere thank to my co-advisor Prof. Pedro Carvalho, for your contributions and for constantly helping me whenever I needed guidance or assistance.

To all my friends and coworkers at INESC for your readiness and expertise in helping me, your warm welcome when I first arrived and your consistency in making every workday a different and fulfilling day. To them, I am proud of calling you a second family: Inês Teixeira, Tiago Costa, José Pedro Pinto, Eduardo Almeida, Américo Pereira and Luiz Pires.

To my family, especially my parents and grandparents, for your endless support during my life and for being role models of live and unconditional love. For you, there is not enough time in the world to repay what I have been given.

Finally, to all my friends for their understanding of my absences and support in less good times. To my close friends, my "brothers", the ones that are continually teaching me to learn from life and not to live from what you know about life. The stories saved in my memory and your friendship are undoubted proof that I am the richest man on earth.

Luís Vilaça

“O que me leva a criar é ser uma fonte de ignição dentro do outro.”

André Neves

Contents

List of Figures

List of Tables

Abbreviations and Acronyms

AI	Artificial Intelligence
CPU	Central Processing Units
CV	Computer Vision
CNN	Convolutional Neural Networks
CHIC	Cooperative Holistic View on Internet and Content
DL	Deep Learning
DPP	Determinantal Point Process
FR	Facial Recognition
GPU	Graphical Processing Units
ICV	Iterated Cross-Validation
K-NN	K-Nearest Neighbours
LFW	Labeled Faces in the Wild
ML	Machine Learning
MLP	Multi-layer Perceptron Network
PCA	Principal Component Analysis
SoA	State of the Art
T-SNE	T-distributed Stochastic Neighbour Embedding

Chapter 1

Introduction

This introduction intends to provide the reader with a first insight into data-based application scenarios, in particular those related to facial recognition. It also presents the grounds for establishing the research questions of this work. The last section of this chapter, presents the structure of this document.

1.1 Background

Data comes from the classical *latin* "datum", meaning "that which is given". It can be generated in the form of descriptions, counts or measures of anything and in any other format. It is anything that, when analysed becomes information, which is the raw material for knowledge [1]. Nowadays, the constant flow of more and better data has transformed modern society due to the pervasiveness of digital technology. Around 3.4 billion people now have access to the internet at home, and there are around four times that number of phones and other sharing devices online [2]. This amount of information empowers predictive analytics to target the tailored needs of each citizen.

Several sectors of activity and several applications benefit from collecting, analysing and acting based on data. For example, by analysing large scale social network applications and by browsing behaviour, it is possible to generate a more thorough customers profile, hence enabling a more thoughtful decision making, and preventing unnecessary stocking costs and waste. Healthcare is another vital area where using data analysis can contribute to improving the quality of services significantly, allowing to provide more effective care by analysing and crossing information from the patient's medical history, its genetics and current lifestyle. Such analysis is only possible through the use of techniques capable of processing these massive amounts of data in real-time, with high performance and efficiency.

Machine Learning (ML) algorithms make use of the representations presented in the input data to infer patterns used to give future predictions. Its popularity was only possible due to the advances in computer architectures, that boosted the amount of available processing power, and to the exponential increase of available data. However, these methods require a hand-engineered dataset for each specific task. Ever since 2012, learning methods based on Deep Learning (DL), which is a subfield of ML, emphasises the automatic acquisition of hierarchical data representations. This can be seen as a multi-stage

distillation of information. The process consists of simple transformations applied consecutively to scale up the information, thus enabling to build a more general concept from lower-level information. From all the DL algorithms, the most popular are Convolutional Neural Networks (CNN) which are mainly used for handling image data. The architecture of CNNs and the nature of the convolutional operation allow dealing with images effectively and can empower applications that are already in use in our everyday life, e.g. automatic inspection systems[1], medical image analysis, autonomous navigation and multimedia archive indexation and organisation.

When using DL, one of the most debated questions is whether we should spend more time collecting large amounts of data or invest in improving the base models through feature engineering processes. For large amounts of data, it is proven that a model with fixed capacity[2] may only improve its performance with exponential increases of data [4]. This is not always feasible since clean annotated data is a scarce resource. Nevertheless, it is possible to reuse a previously trained model in similar tasks[3] to reduce the data size demands. However, there is not a precise estimate of the expected performance for small dataset sizes. Therefore, it is of both academic and practical interest to study how an improvement in the quality of the dataset can help reduce the size of the dataset while keeping the expected performance.

1.2 Motivation

Facial Recognition (FR) is one of the most addressed applications with DL. In a few years, it has gone from a movie reference of a highly advanced security system to a ubiquitous technology in our everyday lives. It has been presented in several ways to the modern society - from consumer electronics made available to everyone, i.e. in smartphones that make use of it for security reasons and in more sophisticated and complex applications that use massive databases which are maintained by government and security agencies. Social media and networking applications such as Snapchat or Instagram do also provide access to facial detection algorithms in the form of image filters.

There is a relevant difference between facial detection and FR systems: the former only needs to detect and locate faces, whereas the latter needs to assign and identify each face. In facial detection problems, one can generalise rules from a small fraction of data, due to a large number of shared facial features among billions of people in the world. On the other hand, in FR problems, the assumption of these resemblances increases the risk of error. Hence, obtaining distinguishable representations for the classification of each identity is imperative. Therefore, for computers is complex to perform such assignments.

Considering a human face, probably only a basic set of features is immediately acknowledged: a face has eyes, a nose and a mouth. But there is more to a face than just these features. By drawing in a piece of paper a set of faces, it is possible to observe the differences between them like the width of the nose, the distance between the eyes, the shape and size of the mouth and so on. In fact, some FR systems consider up to 80 metrics to help identify the most relevant facial features to, ultimately, be able to give an identification.

To obtain reasonable performance scores, FR algorithms are trained with a large amount of data which often include significant variations in pose, illumination, occlusion or expression. Its feature extractor must be robust enough to create/select the best

[1] Quality control.
[2] Set of functions that the learning algorithm is allowed to learn. [3]
[3] Transfer learning.

set of characteristics/representations from which it is possible to distinguish the targets. Furthermore, in production is desirable that these systems are scalable concerning the number of target entities[4], which often forces a cyclical redeployment of these algorithms. Therefore, the system needs to be designed to ensure its versatility while, at the same time, obtaining the best possible performance using the least amount of resources, i.e. images in the training set and training time if applicable.

1.3 Application context

In this thesis, we propose to explore FR approaches for identifying specific people in audiovisual content and, in contrast to existing systems, the implemented strategy aims to obtain the best performance possible while using the bare minimum resources. Moreover, this system must be versatile, so it targets the bulk of applications that use data-based FR systems and not just one in specific, thus widening the range of possible application scenarios for the methods studied in this dissertation. However, to enable testing the results in a real scenario, and given that the media sector has an unquestionable relevance in modern society, we will be addressing some FR applications in this sector. The work presented in this dissertation was carried out within the scope of two research projects for the media-sector which allowed us to experiment and derive our conclusions.

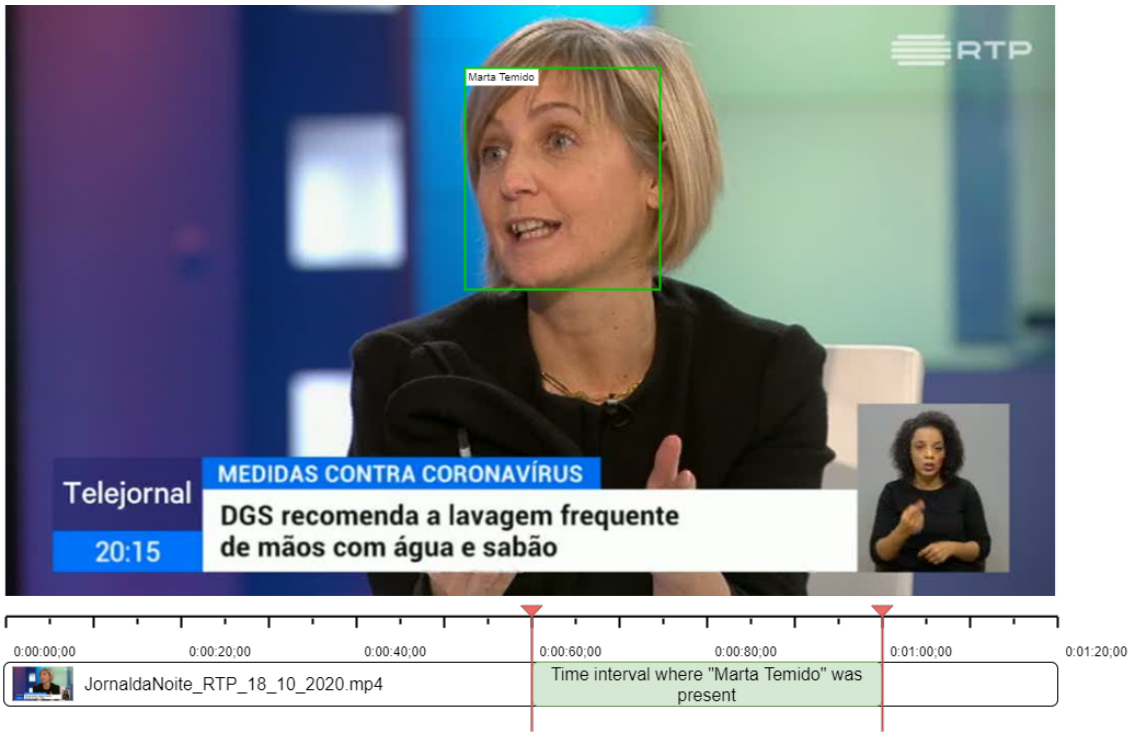
Cooperative Holistic View on Internet and Content (CHIC) is a project that intends to develop systems for facilitating and improving the workflow of content creation and diffusion within a broadcasting/television environment. The challenge is on implementing processes that enable re-using archive content (a very relevant aspect to enable improving financial income but also to enable facing the usual urgent-onto-the-air requests). However, archives are most often deposits of audiovisual content, missing the proper annotations that enable searching, finding, retrieving and re-using them. Given that news is one of the most challenging sub-areas in a broadcasting environment[5] and that finding specific personalities within the archive content is a usual task thereunder, we propose using FR algorithms for automatic content indexation with the smallest possible overhead. Besides enabling content re-use, the results of this timecoded-aware annotations will also contribute to a statistical analysis of the amount of time a given personality is on-the-air. This is another crucial application area when, for example, a political rally is being broadcasted as it may have an impact on viewers. Fig.1.1a illustrates the application of our developments within the scope of this testbed.

FotoInMotion[6] is a European project with the primary goal of developing a system capable of automatically creating dynamic and context-aware video clips from a still photo. It is mainly targeted for marketing and information dissemination in photojournalism scenarios. In this project, automatic content is created based on information drawn from the image itself, context information acquired from mobile sensor-data and collaborative user annotations. The integration of the referred different data sources composes metadata, used for automatic video production, which contributes for more efficient and instant dissemination of information. Our goal is to automatically create relevant metadata associated with the image by integrating the different data sources referred. Therefore, the aim is to identify regions of interest in the image that might contain relevant personalities and where the animation could be centred on. As an example, journalistic coverage of political gatherings usually revolves around some key personalities of the party, as illustrated in

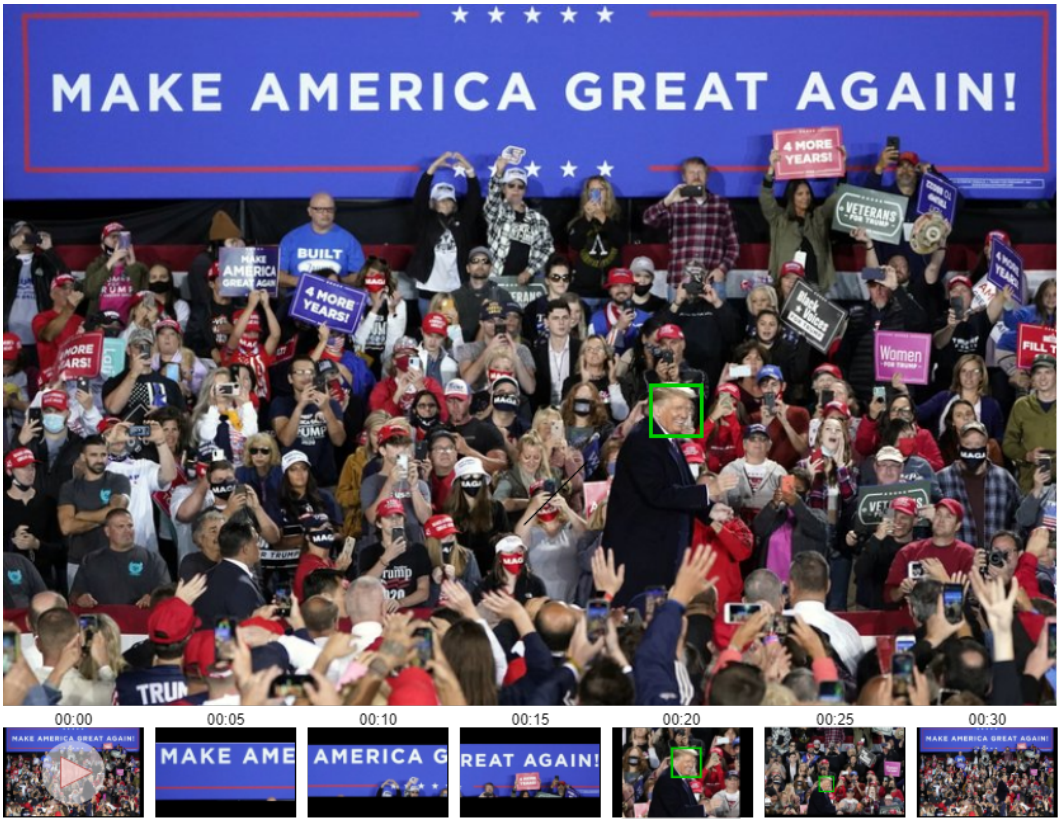
[4]Set of recognition target characters.

[5]The broadcasting environment addressed in this work is Porto Canal, a partner in project CHIC.

[6]https://fotoinmotion.eu/



(a) Timecode associated metadata - Application example (Image drawn from [5])



(b) Regions of Interest - Application example (Image drawn from [6])

Figure 1.1: Research applications.

Fig1.1b. Moreover, the spokesman is always changing, leading to an adaptation of the algorithms to new faces. Therefore, it is of our interest to optimise the pre-processing of data that will be used for this re-adaptation.

1.4 Objectives and Research Questions

The main goal of this dissertation is to investigate methods that enable increasing diversity in FR datasets and to analyse the impact of the approach in the performance of classification models and in the computational costs. Our hypothesis is that the size of the training dataset can be reduced, without compromising the quality of the results, if a sampling heuristic focused on enhancing diversity is implemented.

Therefore, the scope of this dissertation relies on the following research questions:

- RQ1: How to analyse diversity within the training data?
- RQ2: How to select the best samples for the training dataset?
- RQ3: How many training samples do we need to select to enable achieving a similar accuracy?
- RQ4: How long it will take to train with those training samples?

These questions led to establishing the following specific goals for this project:

- Analyse and evaluate the most relevant features for capturing diversity within a FR dataset;
- Evaluate the most relevant methods/heuristics for sampling the most diverse elements from a dataset;
- Analyse the entropy of the sampled subsets of data;
- Define an experimental design which enables to fairly evaluate the dataset reduction and the selected sampling methods;
- Analyse the impact of the dataset reduction, in the classification process, and compare it to the scenario of using the complete dataset.

1.5 Thesis Outline

This dissertation is divided into the following five chapters. The first chapter starts by providing the reader with the first insight into data-based application scenarios, in particular, those related to facial recognition. Focused on the application scenario, it also presents the reasons that led to the research questions. Additionally, it gives a glance of the whole document by presenting its structure.

The second chapter introduces some concepts and literature review on the topics related to the problem addressed. It starts by introducing the concept of inductive learning and its applications in ML and DL. It finishes with a summary/discussion that analyses the most relevant strategies in the scope of this dissertation's goals.

The third chapter presents an initial analysis of diversity within the facial recognition scenario. It discusses the most relevant aspects related to the construction of datasets for

FR data-based applications. Additionally, it analyses how to measure diversity and which metrics are the most relevant.

The fourth chapter presents the analysis of the impact of diversity in the training dataset. It describes the experimental design used to study this problem and to evaluate the selection methods used to sample from the original dataset. Afterwards, it presents the performance impact in DL classification models using the samples drawn with those methods. It finalises with a brief discussion that summarises the previous observations and the inferred conclusions.

Finally, the last chapter provides a synthesis of the results and highlights the main contributions of this dissertation. It concludes by pointing out likely paths able to be pursued as future work.

Chapter 2

Related Work

This chapter intends to present relevant studies on the topics related to the problem to be addressed. It is divided into five subsections. The first introduces the reader to the concept of inductive learning and its applications in ML and DL. The second exposes our area of application, while the third section deepens into our targeted application scenario. The forth section introduces the methods applied for diversity enhancement and the possible applications for our context. Finally, the last section concludes with a brief summary while establishing the pathway for the experimental design.

2.1 Machine Learning

For human beings, simple tasks as identifying a friend in a photograph or notice that someone is sick by observing changes in its voice, are carried out efficiently and without effort. Observations of this kind are made by matching information drawn from a visual stimulus with the one stored in memory (Pattern Recognition [7]). *Atinkson and Shiffrin*, on their work from 1968 [8], introduced the concept of short-term memory and proposed a model that tries to capture the process of information flow from the short-term memory to a long-term stage. This is achieved by continuously providing additional stimulus that increase the knowledge and enables acquiring information in a long-term fashion. The authors claim that this transference only occurs when the short-term memory is occupied with information. Moreover, they have shown evidence that the repeated exposure to similar stimulus strengthens the presence of stored content in long-term memory[1] and its bonds to the short-term component[2]. In other words, experience increases the ability of a human to recognise patterns. This capability is very important since it allows us to read, paint, cherish art and music, read the mood in a room, or even understanding the emotional state of people close to us by observing expressions and actions. Pattern recognition is the main component for logic reasoning as it allows perceiving the environment and all its components, relations and interactions, thus allowing to assimilate what is logic around us.

[1] Permanent storage of information.
[2] Rapid response to new stimulus.

2.1.1 Traditional Machine Learning Overview

Replicating the human capabilities that enable performing simple tasks, as the ones described in the last subsection, is far from being trivial. Earlier studies worked by establishing *a priori* the necessary steps to perform a given action/sequence of actions, or by coding explicit logical rules for making decisions. However, this kind of approach is very rigid and does not perform well in a volatile environment as the real-world. Furthermore, the overgrowing complexity of requirements from all activity sectors led to a search for more efficient and autonomous approaches to reduce human dependency. Therefore, a new methodology, where those explicit rules are recreated autonomously from past experiences in the form of a hypothesis or function was developed. This emergent field is called ML [9].

In ML, computers are programmed through inductive reasoning, from which the generic rules for a problem are synthesised from a set of examples [10]. The system is presented with relevant examples for the tasks in hand, and it iterates over them to find the statistical structure behind the data, thus learning the necessary rules to perform the task. This procedure is typically referred to as the training/learning step. Broadly speaking, the methodology is defined as the following:

> "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."
>
> — Tom Mitchell, 1997 [9]

Following this definition, a ML task is defined in terms of how the system processes the training examples. Each example is one entry or object for a particular set of data, and it is composed by a set of features which describe its main characteristics. Some of the most common tasks tackled are classification and regression. Classification is often modulated as $f : \mathbb{R}^n \rightarrow \{1, ..., k\}$, where an input feature vector is assigned to a code associated with the target class. On the other side, a regression task is modulated as $f : \mathbb{R}^n \rightarrow \mathbb{R}$, where the target space is a continuum numeric value for each combination of inputs.

To evaluate the process, a quantitative measure of the quality of the task being carried out is necessary. Once an hypothesis is inducted, the system should be able to produce useful outputs with unseen data (generalisation capability). This will be used to evaluate the system's performance when exposed to a separate dataset, called the test set.

Several approaches can be used to perform an "experience", according to the degree of human supervision in the training phase [3]:

1. Supervised learning: The input data (X) has been previously annotated so that the desired targets (y) are identified and known to the system. It is called supervised because the dataset (X, y) acts as a supervision which iteratively guides the algorithm to reach the most approximate mapping function between X and y. Some of the most popular algorithms are the ones referred earlier: regression and classification;

2. Unsupervised learning: Only the input data (X) is provided and the system must be able to infer the data's properties/statistical structure. Information drawn is mainly inferred through the relations, scale and value between data points and its features. Some of the most important categories are: clustering, visualisation and dimensionality reduction, and association rule learning;

3. Reinforcement learning: It involves learning a mapping between situations and actions. In other words, learning what to do [11]. A system/learner is placed in an closed-loop environment, where each action will trigger a feedback signal. It will select the appropriate actions through experience based in positive and negative reinforcement (reward maximisation). One pratical example is Google's *DeepMind* [12] which beated the world champion of Go[3], Lee Sedol. It still is a new research area with most pratical applications in games. However, it can be considered in many potential application scenarios in robotics, autonomous navigation or even resource management.

In supervised learning, the relation between X and y might not be deterministic due to other hidden variables that affect y. Without some assumptions about the properties of the data, the relation between test and training error cannot be studied since the test set could be composed of arbitrary values. These fundamental assumptions are [3]:

1. Each sample is independent from each other. Meaning that the occurrence of one sample does not provide information about any other;

2. Train and test splits are identically distributed[4]. For instance, if two persons flip a coin 100 times with the same probability of landing heads, then it is possible to say that both distributions are identically distributed. Therefore, if we sample randomly from either person, then the samples are also identically distributed;

We are assuming that relations between features and between samples do not exist. Hence, it is possible to simplify the procedure of modelling $p(y|X)$. The main approaches for modelling this conditional distribution are discriminative and generative modelling. A discriminative model learns directly $p(y|X)$ by mapping the decision boundaries between classes, while a generative approach learns the joint probability $p(x, y)$ and makes predictions using the Bayes theorem[5] [13]. The above assumptions affect each approach differently. Since the generative model works by using the Bayes rule, it assumes independence between input features. On the other hand, discriminative modelling makes fewer assumptions about the distribution and relies mainly in the quality of representation given by the data. Moreover, with these assumptions it is possible to provide consistent and reliable error bounds of the prediction error through sampling estimation techniques [14].

Small differences between train and test can be commonly observed in the wild, especially if the data is gathered through a long time period. Generally speaking, these assumptions hold for most scenarios. However, it is not guaranteed that a certain error in the test set will lead to the same behaviour in production. Therefore, results achieved with the test dataset should only be used as an indication for estimating future performance. In short, the goal is to minimise the training error while maintaining the smallest gap possible between the train and test error, thus learning by generalisation.

ML describes the process of learning as a search through a space of hypothesis (representational capacity[6]) using a guidance signal to measure success [15]. The goal is to find the model/hypothesis that best fits the training data. However, the final model might be of lower capacity than its maximum. For instance, consider a polynomial regression model capable of representing a function up to the forth degree. If the representation in the data

[3]Chinese strategy board game, where the goal is to surround more territory than the opponent.
[4]All items in the sample are taken from the same probability distribution.
[5]Probability of an event occur based on prior knowledge. $p(y|X) = \frac{p(X|y)p(y)}{p(x)}$
[6]Set of functions that the learning algorithm is allowed to learn [3].

can be approximated using a second degree polynomial, then we are not using the full capacity of the model. Due to most of the search/optimisation methods used not being deterministic, the solution found might be merely the first that outputs a significantly reduced error[7], which might have an impact in the future. If the obtained model exhibits a high error rate in the training set this is a sign of underfit. In other words, the hypothesis space defined by the model (capacity) is not enough to represent the structure lying in the data. In contrast, it is possible for the model to have a low error rate in the training set and not being able to generalise due to an overfit to the training data [14]. It is then important to select the correct model for each application scenario and to understand the data generating distributions which are relevant for real-world problems [3]. In fact, the "No Free Lunch Theorem" by Wolpert *et al.* [16] reinforces this idea by showing that no method's generalisation capability is superior in all possible datasets.

2.1.2 Deep Learning

Traditional ML algorithms typically try to define a set of rules or features in the data, which are hand-engineered. They require transforming the input in order to make it more amenable, through a manual preprocessing step called *feature engineering* [15], which is not scalable in practice. This problem can be illustrated, for example, in a person classification task. To classify it we need to build the concept of a person and the elements that compose it. For each element, a group of relevant and discriminative low level features needs to be selected. Furthermore, as we dive into the problem, at each step, a similar process needs to be made in order to obtain the full description of the classification target. In some scenarios it can be challenging to identify what kind of features to extract due to the ambiguity between sub elements, which might corrupt the final description.

To tackle this problem, a different take on learning was developed: Deep Learning (DL), a sub-field of ML, puts emphasis on learning the underlying features directly from data by building hierarchical representations. These representations are extracted using neural networks, which are the building blocks of DL. The process consists of simple transformations applied consecutively, layer-by-layer, in order to scale up the information, as shown in Fig.2.1 [15].

As the name implies, neural networks are a type of ML model vaguely inspired by the structure of the biological nervous system. They are constituted by a set of neurons[8] densely interconnected [3]. In particular, for each neuron, every input has an associated weight, which is adjusted during the learning process. A bias is added to the weighted sum of the inputs for helping the model fit the data, by giving a degree of freedom along the x-axis.

Since real-life data is mostly non-linear, these models need to be able to learn non-linear representations of the input space. Therefore, the output is defined through a non linear activation function as represented in the following equation: $\hat{y} = \delta(w_0 + \sum_{i=1}^{m} x_i w_i) = \delta(w_0 + X^T W)$. Furthermore, the number of layers and neurons in each one, the degree of connectivity between layers and the presence of recurrent connections[9] define the topology of an artificial neural network and establish the pieces for modifying the underlying pool functions that the model is able to represent [17].

Connecting the layers in a feed-forward fashion, as illustrated in Fig.2.2, makes each neuron to learn a function of a combination of functions learned by the neurons from the

[7] Effective capacity of the "learning" method.
[8] Basic element of a neural net.
[9] Connections that make the information flow in the opposite direction.

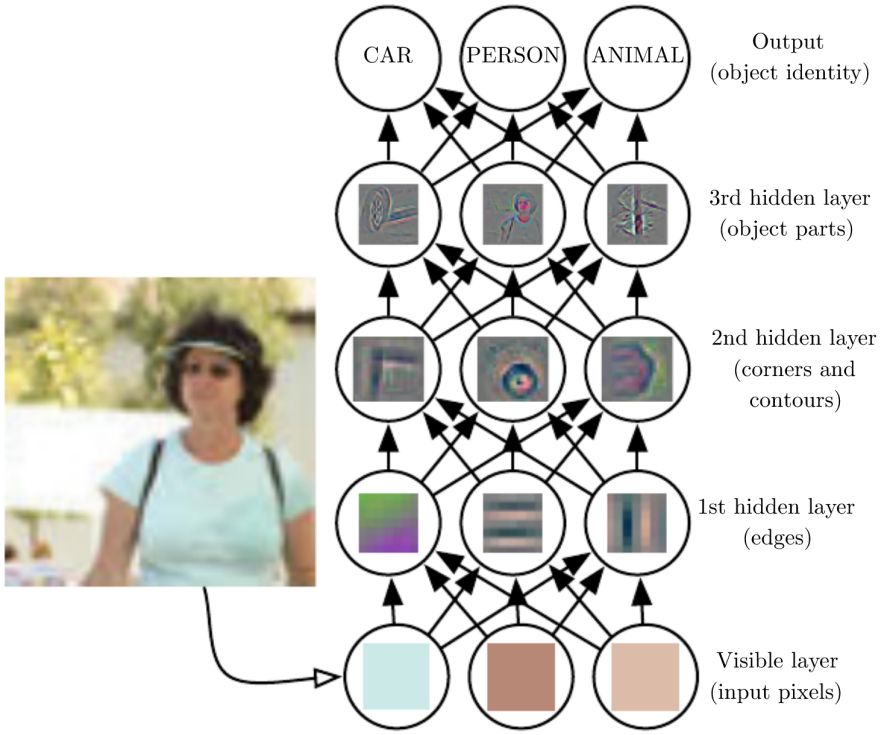


Figure 2.1: Illustration of a possible hierarchical representation extracted using a deep learning model [3].

previous layer. In other words, a Multi-layer Perceptron Network (MLP) model composed functions. Eq.2.1 illustrates this concept through an example of a two layer network model.

$$f(x) = f^2(f^1(x)), f^n = \delta(W^T x + b) \tag{2.1}$$

Typically, to adjust the weights of these models, an iterative gradient-based approach is applied following the back-propagation algorithm for computing it. Back-propagation makes use of the chain rule of calculus to allow the error to flow backwards and derive the gradient with regard to the model's parameters, allowing to estimate and minimise their weights with regard to a given cost function. This cost function is used to indirectly optimise the performance and, in the supervised case, it is defined as a measure of the quality of estimation/prediction by taking into account the current output of the model, its input and the ground truth[10].

Although the core principles and the foundation of DL were already established in the early 90's, it became more popular in the last 8 to 10 years because of the advances in modern Graphical Processing Units (GPU) architectures and the pervasiveness of data. In fact, this research field has benefited from the generalised increase in the amount of available data in several areas and application scenarios, given that one of the requirements to produce relevant results is to have large quantities of data. This requirement, that in the past posed problems in terms of computational cost, can be mitigated by using the massively parallel processing power of new hardware architectures.

Throughout the 2000's, AMD and NVIDIA invested in the development of GPUs targeting the consumer industry of videogames and the multimedia content industry by improving the rendering of realistic 2D/3D scenes in real time. Through the release of CUDA,

[10]Empirical evidence provided in the dataset.

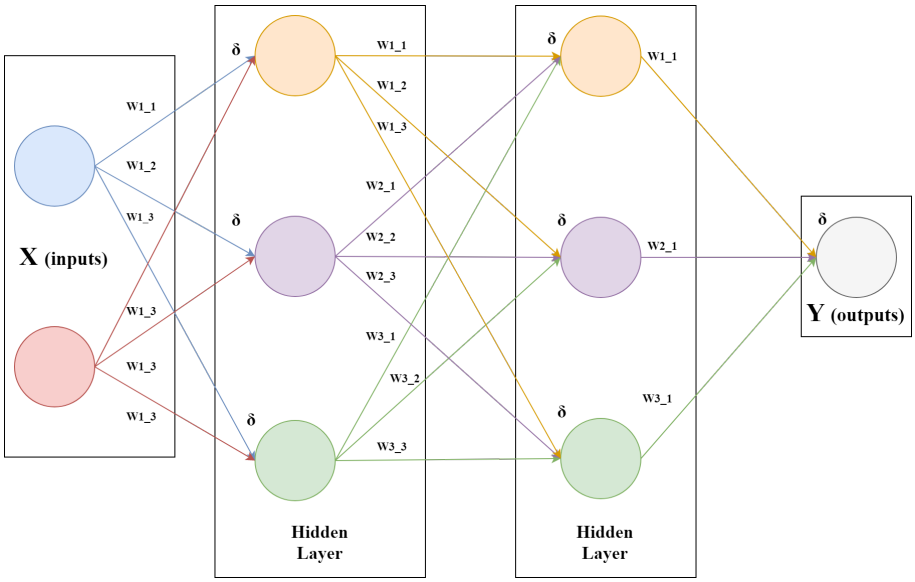


Figure 2.2: Scheme of a Multi-layer Perceptron Network.

NVIDIA's programming interface for GPUs, these devices started to replace clusters of Central Processing Units (CPU) since they incorporate a parallel processing architecture, which is very efficient for DL.

To exemplify the impact that these architectures have, let's take the example of a so called "simpler model" with 18 layers. A forward pass with this model performs around 1.8 billion floating point operations, which is very time-consuming if not conducted in parallel.

Another aspect that led to the widespread advance in the use of DL was the fact that several framework, like *Tensorflow* [18] or *Pytorch* [19], were made available. In the past, C++ and CUDA expertise was a requirement to develop Artificial Intelligence (AI) applications. Currently, by using the available platforms/frameworks, basic Python scripting skills and knowledge of linear algebra are enough to begin experimenting.

2.2 Computer Vision with Machine Learning

Specially in intelligent animals, vision is the biggest sensory system, with up to 67% of the electrical activity in the brain engaged in visual activities [20]. In the early days of Computer Vision (CV), Hubel and Wiesel [21] analysed the mechanics of vision in mammals by observing the stimulus in the primary visual cortex through electrophysiology. They discovered that the brain is composed of different cells, ones being more complex than others. Moreover, they inferred that the visual processing pathway starts with simpler cells, which respond to simple structures and light orientations, and build up the complexity of information, until it matches the stimulus observed. Their work influenced CV by showing how vision can be decomposed.

Several other authors also addressed this field of knowledge. Abhinav Gupta et.al. proposed Blocks World [22] to describe relationships and properties of objects using simple geometric shapes, allowing to reason about its constraints in a 3D scene. Also applied to 3D scenarios, David Marr [23] presented an hierarchical heuristic to portray the visual stages of representation of a 3D object. This kind of approach is also applicable in 2 dimensions. In fact, Binford [24] & Fischler [25] proposed similar methods by stating that each object

is composed by simpler geometric primitives. Respectively, their descriptions used simple cylindrical shapes, and joints where the geometric components are connected. These initial approaches of using hierarchical models for representation established a pathway for the introduction of ML in CV.

CV is an interdisciplinary research area and one of the main fields of application for ML and DL. By definition, it is the study and understanding of visual data. More specifically, DL applied to CV, addresses the problem of how to automatically extract useful information from visual content, powering applications such as image search, autonomous navigation, medical image analysis and multimedia data management.

This is particularly relevant given the ever increasing amount of visual data produced everyday, supported by low cost capturing devices. As an example, Youtube has 2 billion users worldwide [26] and 500 hours of videos uploaded every minute, making it impossible cataloguing that information. CV algorithms can be used to automatically understand the content of visual data. Eventually, it can lead to a more refined user experience in online platforms by promoting more relevant content specific to each user.

This need to understand and structure information, along with availability of faster hardware, had fostered the gradual growth of ML in parallel with CV. As the field evolved, the algorithms started to increasingly incorporate the ML's concept of learning in many important problems of visual recognition such as: image classification, localisation, object detection and segmentation.

2.2.1 Convolutional Neural Networks

Traditional approaches to CV work by extracting invariant features relevant for the task, following a human-driven approach. Until 2011, the ImageNet competition[11] winner approaches used low level feature extraction mechanisms and hierarchical reasoning as separate modules. However, in 2012, Alex Krizhevsky *et al.* [28] used CNN and showed an error rate reduction of more than 10.8% when comparing with previous methods. From that moment on, the winner approaches mainly consist in neural network model based solutions.

As the name implies, CNN's make use of the mathematical operation called *convolution*. Traditional computer vision filters also make use of this operation to highlight specific patterns, such as: blurring, sharpening, embossing or edge detection. It works by applying locally a weighting function (kernel) to transform the input and obtain a map of linear activations. These activations will be used as the input for posterior layers to build a representation following the hierarchical approach referred in section 2.1.2.

Global connectivity patterns, as the one used in MLP models, become computationally overwhelming as the input shape grows. Instead of assigning a weight to each neuron, CNN use the kernel as a weight matrix for the entire input space, benefiting from local connectivity. This allows to reduce the model's memory requirements and increase its efficiency since fewer operations are performed. The filter (kernel) slides over the entire image, computing the dot product between the weights in the feature detector (kernel) and the corresponding input regions as demonstrated in Fig.2.3. Therefore, it allows to maintain the same spacial structure as the input space. Intuitively, the model will be able to identify spacial patterns in the image by learning the most relevant transformations (filters).

[11] The ImageNet Large Scale Visual Recognition Challenge evaluates algorithms for object detection and image classification at large scale [27].

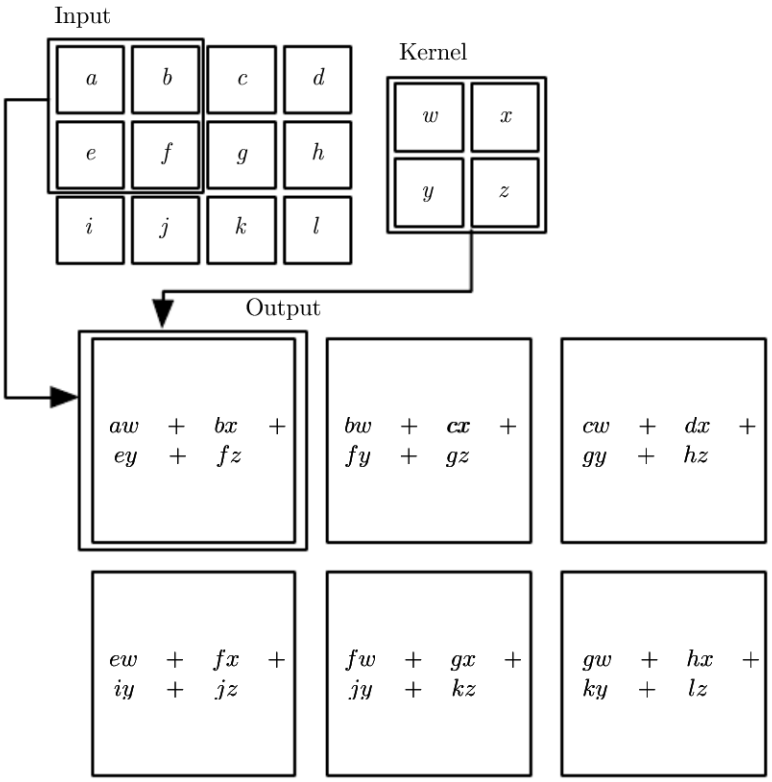


Figure 2.3: Example of 2D convolution.[3]

Considering an input image with three colour channels (Red, Green, Blue) and rectangular shape of 3 lines and 4 columns, a convolutional layer computes the dot product between the weights in the feature detector (kernel) and the corresponding input regions. This is illustrated in Fig.2.3 for one colour channel. The kernel slides through the image according to a predefined stride[12] creating the activation map for that specific filter. At each learning step, the filter is adjusted in order to change the representation to a more meaningful one using the back-propagation algorithm referred in section 2.1.2.

Typically, CNN are built using three stages of operation: convolution layers, non-linear activations, pooling layers. Non-linear activation functions and pooling operations are applied between convolution blocks. Pooling operations create a local summary of the linear activations extracted with the previous layers and allow to reduce the size of the activation map used in further operations, to improve efficiency and memory requirements. However, these elements only comprise the feature extraction module of CNNs.

The usual DL classification pipeline is comprised of the feature extraction module using the basic elements refered previously, and the classification layer. Classification is made through a fully connected layer which processes the feature vector extracted by the previous module. Its role is to transform the feature vector into an amenable input to the loss function. Usually, the loss function establishes the output's activation function. In an image classification scenario the output must be a probability distribution over the n different classes. Therefore, the typical activation function applied in these layers is the *Softmax* function [3].

[12]Space between locally connected regions in a convolutional layer.

2.2.2 CNN Architectures

In 1998, Yann Lecun *et al.* [29] proposed the first supervised training using a CNN for zip code identification in mail letters. This prototype called *LeNet* demonstrated superiority when compared with other State of the Art (SoA) benchmarks. However, it did not prevail due to the scarcity of hardware capable of running its calculations in due time. It was not until 2012, that applications with CNN became prevalent due to the results obtained by *Alexnet* [28] at ImageNet. The architecture of *AlexNet* is composed by the typical CNN structure referred in the previous subsection.

Later, in 2014, *VGG* [30] was proposed. It is built using multiple blocks containing more than one convolutional and max-pooling layers. This architecture introduced the regularisation technique called *Dropout* [31], which consists in randomly deactivating connections in fully-connected layers during training. Tests made using this technique proved that it allows to reduce overfit. Intuitively, it can be described as an approach aiming at allowing the model to learn more robust features by searching for redundancies within the feature space or, from another point of view, by introducing noise (random deactivations) in the training procedure, it becomes possible to avoid memorising non-meaningful patterns.

In the same year, the *GoogleNet* architecture [32] proposed a new concept for CNN with the introduction of inception modules. Its parallel connections allows feature abstraction in different spatial scales. The concept is different from *VGG* since the model is expanded in width, instead of depth, through those parallel connections. Widening the architecture improves training performance since it allows many matrix multiplications to be computed in parallel, while deeper networks require a larger number of sequential operations due to its structure. However, deeper networks can express more complex decision boundaries with fewer parameters, which is an advantage in multiclass problems. Depending on the case, the wide concept suggests that there might exist more efficient representations to learn than just deepen the model [33].

In 2015, the *ResNet* architecture [34] introduced the concept of residual connections to mitigate the problem of vanishing gradient. This problem occurs during training when the value of the partial derivative of the loss function, with regard to the parameters, is very low. Therefore, the proportional update imposed by the optimiser will be negligible, thus inhibiting the training to proceed. To address this issue, residual connections were introduced. They work as signals from previous layers that skip connections only to propagate the gradient forward in the model. This allows to train deeper networks, making it possible for the gradient of the loss function to propagate more easily.

The development of CNN architectures in more recent studies can be divided into two main categories: reducing computational complexity and trying to maintain the same performance to meet with the capacities of small devices [36]; automatically select/learn the best structure of the model [37]; Fig.2.4 illustrates the evolution of these CNN architectures in image recognition. The size of the circle represents the number of parameters, while the different colours represent the relations between architectures.

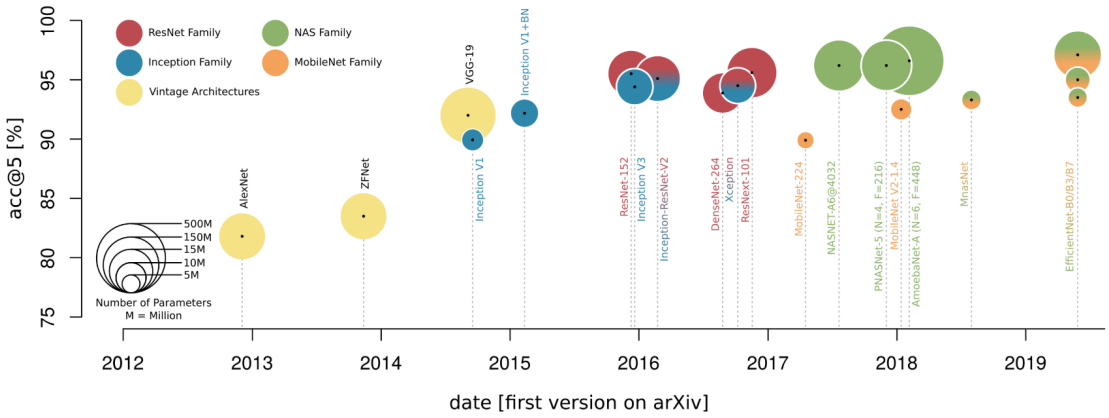


Figure 2.4: Chronological evolution of Convolutional Neural Network architectures for image recognition on ImageNet from 2012 to 2019 [35].

2.3 Facial Recognition

Traditional methods for FR reached a *plateau* in performance mainly due to their deficient representation power. In other words, they lack the capability to extract relevant features which are able to cope with the difficulties when facing new instances. By assembling a group of descriptors, traditional approaches reach 95.15% [38] of accuracy in the Labeled Faces in the Wild (LFW) [39] dataset, which propelled the use of techniques that learn these descriptors by themselves.

DL methods evolved since the initial results reported by Krizhevsky *et al.* at Image Net [28]. Architectures in image classification suffered changes regarding the depth of the receptive fields which lead to bigger, deeper architectures. Although this as the advantage of increasing the discrimination capability, it had an impact on the size of the needed datasets that have to grow in size in order to avoid overfitting.

Different CV applications also follow the usual pipeline used in image classification. However, the only components in common are the feature extractors, which are used as backbones. For instance, object detection tasks diverge from image classification ones due to its complexity, i.e. after computing the image features, the object needs to be detected and then, assigned a label to it. This degree of complexity lead to the development of region-based proposal models and regression models. Respectively, they try to mimic the attention process by focusing in regions of interest, whereas regression models try to directly map image pixels to bounding box coordinates and class probabilities. At the ImageNet Visual Recognition challenge, these kind of approaches for object detection surpassed traditional ones by a large margin and propelled its use for other application fields.

FR methods have a wide range of applications with different training protocols and evaluation tasks. A recent (2019) FR survey of DL methods [40] establishes a taxonomy for these methods which can be divided into two major categories depending on the FR task: facial verification and identification. The first consists of strategies/approaches for comparing faces using multi-input models for feature extraction. The last embraces all methods that depend on a training set or a testing gallery for comparison of inputs. Moreover, both categories can be further divided into subject dependent or independent methods. This ranks if the identities in the testing set appear in the training set, which establishes if the methods are restricted only for those identities or if they can be expanded.

Table 2.1: Commonly used datasets for testing and training [40].

		Training			Testing		
Dataset	Publish Time	#Photos	#Subjects	# of photos per subject	#Photos	#Subjects	# of photos per subject
RFW	2018	-	-	-	40607	11429	3,6
MS-Celeb-1M (Challenge 3)	2018	4M (MSv1c) 2,8M (Asian celeb)	80K (MSv1c) 100K (Asian celeb)	-	274K (ELFW) 1M (DELFW)	5,7K (ELFW) 1,58M (DELFW)	-
VGGFace 2	2017	3,31M	9131	87/362,6/843	-	-	-
UMDFaces-Videos	2017	∼22K	3.107	-	-	-	-
IJB-B	2017	-	-	-	11.754 images 7.011 videos	1.845	36,2
CPLFW	2017	-	-	-	11652	3968	2/2,9/3
SLLFW	2017	-	-	-	13K	5K	2,3
CALFW	2017	-	-	-	12.174	4.025	2/3/4
WebCaricature	2017	-	-	-	12.016	252	-
MS-Celeb-1M (Challenge 1)	2016	10M 3,8 (clean)	100.000 85K (clean)	100	2K	1K	2
MS-Celeb-1M (Challenge 2)	2016	1,5M (base set) 1K (novel set)	20K (base set) 1K (novel set)	1/-/100	100K (base set) 20K (novel set)	20K (base set) 1K (novel set)	5/-/20
Megaface	2016	4,7M	∼672,1K	3/7/2469	1M	690.572	1,4
UMDFaces	2016	-	-	-	367.920	8.501	43,3
CFP	2016	-	-	-	7000	500	14
Google	2015	>500M	>10M	50	-	-	-
IJB-A	2015	-	-	-	25.809	500	11,4
VGGFace	2015	2,6M	2.622	1000	-	-	-
CASIA WebFace	2014	∼494,5K	10.575	2/46,8/804	-	-	-
CelebFaces +	2014	∼202,5K	10.177	19,9	-	-	-
Facebook	2014	4,4M	4K	800/1100/1200	-	-	-
LFW	2007	-	-	-	13K	5K	1/2,3/530

These different approaches lead to different evaluation methods which is reflected by the amount of images per subject in the datasets demonstrated in Table 2.1. For subject dependent approaches, it is expected for the extracted features to be separable because this is defined and enforced by the loss function. However, for subject independent evaluation methods it is desired for the model to obtain general representations which can also be separable. It cannot be performed a mapping between the input space and the targets because the testing set does not contemplate any of the subjects of the training set. Therefore, the methods usually used within this protocol use loss functions which maximize the distance between representations generated by the feature extractor. Thereby requiring a large amount of generic subjects for being able to obtain reasonable performances.

Inspired by the deep neural architectures for feature extraction used at ImageNet, FR models have followed the same trend. Deep Face [41], achieved 97.35% of accuracy on LFW, while being trained with 4 million images of 4000 personalities. It modified the usual classification pipeline, presented in subsection 2.2.1, by applying metric learning[13] (siamese architecture) [42], which accepts two inputs. In 2015, by using a triplet loss function and a dataset of 800 million images, FaceNet [43] achieved an accuracy of 99.63% on LFW. This new loss function, also based on the Euclidean distance[14], considers the absolute distances of the matching pairs of faces and non-matching pairs in the same

[13]Task of learning a distance function over objects.

[14]Distance between two points in a n-dimensional space. It is given by: $d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$

training mini-batch, which went even further in upscaling the resource use.

In the same dataset (LFW), humans only obtained 99.20% of accuracy [44]. Therefore, it is possible to state that these deep FR methods have surpassed human performance (FaceNet). However, this was obtained at the cost of using large amounts of data. The use of large private datasets in FaceNet and DeepFace, not open for the research community, propelled the development of open-source sets to perform benchmarks. Examples resulting from the research community efforts, include MegaFace [45] and VggFace2 [46]. These datasets are classified in "depth" and "breadth" both reflecting the number of images per subject: a dataset of "depth" contains a smaller number of subjects but many intrasubject variations due to the large amount of samples while a dataset in "breadth" is the exact opposite. The authors of VGGFace2 [46] found empirically that one can achieve better results by training a model first in a "breather" set followed by a train in "depth", which can be analogous to transfer learning[15].

Although these datasets provide a way to achieve a fairly good generalisation ability, the majority of them contain considerable label noise, i.e. redundant or wrong labelled information. This propelled the need to analyse the quality of data, to minimise its size and noise. Fei Wang *et al.* [48] analysed label noise in current benchmark datasets and found that, with the increase of noise, one must train a FR model with a few orders more samples to achieve the same results of a model trained with a more reduced, but cleaner, set.

Omkar M. Parkhi *et al.* [49] developed a method for data selection from the internet with low label noise by using weaker classifiers to rank the data before giving it to annotators. The authors improved their work in [46] by including characteristics such as large variations in pose, age, illumination and ethnicity into the dataset. A purity above 96% is claimed, but this result was achieved at the expense of manually removing outliers. To avoid this expensive overhead, a feature engineering pipeline to automate the whole process is required.

The information highlighted in this section covers some of the most relevant studies and results obtained for FR. However, for a more extensive coverage it is recommended to refer to the survey by Mei Wang and Weihong Deng [40]. Furthermore, for a coverage of shallow (traditional) methods for FR it is recommended the survey by Daniel Trigueros *et al.* [50].

2.4 Data diversification

Instance selection is a preprocessing step that can be used to filter noise from a dataset, including unrepresentative samples or redundant information. Existing methods can be divided into two groups: wrappers and filters [51]. Wrapper methods use the accuracy obtained by the model to discard irrelevant items. On the other hand, filters use independent selection heuristics that avoid fitting the model multiple times.

The heuristics used for handling data in the filters approach are based on two types of points: border and interior instances. Border instances are used to provide useful information for discriminating classes, while interior ones are often disregarded due to the lack of relevant information. Moreover, these methods can also be categorised into three subgroups which reflect its underlying heuristic [52]. Condensation methods focus on maintaining border points by eliminating instances that do not affect classification.

[15] ML research problem which intends to improve the performance of a learning model on a given scenario, through the reuse of previous knowledge from different but yet similar domains. [47].

Edition approaches work by removing noise, which often comes in the form of outliers in the distribution of datapoints of each class. Finally, hybrid methods try to find a subset that fulfils the goals of the previous two subgroups.

The majority of instance selection methods is based in the K-Nearest Neighbours (K-NN) classifier due to its simplicity and ability to keep the dataset in prediction. In other words, K-NN uses the training dataset to give predictions by analysing the input's k nearest neighbours, thus not requiring a training procedure.

Using predefined relevant metrics obtained from a feature engineering pipeline, José Riquelme *et al.* [53] and J. Arturo Olvera-Lopéz *et al.* [54] evaluated each example from their training datasets based on the average of their scores. On a less manual approach, Yoel Caises *et al.* [55] and Barbara Spillmann *et al.* [56] both proposed filtering methods based on K-NN using centroids as the selected subset of data. Also using cluster-based selection but with another heuristic, J. Arturo Olvera-Lopéz *et al.* [57] analysed the homogeneity of each cluster and selected for the final subset either border or interior instances depending on the cluster's consistency. However, these approaches do not establish a method for obtaining an optimal size of the selected subset, which is a requirement for running the K-NN algorithm. For that purpose, Venmann and Reinders [58] proposed a method of obtaining that information by maximising the within-cluster variance.

In addition to these methods for instance selection, cluster-based sampling is also often applied to balance sets of data [59], reduce training times [60] and reduce the amount of labour needed to gather data [61]. Furthermore, Wei-Chao Lin *et al.* [62] and Chih-Fong Tsai *et al.* [59] proved that it can outperform random sampling, which is the predefined comparison method.

The information regarding these cluster-based instance selection methods covers only the most relevant work for our research questions. However, for a more extensive coverage it is recommended to refer to the survey by J. Arturo Olvera-Lopéz *et al.* [51].

Existing approaches targeting FR mainly evaluate the training dataset by enhancing its overall image quality. The quality of information is often quantified in two ways: through the use of standard reference images and comparison metrics to identify elements with lower quality [63] or proceed with a feature engineering approach by hand-picking facial properties to predict the final measure [64]. Using traditional feature engineering techniques and to achieve a global measure of quality, Yang *et al.* [64], Gao *et al.* [65] and Z. Wang *et al.* [66] used pose estimation, facial asymmetry and illumination respectively. However, these type of approaches measure image/face quality, which is not always tied to quantity of information.

Another approach by J. Ding and X. Li [67] is based in a validation strategy to detect dataset problems by evaluating the ML system as a whole. They used training sets produced with different methods to verify if the classification results were stable, thus analysing its fidelity, variety and veracity. By comparing the output features values from the last fully-connected layer of a CNN, they analysed the contribution of small images towards the training effectiveness. As a result, they could remove redundant information, but still without quantifying it.

Starting from a previously trained model, the studies of Yazhou Yao *et al.* [68] and Ahsan Habib *et al.* [69] focused on using the performance contribution of each image to select the most distinct ones. Similar approaches have been applied in active learning, which seeks to find the most relevant elements in a pot of unlabelled data [70]. Active learning requires to fit the model multiple times and to to query an external input or measurement in order to label new datapoints with the desired output.

To achieve a diverse set of tags used to describe an image, Baoyuan Wu *et al.* [71]

proposed a Generative Adversarial Model[16] to map the relation between the image features and a set of labels in order to obtain a diverse and distinct set of annotations in their dataset. In this work, diversity of annotations is achieved via sequential sampling from the probability distribution generated from a Determinantal Point Process (DPP) model. DPP is a method that selects the most distinct datapoints by accounting every possible subset [73]. However, this method can only select a small amount of points. Specifically, it can only select subsets with a maximum size equal to the rank of the covariance matrix. It is a method often used for enhancing diversity in image search [74], for document and video summarisation [75][76], and product recommendation [77].

2.5 Discussion

Considering the task of selecting a subset of diverse images from a pool of data, it can be concluded that most applications in facial recognition shown in the previous section make use of predefined low level features to select them by quality or analyse the contribution of each image to the loss function. Moreover, to the extend of our knowledge, the work presented above regarding diversity stimulation has not yet been applied to facial recognition to evaluate its impact in performance. Nevertheless, DPP and cluster-based instance selection methods are relevant methods for adapting to the output of our exploratory data analysis.

After analysing the research previously described, at the time of this research project, it is not known to us of any studies conducted with the goal of quantifying the degree of information and relate it with the model's throughput. Our research intends to show some empirical boundaries of the impact of diversity in a facial recognition system.

[16] Class of ML model used to generate new data with the same patterns and properties as the training set [72].

Chapter 3

Exploratory Data Analysis

This chapter intends to present an initial analysis on how to select the most relevant facial features for measuring diversity. It is divided into four subsections. The first exposes some important questions regarding the construction of FR datasets. The second and third sections explain how the bases for the experimental design were established. Furthermore, the second section also answers some initial questions regarding the most relevant facial features while taking into account other relevant studies from the SoA. Finally, the last section explains how the process for dimensionality reduction was conducted and exposes the results obtained regarding the most significant facial features for capturing diversity.

3.1 Relevant aspects for FR datasets

Datasets used for training ML/DL models have a significant impact on the results achieved. Increasing the size of these datasets is expected to contribute to increase the performance. However, this is achieved at the cost of increasing also the computational costs and the time to reach a stable model. Also, gathering a dataset with the required size might not be feasible for small real-world applications. Given this, being able to reduce the number of images, while keeping accuracy above a given threshold, would contribute to enable implementing real scenario applications. In this work, we argue that by enhancing the diversity of information and removing redundancies using a guided approach to build a FR dataset will enable achieving this goal.

The goal of a FR system is to be able to recognise and understand a wide range of characteristics of human faces. For this purpose, the features describing each one of the targeted classes (person) need to be separable. In other words, the decision boundaries traced in the feature space need to avoid overlapping regions between classes in order to provide the best possible classification decision. However, the large amount of shared facial features creates a significant challenge for these systems.

Creating datasets that answer this request has been the focus of some research, and it is even possible to find proposals that, by maximising the distances between facial representations, can exceed human performance. FaceNet [43] is a well-known solution that achieved this relevant result. Nevertheless, the results reported are achieved at the expense of imposing some constraints where variations of pose, resolution, illumination and

occlusion do not exist. A face which is not tightly cropped creates, however, a challenge for FaceNet to extract relevant and separable representations [43].

One other element that may have an impact on using a dataset in real-life scenarios is how well it represents the world population. Aspects such as race, ethnicity, culture, geography, age or gender need to be taken into account because they also reflect our distinctiveness. These are some of the aspects that should be kept in mind while building a dataset for FR.

Training datasets such as Megaface [45] provide pools of data with a large number of characters that cover different variations of pose and lightning. IJB-C [78] and VGGFace [49][46] went even further by considering metrics as skin colour/type, age and gender in its construction. In other applications, Burns et al. [79] in video captioning and Hee Jung Ryu et al. [80] in face attribute classification have proposed methods for forcing gender neutrality. Wei Wang et al. [81] analysed the impact of ethnicity, pose and expression in facial detection algorithms. Their goal is to increase the coverage of datasets with metrics that represent the real world's distribution, thus having a more balanced performance when the model is placed in production. In other words, by considering these aspects, it is expected that bias-related issues that can lead the model to obtain different performances in different demographic sectors are avoided.

Image quality-related features such as lightning or resolution, which are present in unconstrained scenarios, also have a tremendous impact on the performance of the FR model. This means that this aspect should also be considered when classifying the quality of a given dataset.

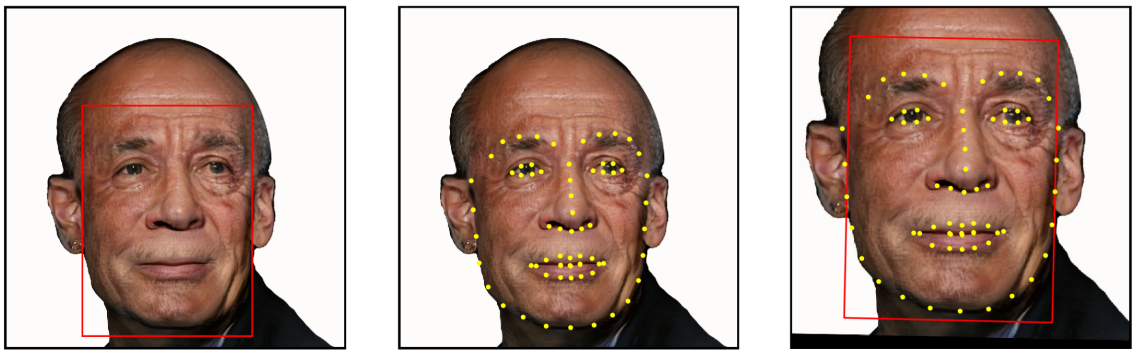
3.2 Facial Landmarks and Coding Schemes

Differences found in the performance of DL models between validation and production phases fuelled the interest of the research community towards analysing dataset's fairness in representing the real world. Studies focused on balancing the representation of a given dataset mainly use metrics that capture information related to the different demographic sectors. However, by using them, it is not possible to analyse the distinctiveness between each person and its variations. Therefore, features that enable to study individual aspects of a given face need to be considered.

Facial landmarks/keypoints are points located in every face, from which it is possible to define salient facial domains. Fig.3.1 illustrates this concept, by using the 68-point facial landmark detector from DLIB[1] [82]. In practice, methods that require to locate facial regions such as alignment, face swapping or filters or head pose estimation, often use them. These key points enable normalising a feature extraction pipeline since they are prevalent to any person. Being able to extract these points enables reducing the impact caused by low-quality image characteristics, such as resolution, making them an useful starting point for standardising the process of computing metrics related to intrinsic variations of the face.

Michele Merler et al. [83] investigated in-depth how the training data could fairly represents the distribution of faces in the world. One of their findings consists of a set of more than 30 metrics and 10 coding schemes that provide the means to compare the diversity between two FR datasets. The coding schemes are comprised of: Craniofacial distances, areas and ratios, facial symmetry and contrast, skin colour, age, gender, subjective annotation, pose and resolution.

[1] C++ machine learning and data analysis toolkit with Python bindings.



(a) face detection (size) (b) face keypoints (pose and iod) (c) face rectification

Figure 3.1: Preprocessing steps for obtaining the region of interest and keypoints [83].

In our work, we use the coding schemes that provide quantitative measures related to intrinsic features. Therefore, from the previous list, we selected distances, areas and ratios, facial symmetry and contrast as these are the most relevant for capturing intra-subject variations. The others were discarded since they are related to shared features across different demographic sectors (skin colour, age, gender), crowd-sourced annotations (subjective annotation) and image quality related features (resolution and pose). These features are not relevant for our purpose because they are not related with the subject itself, thus reflecting characteristics that cannot be used to analyse the diversity within each one.

Towards the calculation of each metric of the selected coding schemes, the process starts by preprocessing each candidate's face using the steps illustrated in Fig.3.1 and that can be summarised as:

1. Detect/locate the face using a facial detector;
2. Obtain the 68-point facial landmarks using the facial landmark detector from DLIB [82];
3. Center the face using an affine transformation based on those keypoints;
4. Discard faces with low resolution (less then 30 pixels) and non-frontal poses.

Fig.3.2 illustrates some of the mentioned metrics that, somehow, link different keypoints: distances, areas or ratios between them, measured using the euclidean distance.

Facial symmetry can be extracted by using the distance between eyes to divide the face into two parts. By applying the Sobel filters[2], the magnitude and orientation of the edges can be extracted from each one. The final symmetry features are calculated using Eq.3.1 and Eq.3.2, which correspond to the density difference and orientation similarity, respectively. Eq.3.1 uses the intensity value I(x,y) (left side) and I'(x,y) (right side) obtained in the grayscale colour space. Eq.3.2 takes into account Φ, which is the angle between the edge orientations. The final density and orientation similarity metrics are obtained by calculating the average value of DD(x,y) and EOS(x,y). Fig.3.3 illustrates the steps for obtaining the symmetry coding schemes.

$$DD(x,y) = I(x,y) - I'(x,y) \tag{3.1}$$

[2]Image processing method used for edge detection. It calculates the gradient at each pixel and tries to find the direction and rate of the biggest decrease of light.

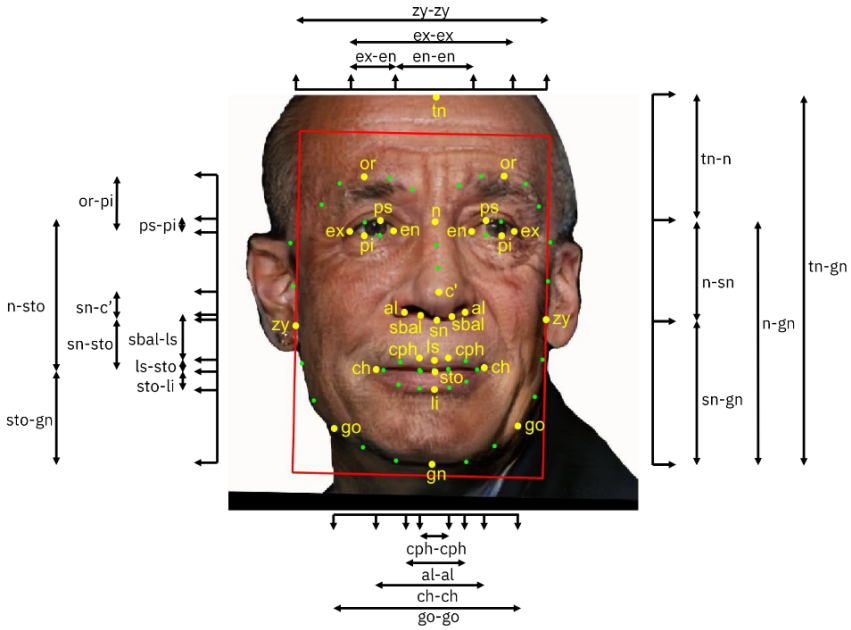


Figure 3.2: Distances, areas and ratios taken into account [83].

(a) (b) (c)

Figure 3.3: Process to calculate the symmetry coding scheme [83]. (a) First step of dividing the face. (b) Edge magnitude. (c) Edge orientation.

$$EOS(x,y) = \cos\left(\Phi(I_e(x,y), I'_e(x,y))\right) \tag{3.2}$$

Contrast, another metric, is obtained for three regions (Eyebrows, Eyes and Lips) as illustrated in Fig.3.4. It is calculated through pixel intensity in the LAB[3] colour space with regard to the inner and outer regions. The outer areas are obtained by expanding the smaller ones by 50%. This coding scheme is calculated per colour channel using the Eq.3.3, which results in a set of 9 contrast metrics.

$$C_{area} = \frac{\sum_{x,y \in outer\_region} I(x,y) - \sum_{x,y \in inner\_region} I(x,y)}{\sum_{x,y \in outer\_region} I(x,y) + \sum_{x,y \in inner\_region} I(x,y)} \tag{3.3}$$

In sum, from the 10 coding schemes proposed by Michele Merler et al. [83], a subset of 5 was selected. Resulting in a feature space with 39 dimensions (specific metrics).

[3]The CIELAB colour space, defined by the International Commission on Illumination in 1976, expresses colour as three values: L* for the lightness from black (0) to white (100), a* from green (−) to red (+), and b* from blue (−) to yellow (+).

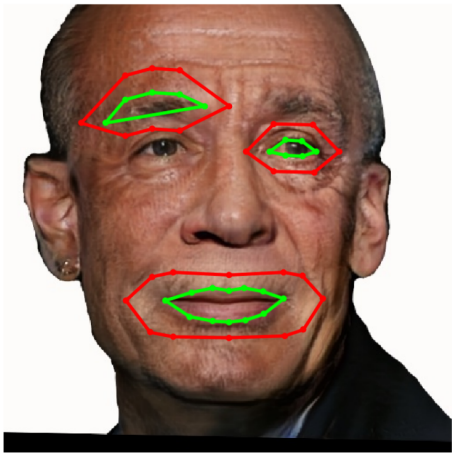


Figure 3.4: Regions considered to obtain the contrast features (eyebrows, eyes and lips) [83].

3.3 Dataset Construction

Towards the construction of a dataset for using in the experimental design of this dissertation's work, we defined a subject's list that intends to capture some of the variety present in the world's population. For that, the world population share (population density by continent) and gender was the primary concern. Other important aspects, such as age, were not considered and shall be analysed in future work.

Table 3.1: World Population Share [84].

Continent	World Population Share
Africa	17,20%
Asia	59,54%
Europe	9,59%
North America	7,60%
South America	5,53%
Oceania	0,55%

Table 3.1, presents the world population distribution over all continents [84]. For the gender distribution, a 50/50 distribution is assumed. Taken this into consideration, 12 personalities (from the VGGFace2 dataset [46]) were selected. Table 3.2 presents the list of subjects and the classification concerning its continent. The percentages of population share in the gathered dataset are calculated using the number of personalities for each class. Thus, as it is shown, it fairly guarantees a similar distribution of the world population.

Since the goal is to define a methodology that, by capturing the variability/diversity between datasets, can select the best elements, two distinct sets of data for each one of the chosen subjects were gathered. The goal of this pair of data is to demonstrate the contrast between a more variable dataset ("different" dataset), with a large number of intrinsic variations per character, and a less variable one ("same" dataset).

We assume that during a video interview, the facial characteristics captured do not change much. Therefore, the "same" dataset was obtained by grabbing random frames from interview videos for each character, thus obtaining a set of very similar items. In

Table 3.2: Dataset Population Share

Subject	Continent	Dataset Population Share
Alain Traoré	Africa	14,5%
Angélique Kidjo	Africa	
Abdullah II of Jordan	Asia	60,3%
Aditya Seal	Asia	
Aishwarya Rai	Asia	
Aya Miyama	Asia	
Dalai Lama	Asia	
Anne, Princess Royal	Europe	11,4%
Cavaco Silva	Europe	
Conan O'Brien	North America	7,6%
Alex Gonzaga	Oceania	0,5%
Zélia Duncan	South America	5,6%

Table 3.3: Datasets gathered for the experimental design

Subject	Size (nº of images)	
	"Same" Dataset	"Different" Dataset
Dalai Lama	462	404
Abdullah II of Jordan	410	453
Aditya Seal	401	458
Aishwarya Rai	312	505
Alain Traoré	312	267
Alex Gonzaga	299	419
Angélique Kidjo	423	349
Anne, Princess Royal	293	439
Cavaco Silva	542	440
Aya Miyama	276	303
Conan O'Brien	423	448
Zélia Duncan	491	319

contrast, the "different" dataset was gathered from Google Search and VGGFace2 [46] to obtain images with larger intra-variations. From the datasets analysed in Section 2, VGGFace2 was selected because it covers an high amount of relevant aspects for FR, as the ones described in Section 3.1. This group of images, gathered from VGGFace2, was complemented through a search in Google Images to include images from different epochs, to increase even more the diversity for each character. Table 3.3 shows both datasets gathered using these approaches.

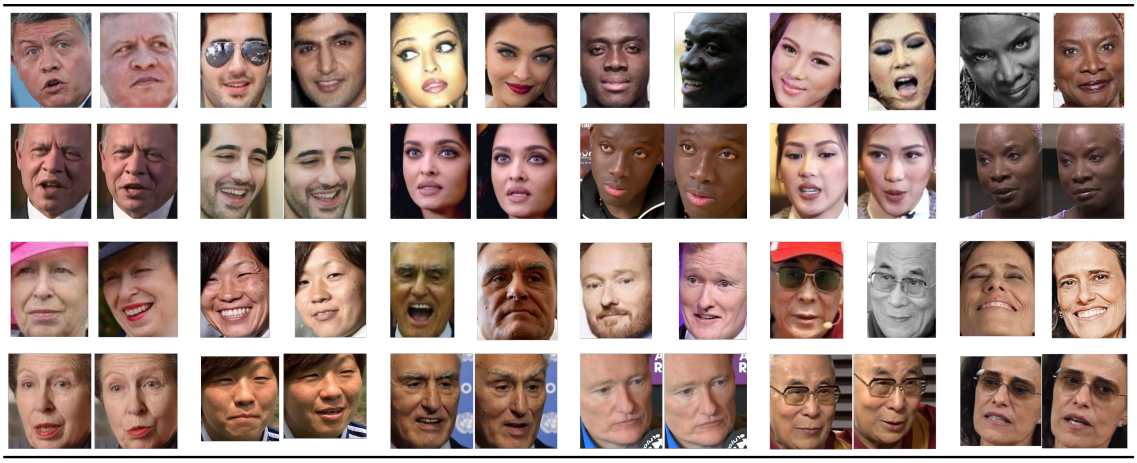


Figure 3.5: Contrast between the two types of datasets gathered for each subject (Lines: 1 & 3 - “Diff”; 2 & 4-“Same”)

Fig.3.5 illustrates this aspect for the 12 subjects that make up our pool of personalities. The images were picked randomly from the full package and, as expected, some less diversity is observed in the ”same” dataset.

3.4 Dimensionality Reduction

From the coding schemes selected in section 3.2 a set of 39 facial metrics is obtained. These metrics were extracted for each image in both datasets defined in section 3.3 to analyse their diversity. However, the sampling methods considered in chapter 2 tend to overfit with a large number of features. Dimensionality reduction methods are often applied to big feature spaces, similar to our own, to identify a subset of elements that contribute the most to the overall representation in the original space.

The goal of our analysis is not to transform the selected feature space into a more reduced version of it. Instead, we want to identify the most relevant features taking into account their relevance for our concept of diversity, which can be quantified with variance. To cope with this challenge, some dimensionality reduction methods were taking into account such as random forests, high correlation filters or low-variance filters. Principal Component Analysis (PCA) was selected because it uses variance as the measure of how important a particular dimension is, as required in this study.

PCA was applied, for each subject in both datasets, to decompose the data and find the combination of features that contribute the most for the total variance. To apply it, each variable needs first to be standardised to 0 mean and standard deviation of 1, so that scale does not influence the variance analysis performed by PCA.

3.4.1 Principal Component Analysis

PCA is a dimensionality reduction technique used for summarisation of high dimensional data. It allows identifying a set of orthogonal directions (axis) in the feature space that contains the most significant amount of information (variability).

Assuming an initial dataset X of $n \times p$, the goal of the algorithm is to find the linear relationship (loading vector) $z_{i1} = \theta_{11}x_{i1} + \theta_{21}x_{i2} + ... + \theta_{p1}x_{ip}$, subject to $\sum_{j=1}^{p} \theta_{j1}^2 = 1$, that maximises the coefficients θ_{p1} with regard to the sample variance of z_{i1}.

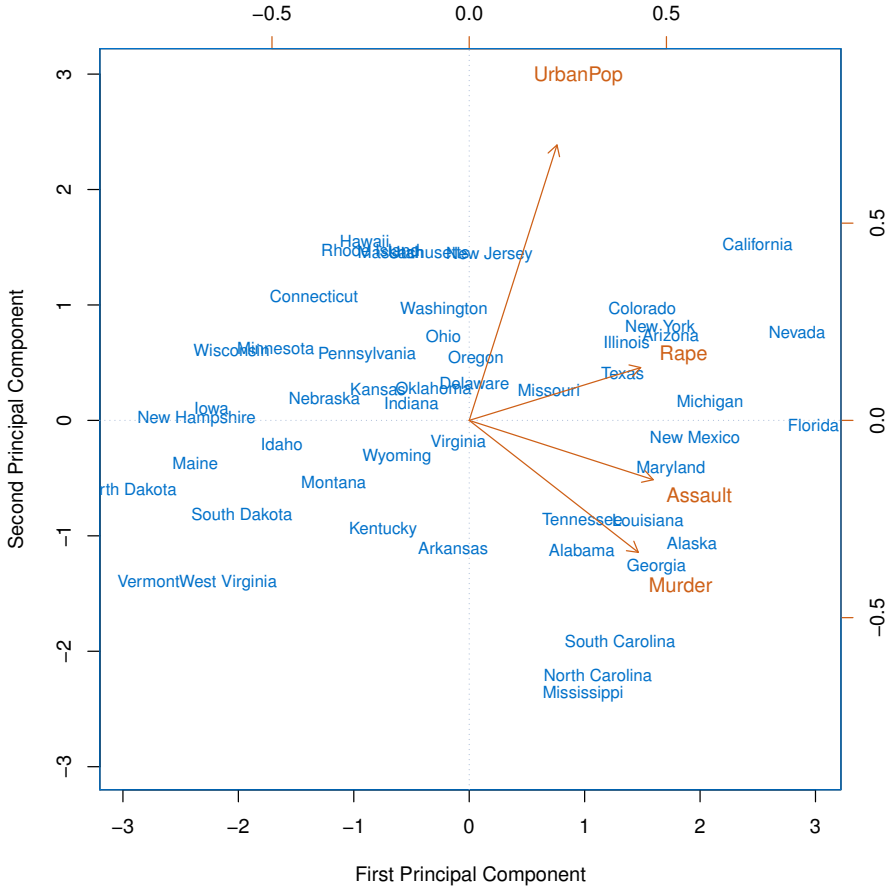


Figure 3.6: First two components for the USArrests dataset. [86]

After determining the first component, a second constraint is added to the optimisation problem, which establishes that the following component must be uncorrelated with the previous one. This is analogous to state that both components are orthogonal. This process is repeated until the number of features is reached.

To summarise, PCA computes different orthogonal linear transformations of the data, each one explaining a percentage of the total variance of the full dataset [85].

The principal components of the initial dataset are obtained using eigendecomposition of the covariance matrix (Σ). Since Σ is symmetric, it is also orthogonally diagonalisable, thus assuring the normality restriction between eigenvectors. As a result, eigendecomposition of the covariance matrix fulfils the goals of principal component analysis. Afterwards, the principal components are selected in descending order with regard to the eigenvalues because they express the proportion explained variance[4] represented by each component.

Fig.3.6 illustrates an example obtained after applying PCA to the *USArrests* dataset [86] by projecting the data into the first two principal components. The orange arrows represent the loading vectors for each dimension in the feature space, and the state names represent the projected data. The UrbanPop dimension represents the percentage of each state's population living in urban areas, which captures a relation with the overall crime

[4]Measurement of proportion to which a mathematical model accounts for the variation (dispersion) of a given dataset.

rate in each state. Therefore, its loading vector is far from the others, showing that this variable is less correlated with them. On the other hand, the results from this example also show a correlation between the other crime rates, which indicates that a state with high assault rate also has high rates for the other two. This example illustrates the advantages of PCA for data analysis and its capabilities for understanding data. This practical example and the information provided in this section establish the necessary background for understanding the results exposed in the next subsection.

3.4.2 Feature Selection Results

The selected components from the PCA analysis account for 90% of the explained variance. For each component's loading vector, we took the absolute value of each parameter (metric) and calculated the weighted average taking into account its explained variance.

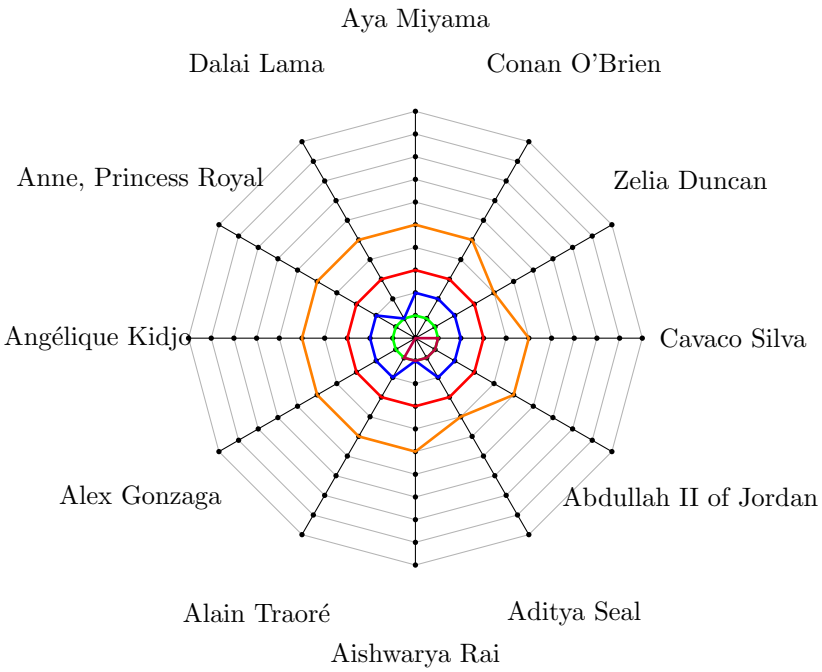
Fig.3.7 shows the variance of the metrics grouped by coding schemes for all the 12 classes and both datasets ("same" and "different"). For each metric, the weights from the previous weighted average have been summed, and its output quantised in a scale from 1 to 10. This method allows visualising the sensitivity of each coding scheme over all the 12 classes.

As it can be observed, the contribution of the contrast measures in the "different" dataset is higher than in the "same" dataset. This might be due to the diverse image capturing devices and conditions, which result in a more disperse range of contrast measures that lead to higher weights. Nevertheless, there are two coding schemes that show higher values than the others: Contrast and Ratios.

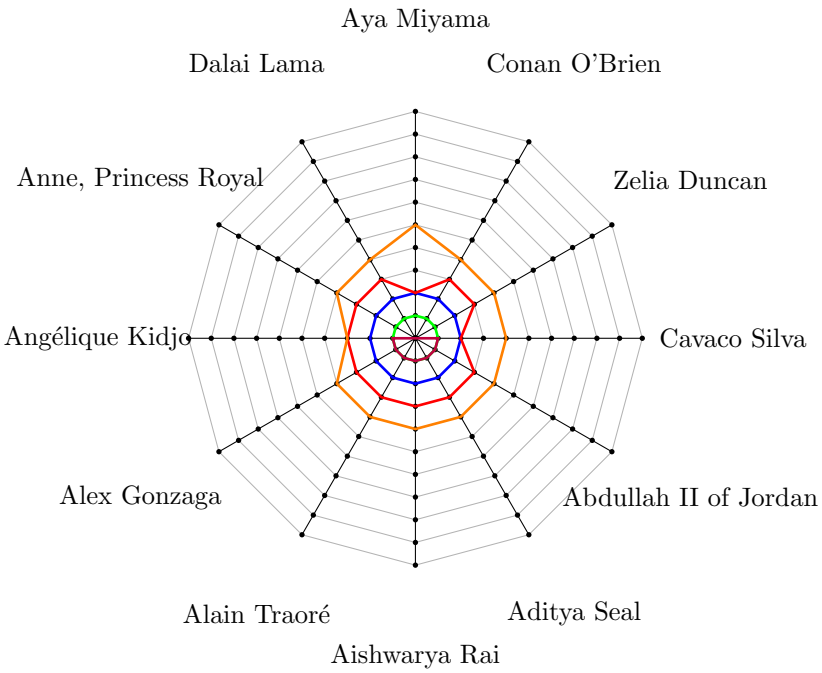
As a next step, we analysed each specific metric within each selected coding scheme and selected the group of metrics which meets two requisites: are common for every sample subject and have the higher weights. Fig. 3.8 shows the prominence of each metric in these two coding schemes by demonstrating the average sensitive value for each feature in both datasets. This resulted in choosing (ex-en)/(en-en)[5] and (ls-sto)/(sto-li)[6] among the ratios and all the contrast measures due to the tendency of these metrics being the most relevant for all characters.

The variance of each metric across the two types of datasets (Different & Same) was then analysed. Fig. 3.9 shows the differences between each one by plotting the sum of variances across all subjects. The existence of a gap in the variance plot can be observed, which demonstrates that it is possible to distinguish a variable set of images from a non-variable one by using the selected metrics. Therefore, it is expected that the sampling methods selected from chapter 2 alongside the heuristics applied with them are able to select subsets of data by enhancing diversity.

[5]Ratio between the diameter of the eye and the distance between both eyes.
[6]Ratio between the upper and lower lip distances.



(a) "Different" Dataset



(b) "Same" Dataset

Figure 3.7: Variance Contribution of each Coding Scheme in both datasets.

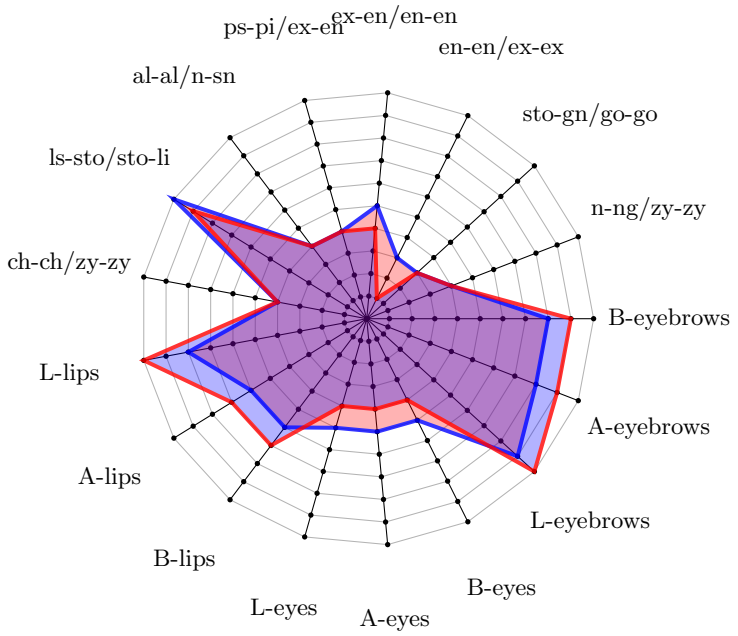


Figure 3.8: Metric weights for each of the chosen coding schemes

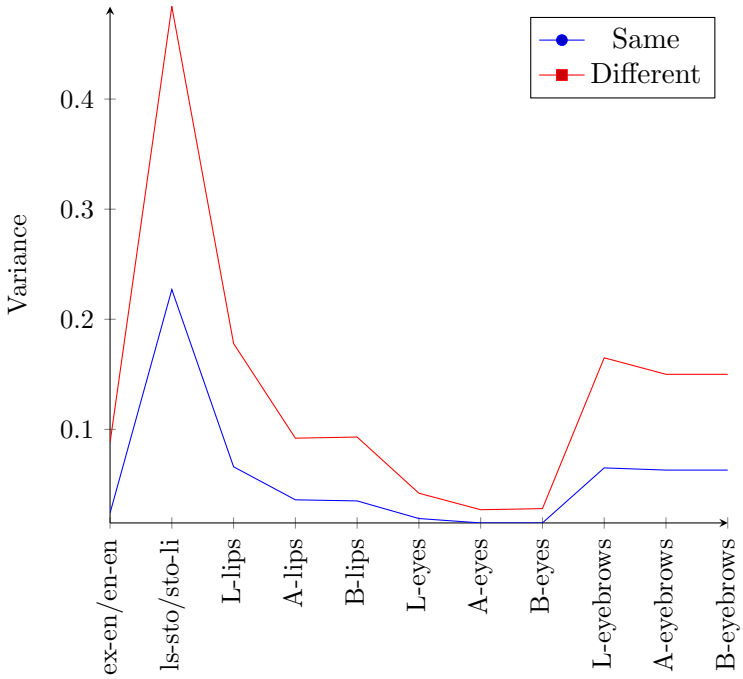


Figure 3.9: Variance of each metric

Chapter 4

Impact of Diversity Enhancement

This chapter presents the analysis of the impact of diversity in the training dataset. It is divided into five subsections. The first reintroduces the research problem, strengthening the goals established and conclusions drawn up to this point. The second exposes the experimental design used to evaluate the implemented approach. The third subsection presents the formal definition for the two proposed methods and its heuristics for selecting data. Moreover, it displays the visual results obtained for the subsets of images. The fourth reveals the performance impact in DL classification models using the built datasets for training. The final section provides a brief discussion that summarises some of the inferred conclusions.

4.1 Objectives and Problem definition

As already stated, the main goal of this dissertation is to study the empirical boundaries of the impact of diversity's stimulation in standard CNN classifiers applied for FR. In detail, we want to analyse the impact of using the methods identified in chapter 2 for selecting, from a bigger pool, the images that enable building the best training dataset. We propose applying these selection methods in the feature space defined by the group of metrics identified in chapter 3.

To support our research, the impact of the resulting subsets in a CNN classification problem performance will be evaluated to prove the usability of the sampling methods. The selection methods were compared against a random selection approach, by assessing the percentage of times each selection method produces better results.

One final and additional goal towards answering the research questions identified in chapter 1 is to set up the borderline concerning the size and content of the constructed datasets that enable achieving similar performance as the one obtained with the full dataset.

4.2 Experimental Design

Comparing the same learning algorithm with different datasets during the training cycle requires defining a fair validation scheme and the performance metrics to be used.

Learning curves can be used to monitor the performance of a DL model and provide the mathematical representation of the learning process, making it possible to determine the training dataset size needed to achieve a given performance.

Prior work has shown that accuracy improvements can be predicted empirically [87] and proved that, in image classification applications, CNN show power-law regions in its learning curves and an unpredictability behaviour when using reduced dataset sizes. Their findings also show that by reducing the training dataset size, the evaluated performance uncertainty tends to increase, following the Bernoulli's theorem of large numbers. Our proposal comes from a hypothesis, built on top of these findings, that a similar pattern will also be found in our results.

When dealing with small dataset sizes, the efficient use of data has to be assured. Several authors have investigated this problem in detail. Mitchell Lyons [88] compared Iterated Cross-Validation (ICV) and Bootstrap for small samples in a binary classification problem and recommended ICV due to the better trade-off between the variance of results and computational demand. Deepak Soekhoe *et al.* [89] studied the impact of reducing the training datasets when applying transfer learning by randomly drawing samples with reduced size, and using a fixed-sized validation set. Their strategy, however, is not sufficient to draw precise conclusions about the expected performance of the classification model due to the low sample size used in their experiments. Their main concern was to establish the randomness of training data while reducing the number of samples to maximise the likelihood that each one has a similar data distribution. A closer look by Beleites *et al.* [90] applied ICV, using 5-fold with 100 iterations, to map learning curves of linear discriminant analysis[1] classifiers with dataset sizes ranging from 5 to 25 samples per class. Among her findings, a method to plan the sample size of the experiments by using a Bayesian approach to establish the confidence interval for the performance measure is proposed.

Summing this up, the experimental design needs to take into account the aspects described regarding the efficient use of data, the likelihood of distributions of the datasets used for direct comparison and the uncertainty of low sample sizes. Therefore, our experimental design is divided into three phases, each one covering a portion of this dissertation's goals:

1. Evaluation of the selection method;
2. Evaluation of the usability of the selection heuristic;
3. Evaluation of the results obtained and comparison with the full dataset.

4.2.1 First Phase: Selection Methods Evaluation

The first phase intends to evaluate and compare both selection methods by analysing the selected subsets of data. This is achieved by doing two types of analysis for each dataset size and heuristic: a quantitative study that tries to quantify the degree of information by attempting to measure the entropy of each subset of data and a qualitative (visual) examination of dispersion and coverage.

[1] Statistical method intending to find the linear combination of features that can separate the target classes [85].

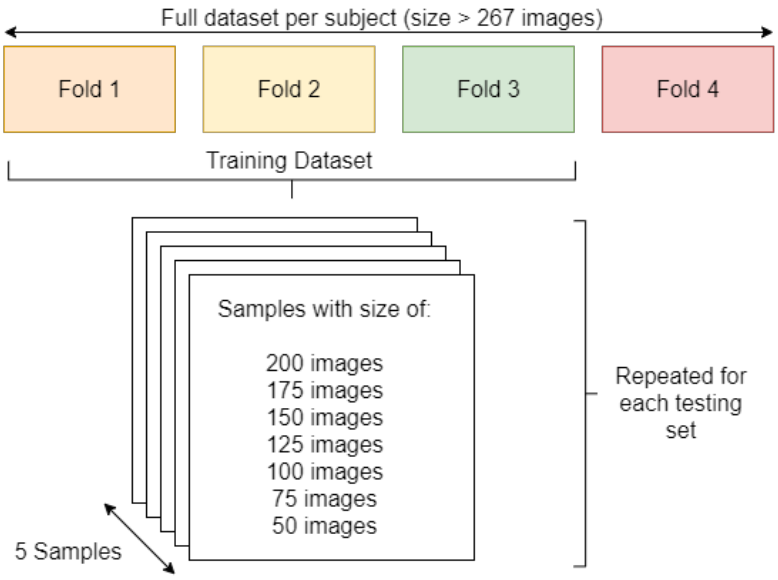


Figure 4.1: Used validation scheme

Shannon's Entropy, calculated using Eq.4.1, measures the average level of information of a random variable. The hypothesis that supports our approach is that entropy can be used to measure the diversity of a given dataset. Given this, we compute the average value over all the features for each subject.

$$H(x) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{4.1}$$

To visually analyse each selected subset, the T-distributed Stochastic Neighbour Embedding (T-SNE) algorithm [91] is used to project the data into two dimensions. T-SNE only focuses on maintaining the relations between data points. Therefore, it is only used to visually observe that the coverage and dispersion of datapoints increases with the sample sizes.

4.2.2 Second Phase: Usability Evaluation of both Sampling Heuristics

The second phase has the goal of evaluating the datasets built by each method and comparing them with a random selection of data. For that, we use the maximum accuracy achieved by each model during a pre-defined training time (epochs[2]). This metric (as defined in Eq.4.2) provides the percentage of images correctly classified by identifying the True Positives (TP) - the images correctly classified - and the ones that were wrongly classified (False Positives - FP).

$$Accuracy = \frac{TP}{TP + FP} \tag{4.2}$$

The validation process was implemented using the approach illustrated in Fig.4.1. In detail, a 4-Fold Cross-Validation process is applied to each character's dataset to assure the efficient use of the data. For each validation fold, samples were picked from the training folds pool with the sample sizes presented in Fig.4.1.

[2]Each epoch is one pass through the entire training dataset.

The evaluation was performed on 5 samples for each of the sizes under assessment to satisfy two requirements: to guarantee a fair scenario when evaluating the random approach and to avoid being stuck in a bad experience associated with a random process. Moreover, this assures that results were not obtained by chance.

Training was performed using each of the 3 datasets (randomly generated, K-means and DPP based) for each of the following sizes: 50, 75, 100, 125, 150, 175 and 200 images. Performance was calculated at every epoch, up to 20, for each of the scenarios, by computing the maximum accuracy in the validation fold at that training stage. This information allows evaluating the relation between accuracy, training time (epochs) and the size of the datasets, mapped by its learning curves.

In contrast with ICV, which iterated the whole cross-validation process multiple time, in our experimental design, the iterations are applied in the form of various selections of data within each training pool (training folds). This allows ensuring the same validation conditions (same validation data) as the dataset decreases.

For all the described experiments, the dataset named "different", in chapter 3, was the only one used, as the "same" dataset's purpose was only to evaluate the facial landmarks and coding schemes.

An additional goal was, however, defined: "How would these methods perform if the baseline (full) dataset is of "bad" quality?" This question was raised by the fact that for real application scenarios, training data may be collected from simple internet searches, that will produce a group of repeated images, creating a lot of redundant information.

To cope with this question, a new dataset ("corrupted" dataset), which attempts to simulate this automatic image web scrapping scenario, where the gathered dataset is likely to have copies of the same image, was created. This was achieved by drawing with replacement a dataset with five times the total amount of images from the training pool (training folds).

We hypothesise that the two guided selection methods will be able to eliminate any copies, while it is expected that the random approach will include duplicated information. If confirmed, the impact of the proposed approach might be even more relevant in this real scenario application where "bad" datasets are expected to exist.

4.2.3 Third Phase: Performance Impact Evaluation

The third phase intends to compare the achievements/results of the two selection methods with the use of the full amount of available data. Results drawn with the full dataset are obtained by applying 4-fold cross-validation. Comparing accuracy allows evaluating the advantage of using either method for sampling image data in a facial recognition scenario. Furthermore, it also allows to answer the research questions established in section 1.4 of chapter 1.

As shown in Table 3.3, the "different" dataset is unbalanced. Therefore, it only allows establishing the sample sizes in the previous phases of the experimental design up to a maximum of 200. However, these sample sizes might not allow the trained model to obtain a similar performance as with the full dataset. Should this happen, the sample size list can be expanded by considering the full dataset for a specific character when the sample size is bigger than the number of available images for that class. This allows to fairly compare both methods while not being affected by the unbalanced dataset.

4.3 Guided Selection

To implement the core process of selecting several representative images from a large dataset, two of the methods described in chapter 2 were set: a cluster-based method and DPP.

Several approaches for clustering can be identified. However, K-means was selected because it generates groups of similar geometrical structure, which satisfies the requirements imposed in our heuristic by theoretically assuring the maximum possible distance between data points chosen.

For evaluating a different approach, DPP was selected as it assures maximum diversity considering all random subsets using the covariance matrix, which gives a global measure of similarity between pairs of items.

4.3.1 K-Means Clustering

K-means clustering is a method for partitioning a given dataset into K non-overlapping and distinct clusters [86]. It attempts to find the group of clusters which minimises the loss function in Eq.4.3.

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \tag{4.3}$$

This represents the within-cluster variance, which symbolises the overall coverage of the solution acquired. In other words, by measuring the amount by which the observations differ from each other within each cluster using the squared euclidean distance it is possible to infer the total coverage of the selected subset within the full dataset.

Algorithm 1: K-means Clustering [86]

1. Randomly assign a number, from 1 to K, to each of the observations. These are used as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations of the kth cluster.
   (b) Assign each observation to the cluster whose centroid is the closest (defined by the Euclidean distance).

As described in the first step of Algorithm 1, each observation is randomly assigned to a cluster. Then, for each cluster its *centroid*[3] is computed and each observation is assigned to the nearest cluster. This process is iterated until these final assignments do not change. Fig.4.2 illustrates the evolution of a K-means clustering process.

The hypothesis drawn for our research objective is that K-means will enable minimising the intra-class variance while maximising their distance. Given this assumption, for each cluster, the image that is closer to the *centroid* will be selected to be included in the training dataset. As a consequence, the number of clusters defines the size of the dataset.

[3]Mean of the feature vector for the given cluster.

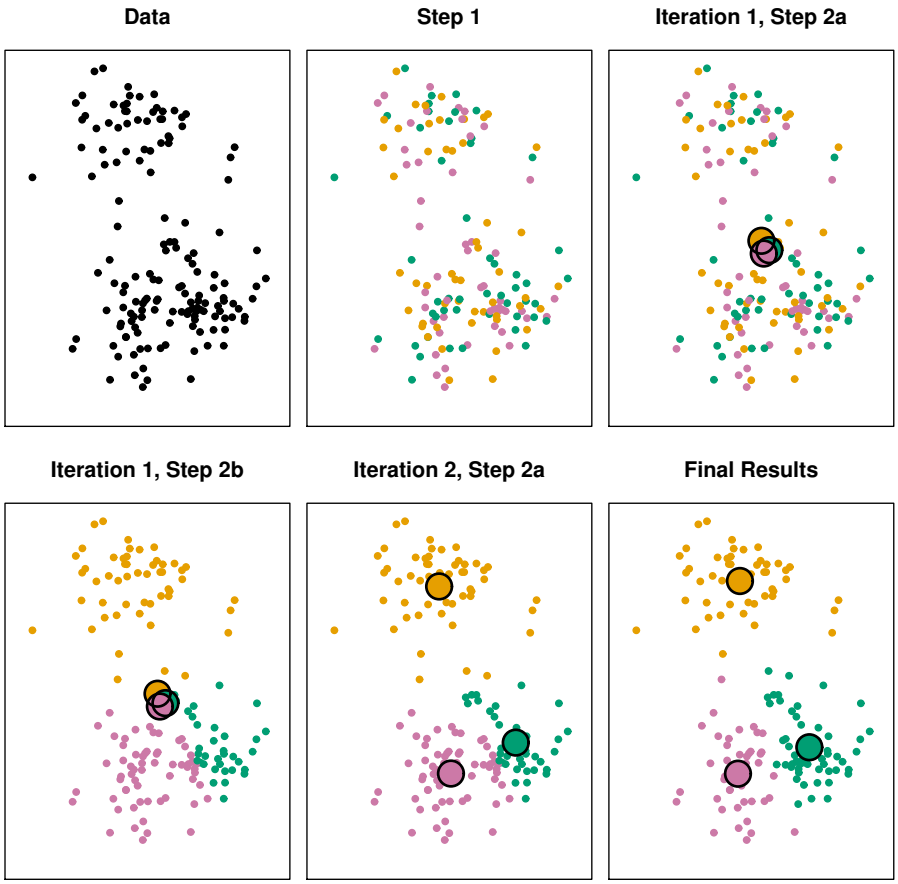


Figure 4.2: Steps of the K-means clustering algorithm [86].

Due to the way the algorithm is initialised and optimised, the solution found might be a local minimum. To cope with this problem, a voting system which will run the randomly initialised clustering until the selected subset has n (clusters) images with at least 5 votes each was included. Alongside this voting system, K-means was configured with 100 iterations and 10 random initialisations, because its output converges to the same group of images with these settings. This led into drawing, by fold, only one clustered sample of each size, following the validation scheme established in Fig.4.1.

4.3.2 Determinantal Point Process

Determinantal Point Process (DPP) is a method of probabilistic modelling. More specifically, DPP is a distribution over subsets of items from a larger global set. Given a global set $\mathcal{Y}$ composed by K items in their feature representation ($\mathcal{Y} = \{a_i\}^K$), the probability of a subset of items Y being drawn according to a DPP distribution is given by Eq.4.4.

$$p(Y \subseteq \mathcal{Y}) = \frac{\det(L_Y)}{\det(L+I)} \tag{4.4}$$

L is the covariance matrix of all items in Y and L_Y is the submatrix indexed by the elements of Y. $\det(L+I)$ is the normalisation constant, which represents the sum of probabilities of all possible subsets of items of Y ($\sum_{Y \subseteq \mathcal{Y}} \det(L_Y) = \det(L+I)$).

Intuitively, its geometrical representation can be seen as a measure of similarity between items. In other words, the matrix L (covariance matrix) measures similarity using the dot product between feature vectors. Thus $\det(L_Y)$ is the volume given by the subset Y. Therefore, DPPs favour diversity by giving higher probabilities to more diverse subsets. This geometrical view is illustrated in Fig.4.3, which demonstrates the difference between similar and distinct elements. Fig.4.3a demonstrates the probability of the subset Y, composed of two elements, which is given by the volume spanned by its feature vectors. Fig.4.3b and Fig.4.3c illustrate the contrast between similar items (c) and distinct items (b).

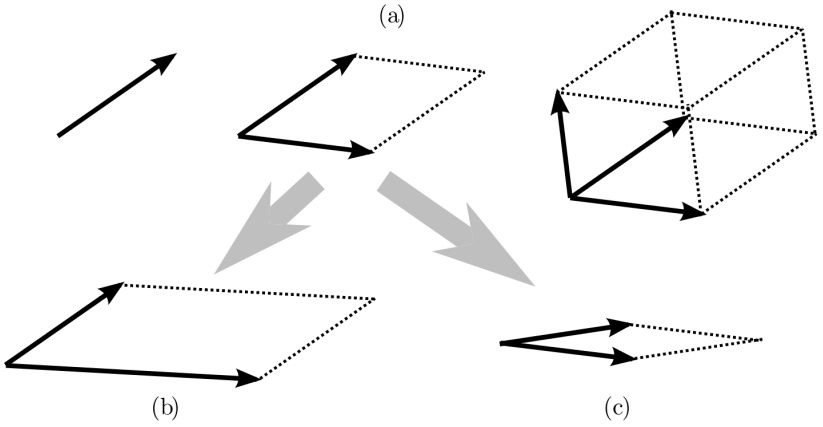


Figure 4.3: Geometrical view of Determinantal Point Processes [73].

Algorithm 2: Sampling from a DPP [73]

Input : eigendecomposition $\{(v_n, \lambda_n)\}_{n=1}^{N}$ of L
J $\leftarrow \emptyset$
for *n = 1,2,...,N* **do**
  J $\leftarrow J \cup \{n\}$ with prob. $\frac{\lambda_n}{\lambda_n+1}$
end
V $\leftarrow \{v_n\}_{n\in J}$
Y $\leftarrow \emptyset$
while $|V| > 0$ **do**
  Select i from $\mathcal{Y}$ with $Pr(i) = \frac{1}{|V|}\sum_{v\in V}(v^T e_i)^2$
  Y $\leftarrow$Y $\cup\, i$
  V $\leftarrow$V$_\perp$ an orthonormal basis for the subspace of V orthogonal to e$_i$
end
Output: Y

Sampling using DPP is performed using Algorithm 2. The first loop selects the elements (eigenvectors) according to its associated eigenvalue. After that, the elements are chosen sequentially. In each iteration, the one which maximises the distance between elements concerning the ones already chosen is selected. Fig.4.4 illustrates the results of this process in a two-dimensional example.

In practice, to follow the validation scheme and the sample sizes established in the experimental design, sample size has to be fixed. Therefore, a modified version of DPP for samples of fixed cardinality (k) was used (k-DPP). The implementation used is available

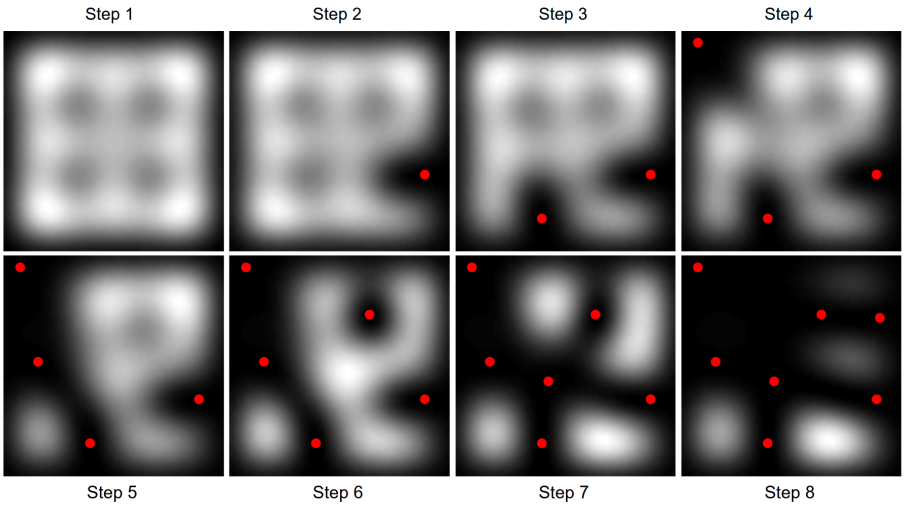


Figure 4.4: Two-dimensional sampling with DPP [73].

in DPPy [92], a Python library for sampling data points using DPP.

The goals of DPP are strictly aligned with the purpose of our work (diversity stimulation). However, it can only select subsets with a maximum size limited by the rank of the matrix L. Consequently, it is not possible to directly select subgroups of images with the dimensions defined in the previous section. Therefore, a voting system that runs k-DPP until n different images are selected was added. Although allowing to bypass this challenge, this might not be the most suitable method for sampling with DPP because it adds a condition that imposes a long processing time.

DPP was only taken into account as a comparison selection method because of its properties. Its comparison with K-means and the conclusions derived from that analysis can only be viewed as a possible future implementation which requires a modification of the underlying heuristic to make it valid from a practical point of view.

4.3.3 Selected dataset results

The results and observations presented in this subsection are related to the first phase of our experimental design. Subsets with the following sizes were selected: 50, 75, 100, 125, 150, 175 and 200. For K-means clustering and DPP, only one sample per size for each training pool (training folds) was chosen, while random sampling requires 5 examples per size.

T-SNE was applied to project the selected datasets into two dimensions to enable visualising how their dispersion and coverage evolves as sizes increase. Fig.4.5 illustrates this process by plotting results for sizes 50, 100, 150 and 200 for the character Cavaco Silva. In these charts we represent the full dataset and highlight, in a different colour, the samples that were selected for each sample size. As it is possible to see, the coverage increases with the size.

A visual comparison between both methods allows to perceive that, for more reduced sample sizes, K-means can select more disperse configurations than DPP, as illustrated in Fig.4.6 for a sample size of 100. In this example, both regions, highlighted in orange and blue, show a more disperse selection in K-means selection. For instance, in the blue region, DPP selects a group of overlapped datapoints, while K-means gives a more spread out selection. The orange region reflects a similar case, where DPP fails to cover a more

spread out area. These disperse configurations might result in better entropy values that can lead to higher classification results.

For the higher sample sizes it is not possible to draw any conclusion regarding dispersion due to the cluttering of data points. Furthermore, K-means fails to select more data points in some of the most areas with high density. This is due to its heuristic, which only considers the distance between points.

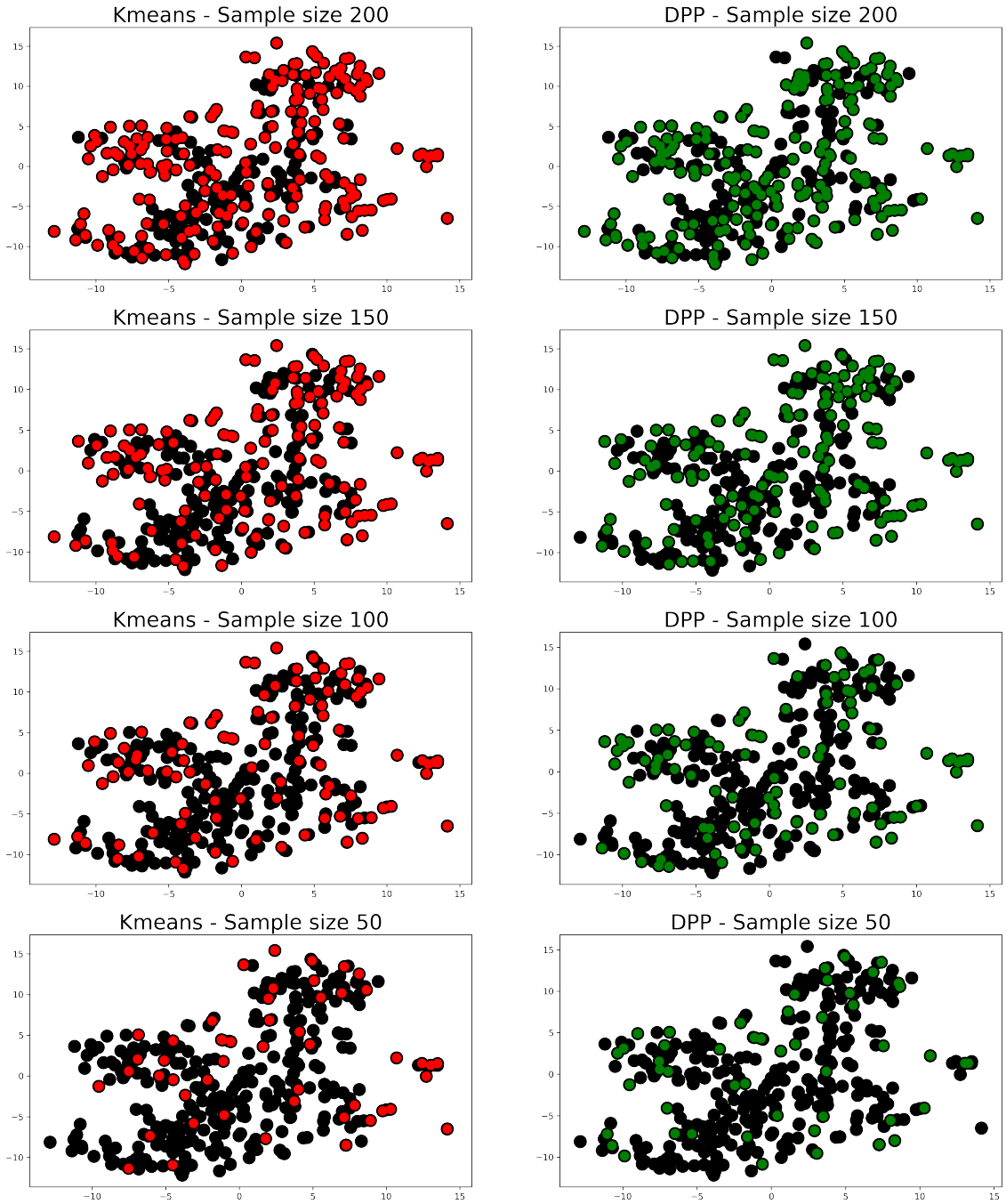


Figure 4.5: Visual illustration of the selected datapoints using both selection methods. Low dimensional space generated used T-SNE [91].

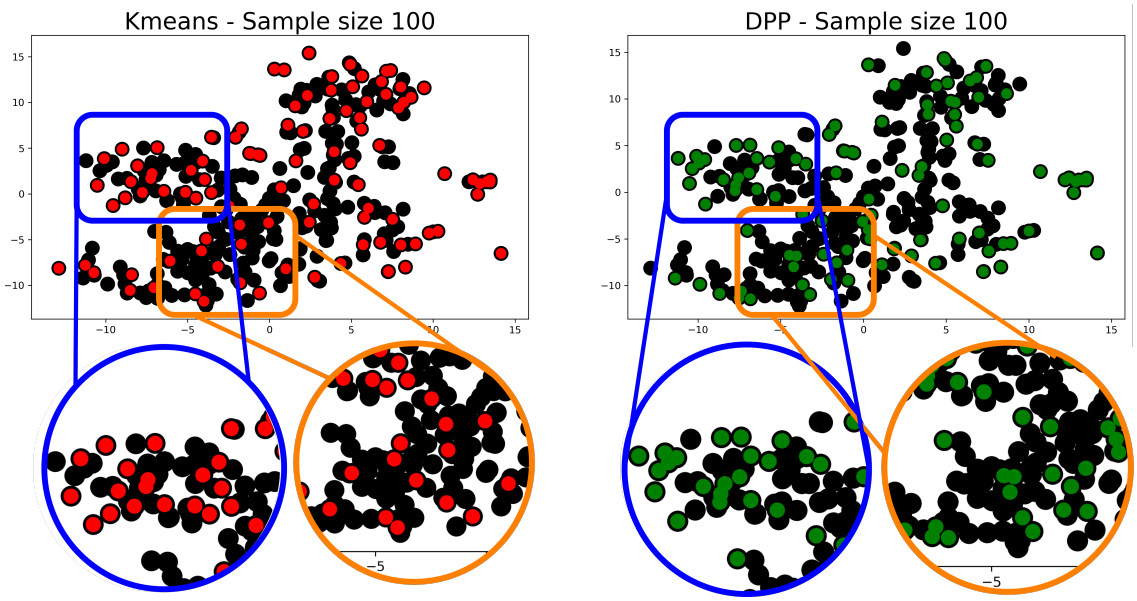


Figure 4.6: Visual comparison of the selected datapoints using both selection methods. Region Analysis.

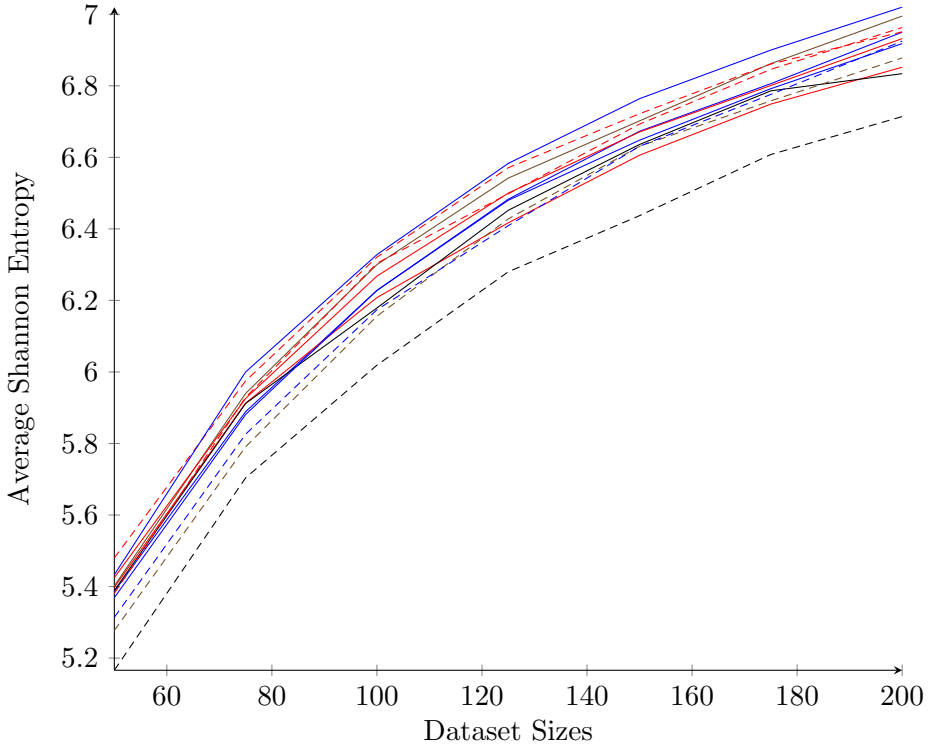


Figure 4.7: Average Shannon's entropy per subject

This analysis was only intended to provide a first insight on the sampling method's results. Therefore, to infer the effectiveness of the selection method more precisely, Shannon's entropy[4] was used to measure the quantity of information for each of the selected

[4]Average level of information of a random variable.

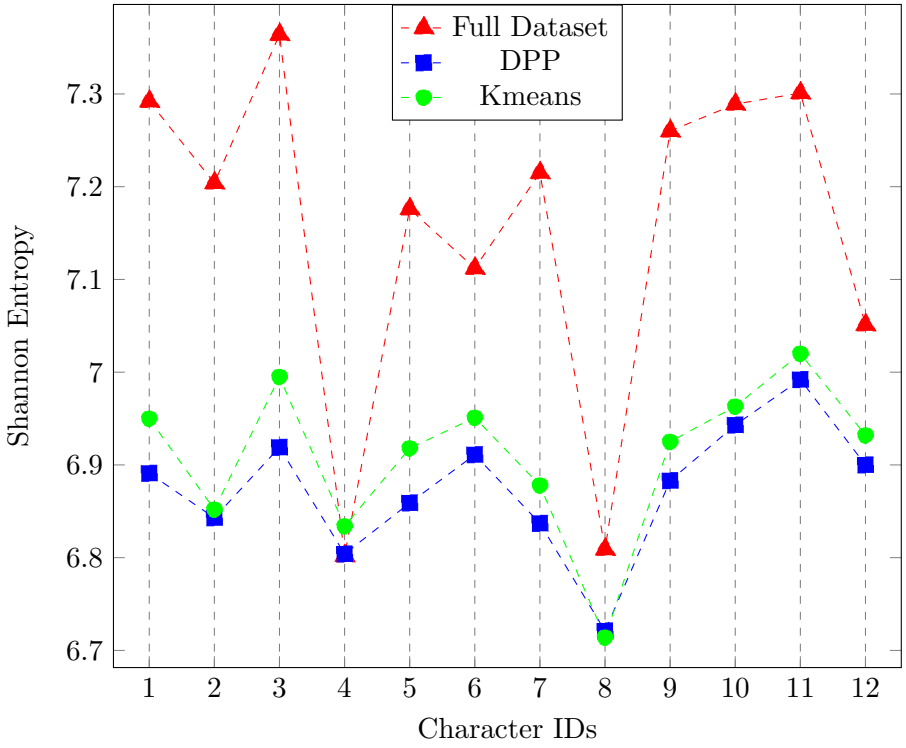


Figure 4.8: Maximum entropy comparison between Kmeans, DPP and the Full dataset

subsets. Fig.4.7 illustrates the entropy results obtained by averaging the individual values over all features. The graph shows the entropy over all sizes by stacking the results obtained of all characters. As expected, entropy increases with the sample size. However, each character reveals a different maximum entropy score, suggesting that the sample size must be adjusted differently for each one.

Fig.4.8 illustrates this point of view by plotting the maximum entropy score obtained for each character in both selection methods and comparing them with the full dataset. In fact, it shows that defining the same sample size for each character results in different gaps of information. This supports the idea of selecting an optimal sample size per subject. Furthermore, it shows a relevant aspect regarding character ID 4 (Alain Traoré) which is the subject with fewer images in the dataset. The entropy score obtained with K-means clustering is higher than the full dataset. Since the sample size comes close to the number of images in the training pool (training folds - Fig.4.1), the sampling algorithm selects a group of images with a size close to the maximum available. However, it can effectively reduce the amount of redundant information, which is represented by a higher entropy score.

The information illustrated in Fig.4.8 regarding DPP and K-means enables also assessing the effectiveness of the two methods. As it is possible to observe, the dataset gathered by K-means has higher entropy scores than DPP for most of the characters. These results suggest that K-means might able to obtain better classification scores than DPP. The next section will present the study of the impact of these sampling algorithms using the datasets analysed here.

4.4 Image Classification Results

Given that the amount of data available in the datasets (Table 3.3) is not enough to train effectively SoA FR techniques, standard image classification architectures were used to evaluate the selection methods. Although the use of standard classification layers is not aimed for critical FR applications due to the indirectness and lack of efficiency regarding the separation of classes and its generalisation, both approaches share similar feature extractor architectures.

The performance on the target task (separating representations of different characters and recognising them) depends on the raw information inferred by those backbones. Therefore, by using image classification models, it is expected that, due to this sharing, conclusions drawn in this dissertation can be translated into SoA models.

Joel Hestness *el al.* [87] found that Resnet architectures require model sizes above 3.4M parameters for fitting even with small datasets. Therefore, we chose Resnet-18 for the experiments, using it as a feature extractor due to its balance between computational requirements and generalisation capability. All experiments were normalised by using the same parameter initialisation and hyper-parameter settings for each training moment. Transfer learning was applied using the pre-trained model from ImageNet for allowing to reduce the training time.

Following the second phase of the experimental design, results using the defined validation scheme, for each epoch, size, experiment and selection method were obtained. The maximum accuracy achieved in each learning curve is considered an independent experiment. For every experiment, we compared performances for both methods and calculated the percentages where the selection methods show improvement. Fig.4.9 and 4.10 present the results by comparing equivalent validation folds in each sampling methods against a random selection approach.

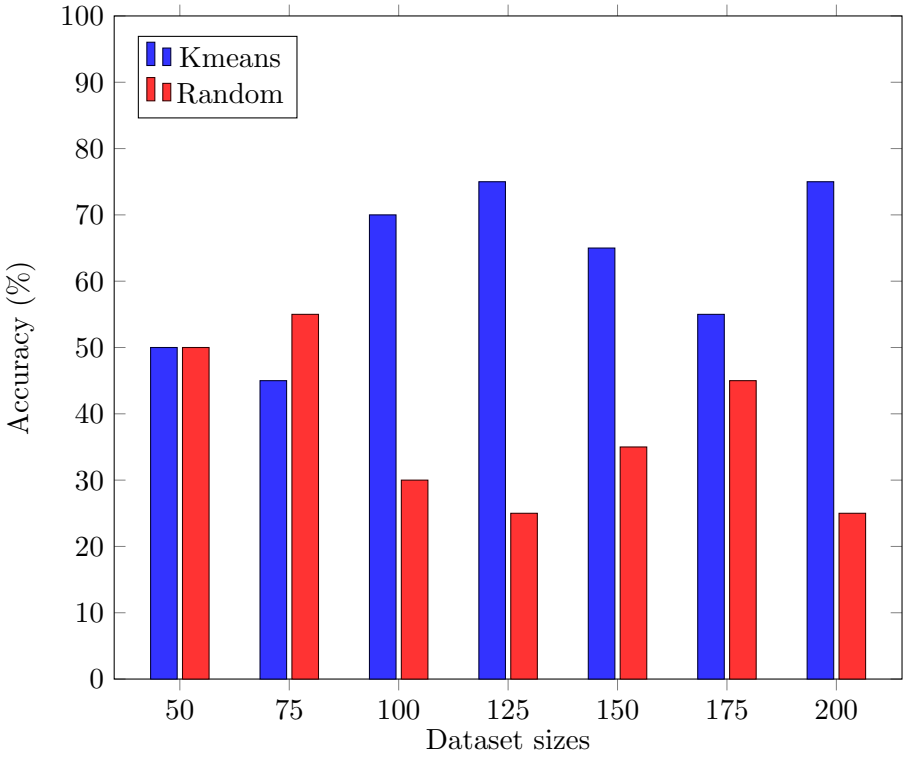
The analysis was conducted on the two datasets mentioned on section 4.2 (the "uncorrupted" and "corrupted" datasets), both based in the "different" dataset presented in section 3.3. Fig.4.9 and 4.10 illustrate the results for the two experiments.

The analysis of the charts show an unpredictable behaviour for smaller sizes (50, 75 and 100), as expected in section 4.2. For these sample sizes it is then not possible to drawn general conclusions due to the inconstancy of results. However, for bigger dataset sizes, K-means shows clear advantage over the random selection. In fact, by analysing the results obtained with the "corrupted" dataset, K-means shows even better performance by being able to select more relevant data.

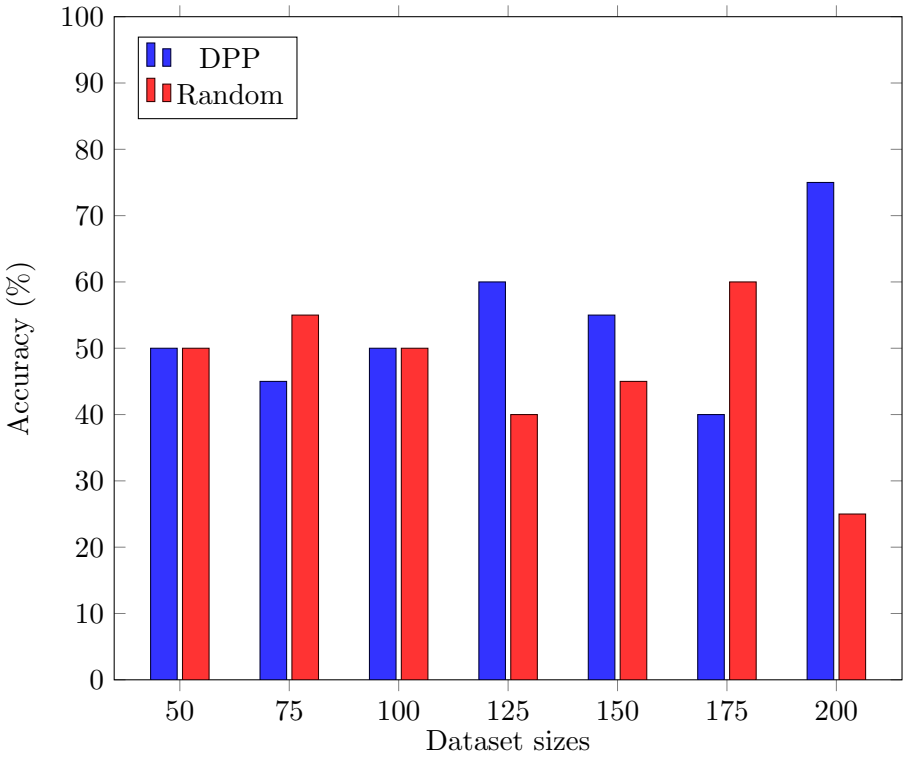
By analysing Fig.4.9, no advantage in using DPP for selecting data is identified. This is even more noticeable in the "corrupted" dataset, which shows an even worst performance. DPP seems not to able to distinguish redundant information from valuable one in this use-case.

These observations can also be supported by the results drawn in the previous phase of the experimental design (section 4.3.3), which suggests that K-means can provide better classification results because of its entropy scores and dispersion of selected data points in the visual results using T-SNE.

In the third step of the experimental design, we compared the full dataset's performance with the selected subsets that obtain the highest scores. Resnet-18 was used as a feature extractor for distinguishing the 12 characters present in the initial dataset (Tab. 3.3). Standard cross-validation was used to estimate its performance, which led to a score of 94,23% of accuracy. This experiment used an average of 298 images per subject, taking around 8 minutes to train the model for 20 epochs.

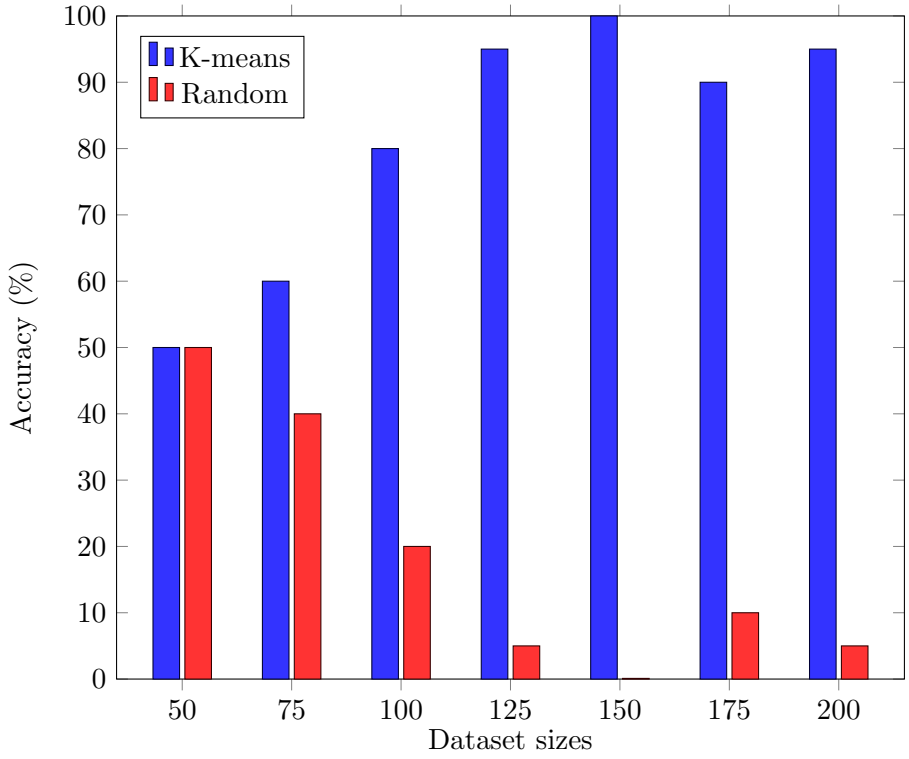


(a) K-means Clustering

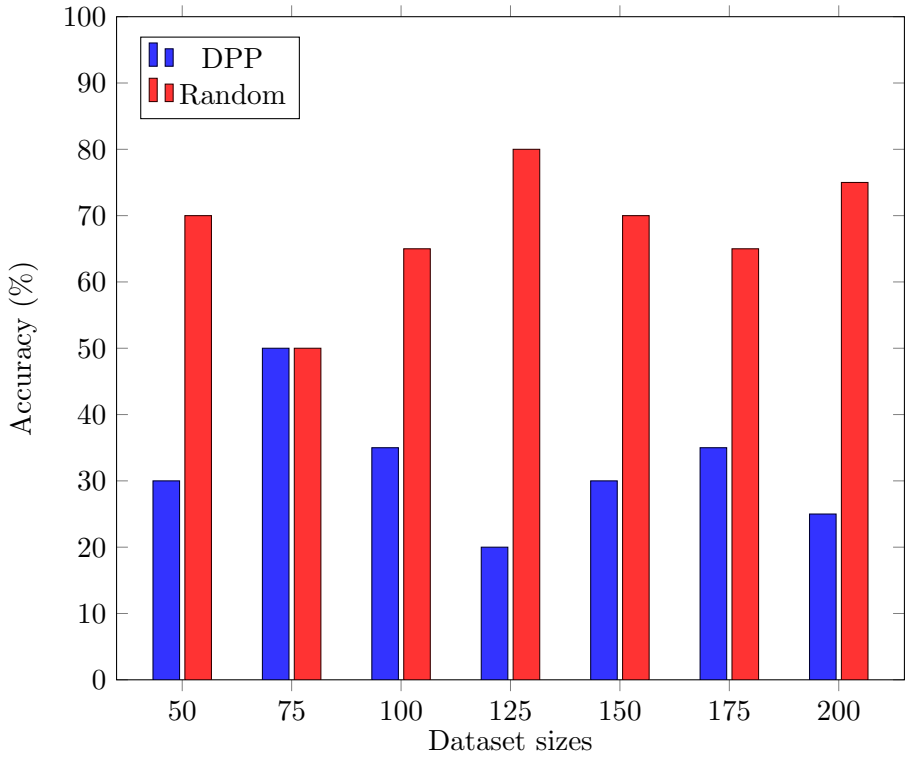


(b) DPP

Figure 4.9: Comparison between selection methods using the "uncorrupted" dataset



(a) K-means Clustering



(b) DPP

Figure 4.10: Comparison between selection methods using the "corrupted" dataset

All experiments were conducted using the same hardware with the following specifications: Intel i7 6700K at 4.1Ghz, 16Gb of Random Access Memory (RAM) at 2400Mhz and a Gigabyte Nvidia GTX 1060 with 6Gb of RAM at 1840 Mhz. Consequently, it is possible to estimate the training time for each dataset since they are proportional to its size. For this use case, with a mini-batch of 12, each iteration took an average of 5ms.

These values will be used as a reference for analysing the dataset reduction. In other words, to find how many images we must select to obtain the closest performance to the baseline and the associated training time.

With the previously defined sample sizes, the maximum accuracy acquired is still far from the baseline. Therefore, the sample size list was extended to 300, according to the heuristic established in the experimental design. Fig.4.11 illustrates the cross-validation estimate obtained for each method with the extended sizes using the "uncorrupted" dataset.

These findings show that K-means clustering enables approaching the baseline performance while still reducing the full dataset. Significant differences between DPP and k-means in what concerns accuracy cannot be identified. However, DPP instability of results, shown in the second phase of the experimental design, make K-means the preferable selection method in this study.

In short, selecting a dataset with 300 images per subject enables achieving a similar performance score as the one obtained with full dataset. Using this sample size, only 4 of the 12 target classes used the full amount of images, while the other 8 used a sample size of 300. This produces an overall reduction of 276 images. With this dataset, the training time is reduced from 8 to 6 minutes.

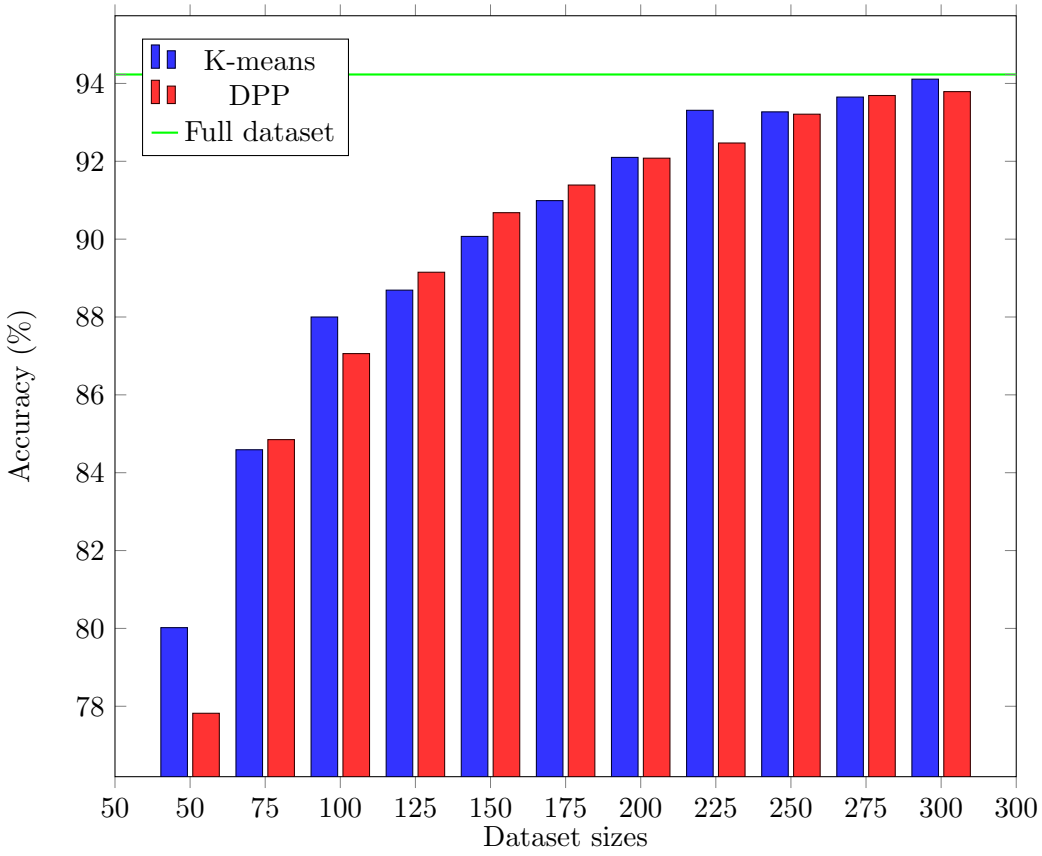


Figure 4.11: Average accuracy scores for each sample size

4.5 Discussion

Several conclusions can be drawn from the experiments described in this chapter. The first conclusion is that is although DPP is a method that theoretically considers all the combinations of data points, it fails to disperse its samples properly when compared with K-means. This is supported qualitatively and quantitatively by the results presented in Section 4.3.3.

The second conclusion is that, when compared against a random selection approach, the cluster-based (K-means) sampling methods outperforms DPP. Differences between both methods are aggravated using the ”corrupted” dataset, in which DPP fails to select relevant data. This is supported by the results described in section 4.4 that also support the previous observations in Section 3.4.2 of Chapter 3.

Additionally, as stated in Section 4.2, it is not possible to draw general conclusions for either method when facing very small sample sizes, due to the instability of results. In a more “real-life” scenario (“corrupted” dataset), results demonstrated in section 4.4 exposed a more steep ambiguity for DPP over all sample sizes, making it not suitable for the aimed use-cases.

The clustering approach has, however, shown a quite good performance when compared with the use of the full dataset, allowing to achieve a similar performance while reducing the training time.

Chapter 5

Conclusions

This chapter draws a set of conclusions, reviews the findings and main contributions of the thesis and points out likely paths able to be pursued as future work.

5.1 Overview

The focus of this dissertation was to study the impact that the quantity of information has on the expected performance of data-based classification models applied for FR. First and foremost, in Chapter 2, relevant research on FR and diversity enhancement was studied, which allowed establishing the most suitable approaches to explore. Additionally, it also presented some relevant concepts to enable a full understanding of the experiments conducted.

After having selected relevant diversity enhancement methods, a way to compute and analyse diversity was proposed. For that, a consistent set of facial features was identified. This feature space is used as a basis to study the impact and advantage of using the proposed sampling methods. Through an analysis of the samples drawn by each method, results show some disparities between techniques. These differences comply with the findings obtained regarding their impact in image classification.

The usefulness of these sampling methods is compared against a random selection. The cluster-based heuristic has proved to be the most effective one, even in more challenging and "real-life" scenarios ("corrupted" dataset). In the final step, the sample sizes in the experimental design were expanded to allow comparing the samples with the full dataset's usage, and still, obtain similar performance scores. Results showed that a dataset reduction is possible, while not having a significant impact on the model's expected performance. Furthermore, the initial hypothesis that the quantity of information has an impact on results was confirmed.

5.2 Contributions of the Thesis

All in all, with this dissertation's findings, it is possible to answer the research questions established in Section 1.4. Concerning RQ1, a methodology for analysing diversity is studied in Chapter 3. The group of facial features selected make it possible to analyse diversity and the intra-subject variations present in a given dataset. In fact, it was shown that they enable to distinguish a more variable dataset from a less variable one.

Then, RQ2 is answered through the results obtained in Sections 3.4.2 and 4.4, suggesting that the proposed cluster-based heuristic is the most relevant for selecting the best training samples.

For RQ3, results presented in Section 4.4 establish 300 images per subject as the size of the dataset that enables approaching the results of the baseline scenario. For our use-case, this resulted in a reduction of 276 images. However, the process of obtaining an optimal number of samples per subject was not considered here and shall be analysed in future work.

Since each training steps takes an equal amount of time to complete, answering RQ4 is straightforward. For our experiment setup, a reduction from 8 to 6 minutes was observed.

Following the specific steps presented in Section 1.4 allowed us to answer the central questions of this dissertation's work and, consequently, it allowed outlining the contributions associated with our findings. Therefore, we emphasise the following contributions:

- Identification of a group of metrics relevant for capturing the diversity of a FR dataset;
- Evaluation of two state-of-the-art heuristics for sampling subsets of data focusing on maximising diversity;
- Analysis of the impact of each sampling approach on a classification model's performance;
- Integration of the proposed approach in the application scenarios of FotoInMotion and CHIC.

Additionally, these contributions resulted in the following scientific outputs:

- Vilaça L., Viana P., Carvalho P., Andrade T. (2020) Improving Audiovisual Content Annotation Through a Semi-automated Process Based on Deep Learning. Advances in Intelligent Systems and Computing, vol 942. Springer, Cham. [93]: Describes the initial experiments regarding the data requirements for a FR application;
- Paula Viana, Pedro Carvalho, Maria Teresa Andrade, Pieter P. Jonker, Vasileios Papanikolaou, Inês N. Teixeira, Luis Vilaça, José P. Pinto, and Tiago Costa. 2020. Semantic Storytelling Automation: A Context-Aware and Metadata-Driven Approach. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). [94]: Integration of this dissertation's findings in the FotoinMotion platform by using FR models in a context-aware video creation;
- Paula Viana, Pedro Carvalho, Maria Teresa Andrade, Inês N. Teixeira, Pieter P. Jonker, Luis Vilaça, José P. Pinto, and Tiago Costa (2020). From a Still Image to a Semantically Aware Video: A Context and Metadata-driven Automatic Media Production Framework, The 17th ACM SIGGRAPH European Conference on Visual

Media Production [95]: Scalable refactoring of FotoinMotion's production environment and impact of these reduced data requirements in the redeployment of DL models;

· Luís Vilaça, Paula Viana, Pedro Carvalho, and Maria Teresa Andrade (2020). Facial Recognition using Guided Data Selection, submitted to a Q1 journal: Details the process and findings of the dataset's construction methodology;

5.3 Future Work

This section discusses some promising directions that can be taken from this work.

Automatic feature extraction with focus on diversity

The feature space defined in Chapter 3 is obtained through a manual feature engineering pipeline. The progress of CV supported by ML/DL has been towards techniques for automatically selecting the best features for a given task. In this way, developing a loss function/heuristic that allows a given model to assign diversity scores to each image or groups of images might enable the model to account for a feature space directly related with diversity. Therefore, this direction might be relevant for scaling the concept presented in this dissertation by taking into account the progress made in this research area.

Optimal sample size

As demonstrated in Section 4.3.3, the entropy scores are different for every subject in the dataset. This suggests that a process for automatically selecting the best sample size for each particular subject is required. Allowing to select the sample size from where the dataset starts to introduce redundant information. As a result, the final subset of images will have the highest amount of information possible and it might even exceed the entropy of the full amount of data, as illustrated in Fig.4.8 and subsequently discussed. Therefore, the development of a heuristic that allows this adjustment can contribute to obtaining even better results with the selected samples.

Translation to state-of-the-art facial recognition approaches

As discussed in Section 4.4, the methods presented in this dissertation use the same feature extractors as SoA FR techniques based on metric learning. It is relevant to analyse the impact of the proposed methodology on those methods by measuring their data requirements since they use datasets with a size larger than 10 times the amount of data used in this dissertation. Therefore, the impact of image sampling methods, as the one proposed, is even greater. These experiments, however, require that the optimal sample size can be obtained for each character to avoid losing a great amount of information due to the unbalanced characteristic of these datasets.

Bibliography

[1] J. Rowley, "The wisdom hierarchy: representations of the dikw hierarchy," *Journal of information science*, vol. 33, no. 2, pp. 163–180, 2007. [cited in page 1]

[2] G. M. D. T. Forecast, "Cisco visual networking index: global mobile data traffic forecast update, 2017–2022," *Update*, vol. 2017, p. 2022, 2019. [cited in page 1]

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016. [cited in page 2, 8, 9, 10, 11, 14]

[4] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 843–852, 2017. [cited in page 2]

[5] RTP - Rádio e Televisão de Portugal, "Covid-19. ministra da saúde fez ponto de situação no telejornal," 2020. [Online; accessed October 21, 2020]. [cited in page 4]

[6] Modern Healthcare, "Health officials seek to block trump rally in virginia," 2020. [Online; accessed October 21, 2020]. [cited in page 4]

[7] M. W. Eysenck and M. T. Keane, *Cognitive psychology: A student's handbook*. Taylor & Francis, 2005. [cited in page 7]

[8] R. C. Atkinson and R. M. Shiffrin, "Chapter: Human memory: A proposed system and its control processes," *The Psychology of Learning and Motivation*, vol. 2, pp. 89–195, 1968. [cited in page 7]

[9] T. M. Mitchell, *Machine Learning*. McGraw-Hill, Inc., 1997. [cited in page 8]

[10] G. W. Rainbolt and S. L. Dwyer, *Critical thinking: The art of argument*. Cengage Learning, 2014. [cited in page 8]

[11] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018. [cited in page 9]

[12] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017. [cited in page 9]

[13] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in Neural Information Processing Systems*, pp. 841–848, 2002. [cited in page 9]

[14] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006. [cited in page 9, 10]

[15] C. Francois, "Deep learning with python," 2017. [cited in page 9, 10]

[16] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997. [cited in page 10]

[17] J. Gama, A. Carvalho, K. Faceli, A. Lorena, and M. Oliveira, "Extracção de conhecimento de dados: data mining," 2017. [cited in page 10]

[18] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th Symposium on Operating Systems Design and Implementation*, pp. 265–283, 2016. [cited in page 12]

[19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, pp. 8026–8037, 2019. [cited in page 12]

[20] S. B. Sells and R. S. Fixott, "Evaluation of research on effects of visual training on visual functions," *American Journal of Ophthalmology*, vol. 44, no. 2, pp. 230–236, 1957. [cited in page 12]

[21] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, p. 106, 1962. [cited in page 12]

[22] A. Gupta, A. A. Efros, and M. Hebert, "Blocks world revisited: Image understanding using qualitative geometry and mechanics," in *European Conference on Computer Vision*, pp. 482–496, Springer, 2010. [cited in page 12]

[23] D. Marr, *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. [cited in page 12]

[24] I. Binford, "Visual perception by computer," in *IEEE Conference of Systems and Control*, 1971. [cited in page 12]

[25] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 67–92, 1973. [cited in page 12]

[26] J. Clement, "Global social networks ranked by number of users 2019," *Statista—The Statistics Portal.*, 2019. [cited in page 13]

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015. [cited in page 13]

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012. [cited in page 13, 15, 16]

[29] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. [cited in page 15]

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. [cited in page 15]

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014. [cited in page 15]

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015. [cited in page 15]

[33] S. Zagoruyko and N. Komodakis, "Wide Residual Networks," *British Machine Vision Conference 2016, BMVC 2016*, vol. 2016-September, pp. 87.1–87.12, may 2016. [cited in page 15]

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. [cited in page 15]

[35] T. Hoeser and C. Kuenzer, "Object detection and image segmentation with deep learning on Earth observation data: A review-part I: Evolution and recent trends," *Remote Sensing*, vol. 12, no. 10, 2020. [cited in page 16]

[36] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018. [cited in page 15]

[37] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8697–8710, 2018. [cited in page 15]

[38] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3025–3032, June 2013. [cited in page 16]

[39] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, *Labeled Faces in the Wild: A Survey*, pp. 189–248. Cham: Springer International Publishing, 2016. [cited in page 16]

[40] M. Wang and W. Deng, "Deep face recognition: A survey," *arXiv preprint arXiv:1804.06655*, 2018. [cited in page 16, 17, 18]

[41] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," 09 2014. [cited in page 17]

[42] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *International Conference on Machine Learning, Deep Learning Workshop*, vol. 2, 2015. [cited in page 17]

[43] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015. [cited in page 17, 21, 22]

[44] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 365–372, Sep. 2009. [cited in page 18]

[45] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4873–4882, 2016. [cited in page 18, 22]

[46] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *International Conference on Automatic Face and Gesture Recognition*, 2018. [cited in page 18, 22, 25, 26]

[47] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, 2020. [cited in page 18]

[48] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, C. Qian, and C. Change Loy, "The devil of face recognition is in the noise," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 765–780, 2018. [cited in page 18]

[49] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference (BMVC)*, 2015. [cited in page 18, 22]

[50] D. Sáez-Trigueros, L. Meng, and M. Hartnett, "Face recognition: From traditional to deep learning methods," *IEEE Transactions on Neural Networks and Learning Systems*, 2018. [cited in page 18]

[51] J. A. Olvera-López, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, and J. Kittler, "A review of instance selection methods," *Artificial Intelligence Review*, vol. 34, no. 2, pp. 133–143, 2010. [cited in page 18, 19]

[52] Z. Abbasi and M. Rahmani, "An Instance Selection Algorithm Based on ReliefF," *International Journal on Artificial Intelligence Tools*, vol. 28, feb 2019. [cited in page 18]

[53] J. A. Olvera-López, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "Prototype selection via prototype relevance," in *Iberoamerican Congress on Pattern Recognition*, pp. 153–160, Springer, 2008. [cited in page 19]

[54] J. C. Riquelme, J. S. Aguilar-Ruiz, and M. Toro, "Finding representative patterns with ordered projections," *Pattern Recognition*, vol. 36, no. 4, pp. 1009–1018, 2003. [cited in page 19]

[55] Y. Caises, A. González, E. Leyva, and R. Pérez, "Scis: combining instance selection methods to increase their effectiveness over a wide range of domains," in *International Conference on Intelligent Data Engineering and Automated Learning*, pp. 17–24, Springer, 2009. [cited in page 19]

[56] B. Spillmann, M. Neuhaus, H. Bunke, E. Pękalska, and R. P. Duin, "Transforming strings to vector spaces using prototype selection," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pp. 287–296, Springer, 2006. [cited in page 19]

[57] J. A. Olvera-López, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "Object selection based on clustering and border objects," in *Computer Recognition Systems 2*, pp. 27–34, Springer, 2007. [cited in page 19]

[58] C. J. Veenman and M. J. Reinders, "The nearest subclass classifier: A compromise between the nearest mean and nearest neighbor classifier," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 9, pp. 1417–1429, 2005. [cited in page 19]

[59] C.-F. Tsai, W.-C. Lin, Y.-H. Hu, and G.-T. Yao, "Under-sampling class imbalanced datasets by combining clustering analysis and instance selection," *Information Sciences*, vol. 477, pp. 47–54, 2019. [cited in page 19]

[60] L. Bogaardt, R. Goncalves, R. Zurita-Milla, and E. Izquierdo-Verdiguier, "Dataset reduction techniques to speed up svd analyses on big geo-datasets," *ISPRS International Journal of Geo-Information*, vol. 8, no. 2, p. 55, 2019. [cited in page 19]

[61] F. U. Nuha *et al.*, "Training dataset reduction on generative adversarial network," *Procedia computer science*, vol. 144, pp. 133–139, 2018. [cited in page 19]

[62] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409, pp. 17–26, 2017. [cited in page 19]

[63] L. Best-Rowden and A. K. Jain, "Automatic face image quality prediction," *CoRR*, vol. abs/1706.09887, 2017. [cited in page 19]

[64] Zhiguang Yang, Haizhou Ai, Bo Wu, Shihong Lao, and Lianhong Cai, "Face pose estimation and its application in video shot selection," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 1, pp. 322–325 Vol.1, Aug 2004. [cited in page 19]

[65] X. Gao, S. Z. Li, R. Liu, and P. Zhang, "Standardization of face image sample quality," in *Advances in Biometrics* (S.-W. Lee and S. Z. Li, eds.), (Berlin, Heidelberg), pp. 242–251, Springer Berlin Heidelberg, 2007. [cited in page 19]

[66] H. Sellahewa and S. A. Jassim, "Image-quality-based adaptive face recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, pp. 805–813, April 2010. [cited in page 19]

[67] J. Ding and X. Li, "An approach for validating quality of datasets for machine learning," in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 2795–2803, Dec 2018. [cited in page 19]

[68] Y. Yao, J. Zhang, F. Shen, L. Liu, F. Zhu, D. Zhang, and H. Shen, "Towards automatic construction of diverse, high-quality image datasets," *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, pp. 1–1, 03 2019. [cited in page 19]

[69] A. Habib, C. Karmakar, and J. Yearwood, "Impact of ecg dataset diversity on generalization of cnn model for detecting qrs complex," *IEEE Access*, vol. 7, pp. 93275–93285, 2019. [cited in page 19]

[70] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu, "Image classification by cross-media active learning with privileged information," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2494–2502, 2016. [cited in page 19]

[71] B. Wu, W. Chen, P. Sun, W. Liu, B. Ghanem, and S. Lyu, "Tagging like humans: Diverse and distinct image annotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7967–7975, 2018. [cited in page 19]

[72] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014. [cited in page 20]

[73] A. Kulesza and B. Taskar, "Determinantal point processes for machine learning," *Foundations and Trends in Machine Learning*, vol. 5, pp. 123–286, jul 2012. [cited in page 20, 39, 40]

[74] A. Kulesza and B. Taskar, "K-dpps: Fixed-size determinantal point processes," in *International Conference on Machine Learning*, 2011. [cited in page 20]

[75] A. Kulesza and B. Taskar, "Learning determinantal point processes," *arXiv preprint arXiv:1202.3738*, 2012. [cited in page 20]

[76] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," in *Advances in Neural Information Processing Systems*, pp. 2069–2077, 2014. [cited in page 20]

[77] J. A. Gillenwater, A. Kulesza, E. Fox, and B. Taskar, "Expectation-maximization for learning determinantal point processes," in *Advances in Neural Information Processing Systems*, pp. 3149–3157, 2014. [cited in page 20]

[78] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, *et al.*, "Iarpa janus benchmark-c: Face dataset and protocol," in *2018 International Conference on Biometrics*, pp. 158–165, IEEE, 2018. [cited in page 22]

[79] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, "Women also snowboard: Overcoming bias in captioning models," in *European Conference on Computer Vision*, pp. 793–811, Springer, 2018. [cited in page 22]

[80] H. J. Ryu, H. Adam, and M. Mitchell, "Inclusivefacenet: Improving face attribute detection with race and gender diversity," *arXiv preprint arXiv:1712.00193*, 2017. [cited in page 22]

[81] W. Wang, X. Alameda-Pineda, D. Xu, P. Fua, E. Ricci, and N. Sebe, "Every smile is unique: Landmark-guided diverse smile generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7083–7092, 2018. [cited in page 22]

[82] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874, June 2014. [cited in page 22, 23]

[83] M. Merler, N. Ratha, R. S. Feris, and J. R. Smith, "Diversity in faces," 2019. [cited in page 22, 23, 24, 25]

[84] "Population: World." `https://www.worldometers.info/geography/7-continents/`. Accessed: 2019-07-16. [cited in page 25]

[85] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009. [cited in page 28, 34]

[86] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013. [cited in page 28, 37, 38]

[87] J. Hestness, S. Narang, N. Ardalani, G. F. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou, "Deep learning scaling is predictable, empirically," *CoRR*, vol. abs/1712.00409, 2017. [cited in page 34, 44]

[88] M. B. Lyons, D. A. Keith, S. R. Phinn, T. J. Mason, and J. Elith, "A comparison of resampling methods for remote sensing classification and accuracy assessment," *Remote sensing of environment*, vol. 208, pp. 145–153, 2018. [cited in page 34]

[89] D. Soekhoe, P. Van der Putten, and A. Plaat, "On the impact of data set size in transfer learning using deep neural networks," in *Advances in Intelligent Data Analysis XV*, pp. 50–60, Springer International Publishing, 2016. [cited in page 34]

[90] C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft, and J. Popp, "Sample size planning for classification models," *Analytica Chimica Acta*, vol. 760, pp. 25 – 33, 2013. [cited in page 34]

[91] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008. [cited in page 35, 41]

[92] G. Gautier, G. Polito, R. Bardenet, and M. Valko, "DPPy: DPP Sampling with Python," *Journal of Machine Learning Research - Machine Learning Open Source Software*, 2019. Code at http://github.com/guilgautier/DPPy/ Documentation at http://dppy.readthedocs.io/. [cited in page 40]

[93] L. Vilaça, P. Viana, P. Carvalho, and T. Andrade, "Improving audiovisual content annotation through a semi-automated process based on deep learning," in *International Conference on Soft Computing and Pattern Recognition*, pp. 66–75, Springer, 2018. [cited in page 50]

[94] P. Viana, P. Carvalho, M. T. Andrade, P. P. Jonker, V. Papanikolaou, I. N. Teixeira, L. Vilaça, J. P. Pinto, and T. Costa, "Semantic storytelling automation: A context-aware and metadata-driven approach," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 4491–4493, 2020. [cited in page 50]

[95] P. Viana, P. Carvalho, M. T. Andrade, I. N. Teixeira, P. P. Jonker, L. Vilaça, J. P. Pinto, and T. Costa, "From a still image to a semantically aware video: A context and metadata-driven automatic media production framework," 2020. [cited in page 51]