Doctoral Thesis

Incremental Learning through Unsupervised Adaptation in Video Face Recognition

Eric López López

2021

UNIVERSIDADE DA CORUÑA

Incremental Learning through Unsupervised Adaptation in Video Face Recognition

Eric López López

Doctoral Thesis

June 2021

PhD Advisors:

Xosé M. Pardo López
Carlos Vázquez Regueiro

PhD Program in Information Technology Research

UNIVERSIDADE DA CORUÑA

Dr. Xosé M. Pardo López   Dr. Carlos Vázquez Regueiro

CERTIFICAN

Que a memoria titulada "*Incremental Learning through Unsupervised Adaptation in Video Face Recognition*" foi realizada por D. Eric López Lopez baixo a nosa dirección no Departamento de Enxeñaría de Computadores da Universidade da Coruña, e conclúe a Tese de Doutoramento que presenta para optar ao grao de Doutor en Tecnoloxías da Información e Comunicación con Mención Internacional.

En A Coruña, 28 Xullo de 2021

Asdo.: Xosé M. Pardo López
Director da Tese de Doutoramento

Asdo.: Carlos Vázquez Regueiro
Director da Tese de Doutoramento

Asdo.: Eric López López
Autor da Tese de Doutoramento

A todas as persoas que honestamente
traballan por un mundo mellor

Recoñecementos

Esta tese representa o conxunto do meu traballo levado a cabo ao longo dos últimos case 5 anos. Nun primeiro momento pode parecer que unha tese se trata dun traballo maiormente persoal. Non obstante, so é necesario reflexionar un instante para decatarse de que a miña contribución so é unha pequena parte de todo o engrenaxe necesario para que esta empresa tivera éxito. Polo tanto, gustaríame utilizar este pequeno espazo para recoñecer o traballo de todos as persoas que directa e indirectamente contribuíron na creación desta tese.

En primeiro lugar gustaríame dedicar unhas liñas á miña contorna persoal. Así, debo comezar por agradecer o apoio (e tamén paciencia) da miña familia durante este período. Gustaríame ademais non limitar o agradecemento a estes últimos 5 anos, e estendelo ao conxunto da miña vida. Un debe ser consciente dos seus privilexios e, no meu caso, o certo é que nunca me podería ter permitido o luxo de realizar esta investigación sen ter nacido e crecido na contorna que eles, en gran medida, crearon para min co seu traballo e esforzo. Paralelamente, tamén debo agradecer a un descubrimento moi especial do meu doutoramento (así como a súa preciosa gata) o seu cariño e apoio continuo, vital para soster este longo esforzo. Finalmente, non me podo esquecer dos meus amigos e compañeiros de traballo que tamén me acompañaron e me axudaron durante toda esta etapa.

En segundo lugar gustaríame recoñecer a todo o entorno académico que fixo posible que levara a cabo toda esta investigación. Dende os meus directores da tese, que me guiaron e aconsellaron durante todo este período; pasando polas dúas profesoras que me guiaron durante os tres meses de estadía na Universidade de Bologna (*Grazzie mile!*); ata o conxunto de profesores que me foron acompañando durante esta longa carreira académica.

E en terceiro lugar, tamén me gustaría lembrar ás institucións públicas (Xunta de Galicia, o Estado Español e máis a Unión Europea) que financiaron tanto o meu tempo de dedicación como as infraestruturas e persoal de apoio que posibilitaron a existencia desta investigación.

A todos, *grazas!*

Eric López López

The measure of intelligence
is the ability to change.

Albert Einstein

Resumo

Durante a última década, os métodos baseados en *deep learning* trouxeron un salto significativo no rendemento dos sistemas de visión artificial. Unha das claves neste éxito foi a creación de grandes conxuntos de datos perfectamente etiquetados para usar durante o adestramento. En certa forma, as redes de *deep learning* resumen esta enorme cantidade datos en prácticos vectores multidimensionais. Por este motivo, cando as diferenzas entre os datos de adestramento e os adquiridos durante o funcionamento dos sistemas (debido a factores como o contexto de adquisición) son especialmente notorias, as redes de *deep learning* son susceptibles de sufrir degradación no rendemento.

Mentres que a solución inmediata a este tipo de problemas sería a de recorrer a unha recolección adicional de imaxes, co seu correspondente proceso de etiquetado, esta dista moito de ser óptima. A gran cantidade de posibles variacións que presenta o mundo visual converten rápido este enfoque nunha tarefa sen fin. Máis aínda cando existen aplicacións específicas nas que esta acción é difícil, ou incluso imposible, de realizar debido a problemas de custos ou de privacidade.

Esta tese propón abordar todos estes problemas usando a perspectiva da adaptación. Así, a hipótese central consiste en asumir que é posible utilizar os datos non etiquetados adquiridos durante o funcionamento para mellorar o rendemento que obteríamos con sistemas de recoñecemento xerais. Para isto, e como proba de concepto, o campo de estudo da tese restrinxiuse ao recoñecemento de caras. Esta é unha aplicación paradigmática na cal o contexto de adquisición pode ser especialmente relevante.

Este traballo comeza examinando as diferenzas intrínsecas entre algúns dos contextos específicos nos que se pode necesitar o recoñecemento de caras e como estas afectan ao rendemento. Desta maneira, comparamos distintas bases de datos (xunto

cos seus contextos) entre elas, usando algúns dos descritores de características máis avanzados e así determinar a necesidade real de adaptación.

A partir desta punto, pasamos a presentar o método novo, que representa a principal contribución da tese: o *Dynamic Ensemble of SVM (De-SVM)*. Este método implementa a capacidade de adaptación utilizando unha aprendizaxe incremental non supervisada na que as súas propias predicións se usan como pseudo-etiquetas durante as actualizacións (a estratexia de auto-adestramento). Os experimentos realizáronse baixo condicións de vídeo-vixilancia, un exemplo paradigmático dun contexto moi específico no que os procesos de etiquetado son particularmente complicados. As ideas claves de De-SVM probáronse en diferentes sub-problemas de recoñecemento de caras: a verificación de caras e recoñecemento de caras en conxunto pechado e en conxunto aberto.

Os resultados acadados mostran un comportamento prometedor en termos de adquisición de coñecemento sen supervisión así como robustez contra impostores. Ademais, este rendemento é capaz de superar a outros métodos do estado da arte que non posúen esta capacidade de adaptación.

Resumen

Durante la última década, los métodos basados en *deep learning* trajeron un salto significativo en el rendimiento de los sistemas de visión artificial. Una de las claves en este éxito fue la creación de grandes conjuntos de datos perfectamente etiquetados para usar durante el entrenamiento. En cierta forma, las redes de *deep learning* resumen esta enorme cantidad datos en prácticos vectores multidimensionales. Por este motivo, cuando las diferencias entre los datos de entrenamiento y los adquiridos durante el funcionamiento de los sistemas (debido a factores como el contexto de adquisición) son especialmente notorias, las redes de *deep learning* son susceptibles de sufrir degradación en el rendimiento.

Mientras que la solución a este tipo de problemas es recurrir a una recolección adicional de imágenes, con su correspondiente proceso de etiquetado, esta dista mucho de ser óptima. La gran cantidad de posibles variaciones que presenta el mundo visual convierten rápido este enfoque en una tarea sin fin. Más aún cuando existen aplicaciones específicas en las que esta acción es difícil, o incluso imposible, de realizar; debido a problemas de costes o de privacidad.

Esta tesis propone abordar todos estos problemas usando la perspectiva de la adaptación. Así, la hipótesis central consiste en asumir que es posible utilizar los datos no etiquetados adquiridos durante el funcionamiento para mejorar el rendimiento que se obtendría con sistemas de reconocimiento generales. Para esto, y como prueba de concepto, el campo de estudio de la tesis se restringió al reconocimiento de caras. Esta es una aplicación paradigmática en la cual el contexto de adquisición puede ser especialmente relevante.

Este trabajo comienza examinando las diferencias entre algunos de los contextos específicos en los que se puede necesitar el reconocimiento de caras y así como sus efectos en términos de rendimiento. De esta manera, comparamos distintas ba-

ses de datos (y sus contextos) entre ellas, usando algunos de los descriptores de características más avanzados para así determinar la necesidad real de adaptación.

A partir de este punto, pasamos a presentar el nuevo método, que representa la principal contribución de la tesis: el *Dynamic Ensemble of SVM (De- SVM)*. Este método implementa la capacidad de adaptación utilizando un aprendizaje incremental no supervisado en la que sus propias predicciones se usan cómo pseudo-etiquetas durante las actualizaciones (la estrategia de auto-entrenamiento). Los experimentos se realizaron bajo condiciones de vídeo-vigilancia, un ejemplo paradigmático de contexto muy específico en el que los procesos de etiquetado son particularmente complicados. Las ideas claves de De- SVM se probaron en varios sub-problemas del reconocimiento de caras: la verificación de caras y reconocimiento de caras de conjunto cerrado y conjunto abierto.

Los resultados muestran un comportamiento prometedor en términos de adquisición de conocimiento así como de robustez contra impostores. Además, este rendimiento es capaz de superar a otros métodos del estado del arte que no poseen esta capacidad de adaptación.

Abstract

In the last decade, deep learning has brought an unprecedented leap forward for computer vision general classification problems. One of the keys to this success is the availability of extensive and wealthy annotated datasets to use as training samples. In some sense, a deep learning network summarises this enormous amount of data into handy vector representations. For this reason, when the differences between training datasets and the data acquired during operation (due to factors such as the acquisition context) are highly marked, end-to-end deep learning methods are susceptible to suffer performance degradation.

While the immediate solution to mitigate these problems is to resort to an additional data collection and its correspondent annotation procedure, this solution is far from optimal. The immeasurable possible variations of the visual world can convert the collection and annotation of data into an endless task. Even more when there are specific applications in which this additional action is difficult or simply not possible to perform due to, among other reasons, cost-related problems or privacy issues.

This Thesis proposes to tackle all these problems from the adaptation point of view. Thus, the central hypothesis assumes that it is possible to use operational data with almost no supervision to improve the performance we would achieve with general-purpose recognition systems. To do so, and as a proof-of-concept, the field of study of this Thesis is restricted to face recognition, a paradigmatic application in which the context of acquisition can be especially relevant.

This work begins by examining the intrinsic differences between some of the face recognition contexts and how they directly affect performance. To do it, we compare different datasets, and their contexts, against each other using some of the most advanced feature representations available to determine the actual need for

adaptation.

From this point, we move to present the novel method, representing the central contribution of the Thesis: the Dynamic Ensembles of SVM (De-SVM). This method implements the adaptation capabilities by performing unsupervised incremental learning using its own predictions as pseudo-labels for the update decision (the self-training strategy). Experiments are performed under video surveillance conditions, a paradigmatic example of a very specific context in which labelling processes are particularly complicated. The core ideas of De-SVM are tested in different face recognition sub-problems: face verification and, the more complex, general closed- and open-set face recognition.

In terms of the achieved results, experiments have shown a promising behaviour in terms of both unsupervised knowledge acquisition and robustness against impostors, surpassing the performances achieved by state-of-the-art non-adaptive methods.

Preface

This preface outlines some of the key ideas and contributions presented in this Thesis: *Incremental Learning through Unsupervised Adaptation in Video Face Recognition.*

Objectives and Work Methodology

The main objective of this Thesis is to develop a method capable of performing online incremental learning in contexts where image quality is relatively low and label availability is extremely limited. In this regard, the problem of face based biometric identification in video-surveillance represent a paradigmatic example of these contexts. Consequently, as proof-of-concept, this will be the main scenario in which we conduct the experiments.

We establish now the main goals (and sub-goals) of the Thesis:

1. Perform a preliminary exploration of the problematic we wanted to approach.
    - Explore the intrinsic differences between some face recognition contexts and their effects in performance
    - Study the behaviour of the existent incremental learning approaches in the absence of labelled data
2. Develop the online incremental learning method.
    - Be able to work in contexts where image quality is specially limited.

    - Design an strategy of continuous adaptation based on incremental learning of the online incoming samples.
    - Overcome the labels limitations to have a successful incremental learning.

3. Validate the method viability in the specific context of face recognition in video-surveillance.

    - Perform a comprehensive evaluation in verification conditions (1:1 identification).
    - Perform a comprehensive evaluation in closed-set and open-set recognition conditions (1:N)

Structure of the Thesis

Following a similar structured as the previous objectives, the Thesis is organised as follows:

- Chapter 1 introduces the central topics approached during the Thesis as well as the main motivation behind them.
- Chapter 2 explores the intrinsic differences between different face recognition contexts and their effects in performance.
- Chapter 3 tackles the simpler face verification problem in video-surveillance problem by presenting the proposed incremental learning system designed to work in very low-labelled conditions.
- Chapter 4 extends the previous system to deal with the more complex open-set face recognition problem under the same context conditions as before.
- Chapter 5 extracts some final conclusions from the Thesis and discusses future research lines in regards to adaptive and incremental learning systems.

Main Contributions

The main original contributions derived from the Thesis are the following:

- About feature representations and adaptation needs on specific environments (Chapter 2):
    - A study of the impact on performance of dataset bias, when combining different face datasets during training phase for face verification purposes.
    - A novel insight into the face dataset bias by a study of their distribution on feature spaces. The approach followed here uses geometrical perspective by looking into the datasets' feature vectors distribution.
    - A comparison of the behaviour of different feature descriptors, both handcrafted and learned, and their robustness against dataset bias. The results shows that even deep feature representations are susceptible to dataset bias.
- About video-to-video face verification in video surveillance (Chapter 3):
    - The proposition of an ensemble-based adaptive biometric system called Dynamic Ensemble of SVM (De-SVM). De-SVM retains the CNN discrimination power while providing for additional modularity, scalability, reversibility, generalisation and robustness of ensemble-based approaches.
    - The use of the self-updating approach to use system's predictions as pseudo-labels to learn and operate simultaneously.
- About video-to-video face recognition in video surveillance (Chapter 4):
    - An extension of the previous approach to unsupervised incremental face recognition: Open-set Dynamic Ensemble of SVM (OSDe-SVM).
    - A strategy to deal with both catastrophic forgetting issues and the effect of mistaken pseudo-labels, taking advantage of ensembles full potential.
    - An approach to instance-incremental learning in the open-set, which could be extended to cope with the class-incremental problem.

  - ◦ A method for person re-identification based on face, which is not directly based on a reservoir of face images and only requires for 5 labelled video frames for initialisation.

Publications from the Thesis

Journal publications

- E. López-López, X. M. Pardo, C. V. Regueiro, R. Iglesias, and F. E. Casado. Dataset bias exposed in face verification. *IET Biometrics*, 8(4):249–258, 2019
- E. Lopez-Lopez, C. V. Regueiro, X. M. Pardo, A. Franco, and A. Lumini. Towards a self-sufficient face verification system. *Expert Systems with Applications*, 174:114734, 2021
- (In Revision) E. Lopez-Lopez, C. V. Regueiro, and X. M. Pardo. Incremental learning from low-labelled stream data in open-set video face recognition, 2020

International conferences

- E. Lopez-Lopez, C. V. Regueiro, X. M. Pardo, A. Franco, and A. Lumini. Incremental learning techniques within a self-updating approach for face verification in video-surveillance. In *Pattern Recognition and Image Analysis*, pages 25–37. Springer International Publishing, 2019
- E. Lopez-Lopez, C. V. Regueiro, and X. M. Pardo. An adaptive video-to-video face identification system based on self-training. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2590–2596, 2021

Developed software

A summarised version of the source code created in this Thesis was made publicly available:

- OSDe-SVM: Open-Set Dynamic Ensembles of SVM: `https://gitlab.citius.usc.es/eric.lopez/osde-svm`

Funding and Technical Resources

For the successful development of this Thesis, it was necessary to rely on series of indispensable means included in the following list:

- Working material, human and financial support primarily by the CITIC and the Computer Architecture Group of the University of A Coruña and CiTIUS of University of Santiago de Compostela, along with a PhD grant funded by Xunta the Galicia and the European Social Fund.
- Access to bibliographical material through the library of the University of A Coruña.
- Additional funding through the following research projects:
    - State funding by the Ministry of Economy and Competitiveness of Spain (project TIN2017-90135-R MINECO, FEDER),
    - The Consellaría de Cultura, Educación e Ordenación Universitaria (accreditations 2016-2019, EDG431G/01 and ED431G/08), and reference competitive groups (2017-2020 ED431C 2017/69, and ED431C 2017/04),
- A three-month research visit to the University of Bologna, Italy; which has play a crucial role in the Thesis development. This research visit was funded by Xunta de Galicia with the grant within the PhD program.

Contents

1. Introduction **1**

2. The Relevance of Context in Face Recognition **7**

2.1. Introduction 7

2.2. The Nature of Dataset Bias 9

2.2.1. Effects of Cross-Dataset Training 11

2.2.2. Datasets' Feature Space 12

2.3. Datasets of Study 13

2.3.1. Dataset Pre-processing 15

2.4. Feature Representation 15

2.4.1. LBP over Landmarks 16

2.4.2. VGG-Face 17

2.4.3. ResNet 17

2.5. Effects of Cross-Dataset Training 18

2.5.1. Experimental Results 19

2.6. Datasets' Feature Space 24

2.6.1. Building the subsets 24

2.6.2. Metrics in the Feature Space 25
2.6.3. Nearest Neighbour Search 26
2.6.4. Experimental Results 27
2.7. Conclusions 31

3. An Adaptive Face Verification in Video-surveillance 33

3.1. Introduction 33
3.2. Related work 35
3.3. Adaptive Biometrics for Face Verification in Video-Surveillance 37
3.3.1. Self-updating Pipeline 37
3.3.2. The Proposed Dynamic Ensemble of SVM (De-SVM) 39
3.3.3. Other Explored Classification Methods 42
3.4. Methodology 43
3.4.1. Datasets 43
3.4.2. Face detection and feature extractor 45
3.4.3. Testing Protocol 45
3.4.4. The *Operational Threshold* 47
3.4.5. Metrics 48
3.5. Experiments and Results 48
3.5.1. Supervised Adaptation 48
3.5.2. Unsupervised Adaptation 52
3.5.3. Post-robustness impostor testing 55
3.5.4. Summary and Discussion 58
3.5.5. Further testing on De-SVM 60

3.6. Conclusions 62

4. Adaptive Face Recognition in Video-surveillance 63

4.1. Introduction 63

4.2. Related Work 65

4.3. Open-Set Dynamic Ensembles of SVM (OSDe-SVM) 68

4.3.1. Ensemble's Decision Functions 69

4.3.2. Update Module: Incremental learning based on Self-Updating 71

4.3.3. Limitation Module 72

4.3.4. Self-healing: Correcting Wrong Updates 73

4.4. Methodology 74

4.4.1. Database Selection 74

4.4.2. Experimental Set-up 77

4.5. Experiments and Results 78

4.5.1. Performance vs. Universe Size 79

4.5.2. Temporal Evolution 81

4.5.3. Comparison against state-of-the-art face recognition models. . 82

4.5.4. Performance vs. Openness 83

4.6. Conclusions 84

5. Conclusions and future work 87

References 91

A. Resumen extendido en galego 107

List of Tables

2.1. Summary of dataset characteristics. 14

2.2. Summary of samples of each dataset used after the face detection. . 16

2.3. Type and dimensionality of each feature detector. 17

2.4. Configuration of *Dataset A* and *Dataset B* in the training phase and in the two different tests (1 and 2). 18

2.5. LBP. Drop in performance (TAR @ FAR 0.001) of Test 1 respect to Test 2 for each combination of *Dataset A* and *B*. 20

2.6. VGG-Face. Drop in performance (TAR @ FAR 0.001) of Test 1 respect to Test 2 for each combination of *Dataset A* and *B*. 21

2.7. ResNet. Drop in performance (TAR @ FAR 0.001) of Test 1 respect to Test 2 for each combination of *Dataset A* and *B*. 22

2.8. Distribution of nearest neighbors over each dataset in % using different features. 28

3.1. Summary of the decision rules with each method. 40

3.2. YouTube Faces distribution of the amount of videos per person after the face detection phase. 43

3.3. How the datasets' identities are divided in order to generate each subset defined in Sec. 3.4.3. 46

3.4. Summary of TAR@FAR1% performances values obtained (values in %). Uncertainty is not represented in previous graphs for the sake of clarity. SU stands for self-updating. 59

3.5. COX: Comparison on the function used in the FDF and SDF. 60

3.6. COX: Study on FAR point of operational threshold De-SVM (template size = 5 frames). 61

3.7. COX: Study on template size De-SVM at operational threshold 5%. . 62

4.1. Performance over different universe sizes, while preserving the ratio with the number of IoI. Values are expressed as $\mu(\sigma)$, where μ stands for mean and σ for standard deviation. 81

4.2. Comparison against state-of-the-art face recognition models: FaceNet [121] and RN100-AF (ArcFace) [26], (for the case of 50 IoI in an universe of 100). 82

List of Figures

2.1. t-SNE representation of features of a random subset of samples drawn from three datasets with three feature descriptors: (Red) IJB-A; (Green) LFW; (Blue) FERET. 10

2.2. Sample images from each dataset used for the experiments. 14

2.3. Extraction of LBP features. Left: locations of the 51 inner landmarks used for feature extraction. Right: 4x4 cell centered on one of the landmarks. 17

2.4. Average TAR drop between Test 1 and Test 2 respect to the FAR point in which we perform the measure, using FS just as *Dataset A*. . 25

2.5. Scheme of the subsets generated for each dataset (Dataset 1, Dataset 2, ..., Dataset D_i, ..., Dataset N_D). Blue triangles represent *probe sets* and green ellipses represent the part of the dataset used to generate the *gallery set*. In this case, $N_p = 5$ *probe sets* were generated for the first dataset, and $N_p = 2$ *probe sets* for the rest. 26

2.6. A t-SNE representation of the 6 *gallery sets* taken from each dataset of study for three feature descriptors. (Red) FS; (Green) IJB-A; (Blue) LFW; (Purple) O2FN; (Olive) MobBIO; (Black) FERET. 31

3.1. The designed pipeline for the implementation of the self-updating strategy with its ideal behaviour. Whenever the target identity (the blue jersey one) appears, the model is updated. Otherwise, the model is maintained. 38

3.2. The pipeline of De-SVM within the self-updating general strategy. The model (M_t) is updated based on its decision over the query sequences (S_{t+1}) to generate a new model (M_{t+1}). Three different functions are involved in the decision of updating (FDF and SDF) and the way of updating (UF). 41

3.3. Samples of both datasets, COX (left) and YTF (right). 44

3.4. Example of division of `cam1` and `cam2` in the COX Face database to generate the query sub-sequences (SS stands for sub-sequence). 44

3.5. COX: Supervised updating performance comparison. 49

3.6. YouTube Faces: Supervised updating performance comparison. 49

3.7. Evolution of the ROC curve and the ROC point associated to the *operational threshold* after each query for the supervised case using the SVM classification model. 51

3.8. COX: Threshold evolution of a same FAR point of the ROC curve (Data for 1%, 5%, 10%, 25% and 50%). 52

3.9. FACE COX: Self-supervised experiment using an operational threshold of 5% initial FAR. 53

3.10. YouTube Faces: Self-supervised experiment using an operational threshold of 5% initial FAR. 53

3.11. Genuine and impostor updates performed by each classification technique in each database. 56

3.12. FACE COX: Self-supervised performance comparison fixing the operational threshold at 5% initial FAR, testing robustness. 57

3.13. YouTube Faces: Self-supervised performance comparison fixing the operational threshold at 5% initial FAR, testing robustness. 57

4.1. Open-Set Dynamic Ensembles of SVM (OSDe-SVM) is able to incorporate new knowledge and correcting wrong updates by adding and removing classifiers in an unsupervised way. The system is designed to work under open-set recognition conditions. 65

4.2. Pipeline of OSDe-SVM. 69

4.3. Examples of Extreme Values distributions and application of RDF. A Weibull function is fitted to the min distribution to see whether the candidate belongs or not to this distribution. First row illustrates examples of *known* identities and second row does the same with the *unknwon* ones. 71

4.4. Pipeline of OSDe-SVM when self-healing is performed. 74

4.5. Performance versus image resolution, scales 1, 1/2, 1/4, 1/8 and 1/16 (that is, 112x112, 56x56, 28x28, 14x14, and 7x7 pixel image sizes). . 76

4.6. Adaptation steps performed during the experiments. INI stands for initialisation, UP for update and SH for self-healing. The last step corresponds to the beginning of the second lap. 78

4.7. Performance under different universe sizes (20, 40, 60, 100, 200, 400, 600, and 1000) and same ratio between number of IoI and universe size (1:2). 80

4.8. Evolution of OSDe-SVM for the case of 50 IoI over an universe of 100 identities. 82

4.9. Performance openness dependence with fixed IoI (50), and universe $\in \{50, 100, 200, 400, 600, 1000\}$. 84

Chapter 1

Introduction

Motivations

Humans acquire information from the outer world through five different senses — vision, hearing, touch, taste and smell — from which vision stands out as the most important human senses. As a matter of fact, the amount of raw information received by the eyes is estimated to be at least two orders of magnitude above the joint information received by the other senses [24]. For this amount of information to be manageable, outer world visual information needs to be compressed, classified and structured. Consequently, the set of mechanisms shaping what we call vision are not mere light receptors (eyes), but involve a specific part of the brain (the visual cortex). If that was not enough complexity, the adult human vision is not a static ability that does not evolve with time. Indeed, part of vision involves pretty stable processes developed during child development or by just evolution hard-wiring. Nevertheless, there is also an essential part of the vision which is *modulated and adapted through active use.*

Computer vision is born with the complex task of replicating the human vision system. Although we do not entirely understand its natural mechanisms, the enormous potential in automation applications of an artificial vision system has broken through. The research in computer vision involves a wide range of inter-disciplinary experts (from engineers, psychologists, biologists, to even philosophers, among others). Since its birth as a summer project [96] in the '60s, the advances have been mostly gradual until the last decade. In 2012, the re-visitation of a specific type of neural networks [67] (the Convolutional Neural Networks, CNN) in conjunction with the increasingly powerful GPU's and the availability of large-scale annotated datasets [25] provoked an important leap forward in performance, in what can be called deep learning revolution. Suddenly, some of the most challenging tasks of computer vision (e.g. car drive assistance, image retrieval, face swap filters, etc.) started to be accessible by using these techniques.

From the biological perspective, CNNs can be associated with the more static part of the human visual system. However, currently, it fails to provide this capability of *adaptation through active use*. When we try to fine-tune a pre-trained network using a small amount of data for adaptation purposes, the network tends to forget its original capabilities in problem denominated as *catastrophic forgetting*. This issue is problematic both from a more theoretical perspective; adaptation should be a core ability of any truly intelligent system; and a more practical one, which we are about to analyse next.

The power of CNNs is closely related to the quality and completeness of the large-scale annotated dataset used in training. In this regard, even the most comprehensive dataset cannot guarantee a good performance in every specific context. While collecting (and annotating) additional data seems like an immediate solution, the truth is that there are contexts in which this endeavour is particularly challenging (due to privacy issues, limitation in resources, cost-related problems, etc.). Therefore to endow recognition systems with adaptation capabilities appears as a practical and efficient way of dealing with these issues. In the literature, this adaptation capability is implemented by different inter-related research lines: *transfer learning*, *lifelong learning*, *domain adaptation*, *continual learning* or *incremental learning* [38].

The challenges of this Thesis

This Thesis precisely aims to contribute to developing the unavoidable adaptation part of computer vision. Due to the wide range of computer vision applications and as a proof-of-concept, we have restricted our study to the task of face recognition.

Face recognition is a task of computer vision that has consistently received significant attention from the scientific community. Roughly, it consists of assigning identities based on facial images. Nevertheless, in practice, face recognition research includes many sub-problems (e.g. face verification or face identification) and the additional tools involved in the identity assignation (e.g. face detection, face tracking or feature representation). Besides, the fact that we distinguish between very similar objects (faces) makes the effects of the context of acquisition particularly relevant. Intuitively, the difference in a person's appearance in video surveillance and an ID mugshot can be more significant than distinguishing them from other people. For these reasons, face recognition appears as an interesting scenario to benefit from adaptation capable methods. With all of this in mind, we now enumerate the main challenges of the Thesis:

The **first challenge** consist of conceptually defining how we plan to retain the CNNs discrimination power while avoiding *catastrophic forgetting*. In this direction, we opted to fix the convolutional layers of state-of-the-art deep networks without the last fully connected layers. This way, we re-take the traditional distinction between feature extraction and classification parts and implement the adaptation capabilities by improving this last part.

The **second challenge** relates to the robustness degree of state-of-the-art face recognition features to context variations. Therefore, the first step in this direction will be to study the performance effects when we mix samples of two contexts during the classifier training. From the face recognition perspective, we try to understand the impact of using auxiliary data sources (often taken in a different context) when we deploy a face verification system in a specific context where availability of labelled data is scarce (i.e. mobile biometrics, video surveillance, etc.).

In this endeavour, we perform a comprehensive study of the *dataset bias* using some well-known face datasets, making an especial emphasis on context-related biases. The results of this study can be seen in Chapter 2.

The **third challenge** consists of the actual design of a novel classification part that implements adaptation. The proposed solution is an ensemble-based incremental learning method built on top of the most advanced deep feature embeddings. The deep network used for the feature extraction is not relevant, and it will be changing during the Thesis development as the state-of-the-art advances. The method is a *Dynamic Ensemble of SVM (De-SVM)* which will be adding new classifiers incrementally to adapt to the target context and improve performance. Ensembles have often been recalled as an efficient, robust and scalable solution for classification tasks compared to the regular fully connected layers of CNN. Besides, by encapsulating updates into individual classifiers, they also provide for update reversibility and better explainability. This especially interesting if we want to operate in label-scarce contexts, as we further explain in the following proposal.

The **fourth challenge** addresses the scarceness of labelled data of the contexts we tackle. In this regard, the proposed *De-SVM* uses the self-training paradigm to use the predictions of the method at the moment as pseudo-labels to add new classifiers to the ensemble. The training process becomes integrated with the system's operation without the need for the usual separation. Thus, after initialisation (using just 5 video frames), the system will perform the incremental learning process in a completely unsupervised way.

Decisions within the ensemble use the raw score output by the SVM classifiers to then fused them using the median function. Finally, acceptance or rejection are determined based on a threshold. The setting of this threshold has proven to be particularly challenging and relates to the *stability-plasticity dilemma.*

Chapter 3 contains the details of De-SVM architecture and also a comprehensive comparison to other incremental learning methods within the self-training approach. The whole set of experiments mimic the video-to-video face verification (V2V-FV) problem in a video-surveillance context, a paradigmatic example of the specific context in which adaption may be required. *De-SVM* stand outs as the best performing method with respect to both other incremental learning techniques as well as other non-adaptable state-of-the-art face verification methods.

The **fifth challenge** relates to the more complex conditions of open-set recognition. In this direction, we will need to extend De-SVM into the so-called *Open-set Dynamic Ensembles of SVM (OSDe-SVM)* in a two-step process:

- The adaptation of the decision procedure to the open-set multi-class problem.

In this setting, the decision needs to expect not only queries from the enrolled identities but also queries from additional *unknown* identities. For this purpose, we use the power of the Extreme Value Theory to discriminate between *known* and *unknown* identities.

- The addition of two new modules that remove classifiers from the ensemble to limit their size (limitation module) and correct possible mistaken updates (self-healing module). This way allows us to exploit the full potential of an ensemble-based approach.

Chapter 4 contains full details of *OSDe-SVM* as well as a comprehensive experimentation in the same video-surveillance context as before.

Main Contributions

The main original contributions derived from the Thesis are the following:

- About feature representations and adaptation needs on specific environments (Chapter 2):
    - A study of the impact on performance of dataset bias, when combining different face datasets during training phase for face verification purposes.
    - A novel insight into the face dataset bias by a study of their distribution on feature spaces. The approach followed here uses geometrical perspective by looking into the datasets' feature vectors distribution.
    - A comparison of the behaviour of different feature descriptors, both handcrafted and learned, and their robustness against dataset bias. The results shows that even deep feature representations are susceptible to dataset bias.
- About video-to-video face verification in video surveillance (Chapter 3):
    - The proposition of an ensemble-based adaptive biometric system called Dynamic Ensemble of SVM (De-SVM). De-SVM retains the CNN discrimination power while providing for additional modularity, scalability, reversibility, generalisation and robustness of ensemble-based approaches.

- The use of the self-updating approach to use system's predictions as pseudo-labels to learn and operate simultaneously.

- About video-to-video face recognition in video surveillance (Chapter 4):
  - An extension of the previous approach to unsupervised incremental face recognition: Open-set Dynamic Ensemble of SVM (OSDe-SVM).
  - A strategy to deal with both catastrophic forgetting issues and the effect of mistaken pseudo-labels, taking advantage of ensembles full potential.
  - An approach to instance-incremental learning in the open-set, which could be extended to cope with the class-incremental problem.
  - A method for person re-identification based on face, which is not directly based on a reservoir of face images and only requires for 5 labelled video frames for initialisation.

Structure of the Thesis

Following a similar structured as the previous objectives, the Thesis is organised as follows:

- Chapter 1 introduces the central topics approached during the Thesis as well as the main motivation behind them.

- Chapter 2 explores the intrinsic differences between different face recognition contexts and their effects in performance.

- Chapter 3 tackles the simpler face verification problem in video-surveillance problem by presenting the proposed incremental learning system designed to work in very low-labelled conditions.

- Chapter 4 extends the previous system to deal with the more complex open-set face recognition problem under the same context conditions as before.

- Chapter 5 extracts some final conclusions from the Thesis and discusses future research lines in regards to adaptive and incremental learning systems.

Chapter 2

The Relevance of Context in Face Recognition

This chapter is heavily based on the contents of this article:

E. López-López, X. M. Pardo, C. V. Regueiro, R. Iglesias, and F. E. Casado. Dataset bias exposed in face verification. *IET Biometrics*, 8(4):249–258, 2019

2.1. Introduction

The outstanding diversity of real-world scenarios makes generalisation one of the trickiest aspects in the development of visual recognition systems [90]. Intuitively, systems designed for a specific application context find it difficult to generalise to more general ones. Notwithstanding, the same is true when transferring systems intended for general context into more specific ones. The underneath explanation for this behaviour relates to the different distributions between source and target domains [131, 125, 141]; and its performance impact is unavoidable to consider. For instance, there have been shown the existence of substantial disparities in gender classification performance across different demographic cohorts due to biases in training datasets [11].

One possible way to study the influence of this problem is through the intrinsic differences between existing datasets or, in other words, through dataset bias.

A wide variety of datasets (either more general or more specific ones) have different built-in contexts of acquisition. In the particular case of face verification, the main task consists of discriminating between identity-carrying features μ (e.g. facial characteristics, ethnicity, gender, etc.), while discarding the identity-irrelevant ones, ϵ (e.g. haircut, makeup, injuries, ageing, illumination, pose, etc.) [15]. The development of more and more extensive face datasets have been of paramount importance for the progress in the general field of face verification. They have paved the way to the development of (data-intensive) deep learning methods, which have achieved impressive performance [64, 25, 55, 15]. Nonetheless, face verification in some data-scarce specific domains remains a challenging task.

Face verification is a useful utility in a wide variety of target domains (e.g. biometrics in mobile devices, video surveillance, or mugshot verification) whose data distributions are different from those in the public domain face datasets but are also different among them. For instance, images generated by smartphone users (equipped with a non-collaborative face verification system) depend on users' behaviour, habits, or transitory appearance changes. For its part, in a video-surveillance context, the spectrum of camera poses and resolutions, scales or blurring effects creates a very specific sample's distribution compared to one of the high-quality and pose-constrained mugshots.

Suppose we train a system to authenticate identity in these contexts using negative samples drawn from a general (i.e. with a diverse pose, illumination and capturing conditions) face dataset. In that case, some identity-irrelevant features could be misleadingly identified as relevant ones. In other words, some μ features would be considered as part of ϵ features, and verification would be partially based on context-specific features, thereby leading to poor performance.

To tackle these issues, datasets are usually augmented by different means (generating synthetic faces images [65, 53]; applying several transformations to enrich the dataset with new simulated poses, resolutions, blurring effects, or lighting [97, 21, 108]), or combined through domain-adaptation approaches [124, 145, 82, 141].

Dataset bias has received surprisingly little attention from the scientific community. Nevertheless, there are some previous works for the case of object recognition [131, 129]. In the case of face datasets, detecting bias represents a more significant challenge than doing so in object datasets. While finding object datasets sharing categories (e.g. car, tree, or building) is relatively easy, this is not the case for face datasets, as they usually do not share identities (categories). Hence, performing a cross-dataset analysis to explore how each dataset represents some common identities and how to exploit their differences is not possible, in contrast to what happens with objects datasets [115, 59].

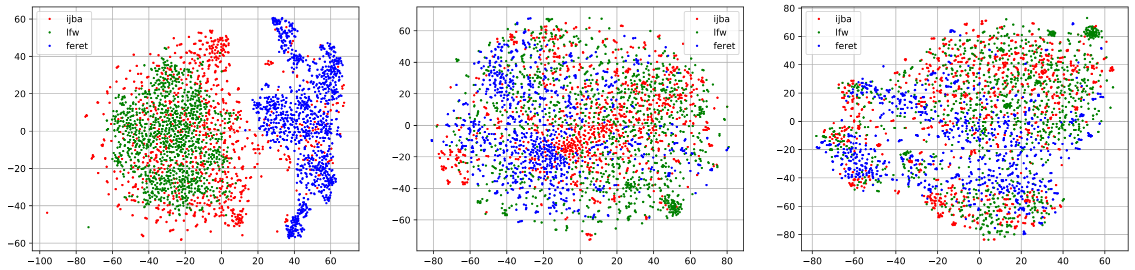
In summary, while many face verification methods assume that training and testing sets are composed of independent and identically distributed samples, this assumption does not hold in many real-world applications. In most cases, these sets contain samples drawn from different populations (and distributions). Motivated by these facts, the main goal of this chapter is to shed light on these differences and their potential harms. In this regard, the main contributions are:

- A study of the impact on the performance of dataset bias, when combining different face datasets during the training phase for face verification purposes (Sections 2.2.1 and 2.5).
- A novel insight into the face dataset bias by studying their distribution on feature spaces. The approach followed here uses geometrical perspective by looking into the datasets' feature vectors distribution (Sections 2.2.2 and 2.6).
- A comparison of the behaviour of different feature descriptors (Section 2.4), both handcrafted and learned, and their robustness against dataset bias.

2.2. The Nature of Dataset Bias

The usual performance differences between the lab and real-world are clear proof of the dataset bias existence. Of course, it could be said that biases are everywhere, but are there any specific causes of bias? Is there any way to palliate its effects?

Pioneering the studies on dataset bias, [131] centres its attention on the object recognition field. This study released moments before the deep learning revolution [69]; therefore, restricted to handcrafted descriptors, namely the HOG descriptor. Posterior works include learned features and continue to remark dataset bias as a



(a) LBP. (b) VGG. (c) ResNet.

Figure 2.1: t-SNE representation of features of a random subset of samples drawn from three datasets with three feature descriptors: (Red) IJB-A; (Green) LFW; (Blue) FERET.

relevant problem for object recognition [129]. They both distinguish four different kinds of bias according to their nature:

- *Selection bias* is related to how images are collected (keywords search, manual selection, crowd-sourcing collection, etc.).
- *Capture bias* comes from how images are captured (type of device, the context of the acquisition, etc.).
- *Label bias* relates to a poor semantic annotation of the dataset, where any labelling mechanism could assign a different label to the same object ("screen" vs "TV"; "grass" vs "lawn"). This bias is not applicable in face verification since labels here are perfectly defined.
- *Negative set bias* relates to how the dataset samples the *rest of the world* (the *rest of faces* in our specific case). If this set is imbalanced or not representative, the model generated with it will have problems generalising.

In regards to the second question, domain adaptation methods are often presented as a way to tackle dataset bias issues [35, 132]. Notwithstanding, a domain is a more general concept than a dataset. For example, in face verification, one can define a domain as the one which only includes grey-scale images, the one with comprehensive coverage of face poses, the one specially built for video surveillance recognition, etc. Domain adaptation tools help to move or transfer knowledge between different domains. In the specific case of face recognition, face-frontalisation methods (methods that convert any face to its frontal view, allowing to eliminate the pose bias) [48, 49, 2] represent a way of domain adaptation to work just with frontal faces.

Conventionally, a domain could also be defined as the one representing the acquisition procedure of a particular dataset. Thus, a domain adaptation technique could help to 'hop' between datasets. Nevertheless, we cannot say that domain adaptation techniques solve dataset problems. For example, suppose we have two different datasets with two different biases. In that case, one could use a domain adaptation technique to transfer knowledge from one of them to another (or even to a common one). However, we could still end with a biased model. At the end of the day, since the final aim should be to represent the visual world correctly, an unbiased dataset would be still needed.

2.2.1. Effects of Cross-Dataset Training

In the development of a face verification system for a specific context, the ideal scenario would be to build an ad-hoc representative dataset of the target domain. Nevertheless, there often are barriers (financial cost, time or privacy-related issues) that convert its construction as something unreachable in many scenarios.

Consider the previous specific scenario of face verification in mobile devices. During operation, the collection of positive (identity of interest) samples to build the verification model is necessarily performed in the actual operational domain (henceforth *Dataset A*). Notwithstanding, when the construction of a complete dataset is not an option, one common strategy is to rely on images from other auxiliary sources (henceforth *Dataset B*) to use as a negative set during the model creation phase. In this setting, the acquisition context of the positive and the negative sets will hardly be the same. Therefore, the first objective will be to study how this configuration affects generalisation on the target domain.

This set-up resembles the one used in [131] for the analysis of the negative set bias in the frame of the object recognition field. This work explores the performance impact of using samples drawn from different datasets to build the negative training set.

2.2.2. Datasets' Feature Space

Conventionally, the image classification pipeline can be divided into feature extraction and actual classification. This dissociation is quite common in the literature. For example, in a face verification context, some works use pools of feature extractors to establish a benchmark [85, 127, 70]. In [16], the author exchange handcrafted and learned features within the same system to push performance to the state-of-the-art. Both parts are not innocuous in relation to dataset bias and can potentially introduce or (desirably) eliminate underline bias in data.

In the previous section, the main focus was to observe dataset bias by its effects on actual classifier performance. Here, the aim is to perform a more direct observation by staying with the feature extraction part. One initial approach to observe biases in the feature vector data could be to look into their representation. Nevertheless, the high dimensionality makes the actual shapes of the samples' distribution difficult to determine, as t-SNE [84] representation of Fig. 2.1 shows. As an alternative, (dis)similarity among datasets can be estimated from distances among feature vectors of different datasets. The hypothesis is that samples from a particular dataset should tend to have the same probability of finding nearby vectors from other datasets with similar distributions as from their own dataset.

Besides, distances are something especially relevant in the case of face verification. The distance between two feature vectors is common practice in identity verification. Feature extractors (especially the deep learning-based ones) are designed so that the same identities tend to be closer [121, 99, 85]. Distances between feature vectors are also used in [127] for both verifications as other related tasks as attribute detection (ethnic, male/female, age, etc.). A cosine dissimilarity metric is also used in [70] over different kinds of descriptors to compare the performance of a certain face descriptor. Consequently, anomalies in these distances provoked by datasets will directly lead to problems in real-world performance.

This neighbour search can be seen from multiples points of view. First, it can serve as an insight into the dataset samples distribution. Second, it also can be seen as the *Name that Dataset!* experiment of [131] where the belonging dataset is guessed using just its neighbours. Finally, another interpretation relates to the dataset origin of hard negatives. In other words, the harder to classify samples.

2.3. Datasets of Study

This study includes a total of six different face datasets (Fig. 2.2). Half of them were built gathering images captured with mobile devices (two with and one without users' collaboration), and the other half contains images taken in a range of different (general) contexts. Next, the specific characteristics of each one (see Tab. 2.1) are described:

- **FERET** [103] is one of the first datasets that tried to become a standard for face recognition, both for training and testing processes. It consists of a total of 8 525 images of 1 109 people with a range of different (annotated) poses taken in a highly controlled environment. This study is restricted to just the *dvd-1* data.
- **Labelled Faces in the Wild**[1] (LFW) [55] contains more than 13 000 face images of a total of 5 749 people collected from the web. Each face was annotated with the name of the person, and 1 680 of the identities have two or more distinct photos in the dataset. It is one of the first datasets aimed at coping with the *unconstrained* face recognition problem. Images were gathered from the internet, and faces were detected using the Viola-Jones detector [136], which introduced a bias in the range of possible poses.
- **IARPA Janus Benchmark A*** (IJB-A) dataset [64] contains a total of 5 712 images of 500 identities (≈11 images per subject). The most distinctive characteristic of this dataset is the elimination of the bias in face detection due to the fact that the complete dataset was manually annotated using crowdsourcing methods. It has been recently updated with the **IARPA Janus Benchmark B** dataset [142] in which the number of images has been increased to 21 728 (1 845 different identities), and the **IARPA Janus Benchmark C**, which even added up more images from video frames.
- **O2FN** dataset [111] contains 2 000 face images taken from 50 different subjects predominantly of Asian ethnic. Images are self-taken photos using a mobile phone in a collaborative context. Subjects were asked to take approximately

[1] It was necessary to eliminate the overlapped identities between IJB-A and LFW datasets. The procedure was to eliminate from LFW the 183 identities also present in IJB-A.

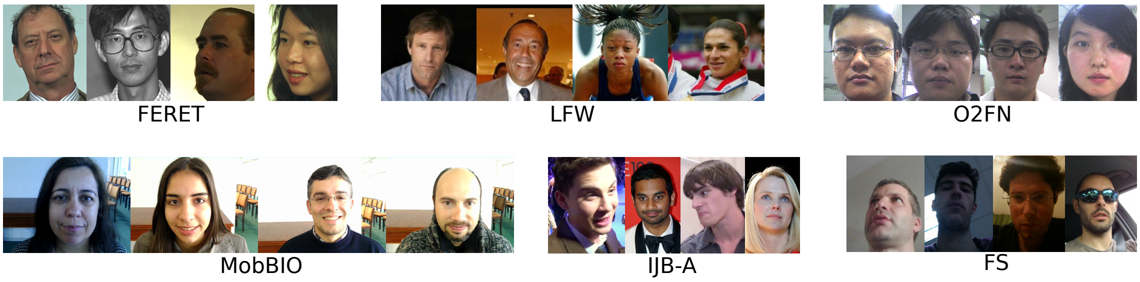


Figure 2.2: Sample images from each dataset used for the experiments.

Table 2.1: Summary of dataset characteristics.

Dataset	**Context**	**Pose Variation**	**Controlled Environment**	**Illumination**	**Ethnicity**
FERET	General	Full	High	Indoors	Caucasian
LFW	General	Limited	Low	Varied	Varied
IJB-A	General	Full	Low	Varied	Varied
O2FN	Mobile	Limited	Intermediate	Varied	Asian
MobBIO	Mobile	Limited	High	Indoors	Caucasian
FS	Mobile	Limited	Low	Varied	Caucasian

20 indoor images and 20 outdoor images, with limited variations in facial expression and out-plane rotations.

- **MobBIO** multimodal dataset [122] was specifically designed for biometrics. It contains data of faces, voice and iris of 105 identities. In the case of facial images, which is the part of our interest, there is a total of 1 640 photographs (≈16 images per identity) taken with mobile devices in a controlled environment, with a limited pose and illumination variations.

- **FaceSampler** (FS) dataset has been created using the frontal camera of mobile phones in a non-collaborative context. It consists of a total of 2 102 images of a total of 15 different identities. The acquisition system was designed to use inertial information in order to maximize the probability of the existence of faces in the frame, although only the images with a detected face were gathered. Image collection was running in the background while users were using the mobile phone [12]. Up to our knowledge, this is the unique dataset that was built on the operational domain for the problem of non-collaborative verification of users of mobile devices.

2.3.1. Dataset Pre-processing

For face detection purposes, the election was the tool provided by Dlib library [61] (based on HOG features) due to its perfect integration with facial landmark detector also implemented in the same library. Here, indeed, an important constraint was introduced. For example, the main contribution of the IJB-A dataset is the elimination of the frontal faces bias (in most cases due to Viola-Jones detection [136]) by performing a manual annotation of the data.

However, this decision is justified by, mainly, two reasons. First, the vast majority of face recognition methods (including those tested here) contemplate some face detection method on their pipelines. Second, by setting a fixed face detector, it can be assumed that any behaviour of the data will not be related to the face detector. In this sense, we could consider it as part of the context of the study, which could be called *universe of possible faces* detected by the face detector. Tab. 2.2 shows the aftermath of the detection process.

2.4. Feature Representation

Feature extractors can be more or less sensitive to specific characteristics, so their election is of paramount importance. For example, a hypothetical feature extractor that is perfectly robust to face pose variations would not reflect the feature space differences between a dataset with only frontal faces and a dataset richer in poses. Of course, this does not mean that the first dataset is not biased towards a particular pose, but the feature extractor can ignore this fact. Besides, this behaviour can be desirable (in a general context) or not (in a specific application's context, poses can be biased, for instance, when looking at a smartphone screen).

This degree of robustness is almost impossible to achieve with general-purpose hand-crafted features since they were not designed for this specific application domain. When using these features, only the classifier can achieve some degree of adaptation throughout the training phase. On the contrary, deep features are trained, so it will be possible to create a more adjustable feature set depending on the application.

This study includes three feature representations with different abstraction and generalisation abilities: LBP over landmarks, VGG-Face and ResNet. All of them have been used in state-of-the-art of face verification.

2.4.1. LBP over Landmarks

Bayesian approaches are some of the few methods which, while using handcrafted features to encode information of faces, still obtain state-of-the-art results ([16] 96.33% accuracy in LFW). So, we have included LBP features over facial landmarks in our analysis.

To obtain a face description, first of all, a facial landmark detection [57] was performed in order to locate a total of 68 points on face images. These landmarks serve as key points for the similarity transformation that rectifies the image. After that, patches centred around just 51 of the inner landmarks (Fig. 2.3 left) are extracted at two different scales. The side lengths of the image at the two scales were 180 and 118 pixels. The patch size is 40 × 40 in both scales. Each patch was divided into 4 × 4 non-overlapped cells (Fig. 2.3 right). Finally, a uniform-LBP vector was computed for each cell, and all of them were concatenated on the final feature vector of 96 288 dimensions (see Tab. 2.3), following [93].

Table 2.2: Summary of samples of each dataset used after the face detection.

Dataset	**Identities**	**Images**
FERET	739	4 929
LFW	5 566	10 966
IJB-A	500	19 427
O2FN	50	1 720
MobBIO	100	1 599
FS	15	2 102

Table 2.3: Type and dimensionality of each feature detector.

	Type	Dimensions
LBP OVER LANDMARKS	Handcrafted	96 288
VGG-FACE	CNN Learned	4 096
RESNET	CNN Learned	128

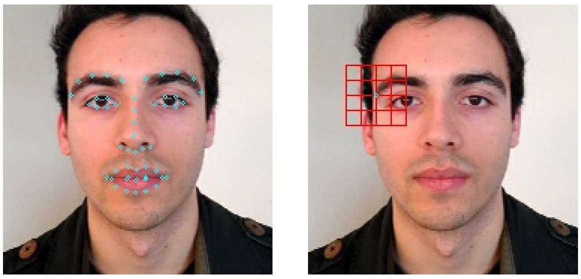


Figure 2.3: Extraction of LBP features. Left: locations of the 51 inner landmarks used for feature extraction. Right: 4x4 cell centered on one of the landmarks.

2.4.2. VGG-Face

The VGG-Face CNN proposed in [99] for general face recognition has achieved an extraordinary value of accuracy of 98.95% on the LFW dataset. This CNN was trained using a large-scale dataset of 2.9 million images of 2 600 people. The first layers of its architecture, discarding the last fully connected layer, were used to extract 4 096 dimension features vectors (Tab. 2.3).

2.4.3. ResNet

Recently, deeper networks with shorter features vectors have improved their predecessors' performance. As a representative of these approaches, we took a modified version of the ResNet-34 [52]. The modification consisted of removing a few layers and the number of filters per layer by half. This model was trained on the same dataset of the VGG-Face [99], and the face scrub dataset [91], apart from additional images taken from the Internet, amounting to a total of about 3 million images of 7 485 different identities. Again, we opted to use the pre-trained models provided by Dlib library [61] reported to achieve a 99.38% accuracy in the LFW. As for the

Table 2.4: Configuration of *Dataset A* and *Dataset B* in the training phase and in the two different tests (1 and 2).

Phase	Positive	Negative
Train	A	B
Test 1	A	B
Test 2	A	A

case of the VGG-Face, we discarded the last fully connected layer to obtain a 128 dimension feature vector (Tab. 2.3).

2.5. Effects of Cross-Dataset Training

With the scenario described in Section 2.2.1 in mind, this experiment aims at determining the potential performance damage when distinguishing samples of the same dataset when training uses outer negative samples.

In other words, could the trained classifier find inter-dataset differences more relevant than intra-dataset ones? To address this question, two datasets are defined as follows:

- *Dataset A* is a small dataset gathered in the target domain. It consists of a set of identities' faces with at least 10 images per identity, taken from one of the datasets presented in Section 2.3. Considering that each dataset has a different number of identities, it would be desirable to maximize uniformity in this sense. At the same time, it would also be desirable to have several identities as relevant as possible to perform statistics. Thus, although the minimum of identities is 15 in FS, we have set a maximum of 50, the available identities of O2FN (the second-lowest value). With this maximum, when we use FS as *Dataset A* the results are averaged just over 15 identities. Besides, images of each identity are split by half into train and test subsets.
- *Dataset B* is a large dataset used as a negative sample source. We have generated it using one dataset of Section 2.3, different from the one used to

create *Dataset A*. This way, each *Dataset A* identity has an associated *Dataset B* generated with the rest of the identities of the dataset. After that, this set is split into train and test subsets without shared identities and containing the same number of identities.

For each identity in *Dataset A*, we have trained a Linear-SVM model using the training subset of the specific identity in *Dataset A*, as positive samples, and the training subset from *Dataset B* as negative samples.

We have tested this model in two different ways to compare the performance:

- TEST 1. Testing against other identities in *Dataset B* (same configuration as the training phase).
- TEST 2. Testing against other identities in *Dataset A*.

For verification, SVM models depend on a threshold. Depending on this threshold's choice, the system's efficiency system varies both the False Acceptance Rate (FAR) and the True Acceptance Rate (TAR). Both measures are complementary. So, it is common to combine them into a single one. In this work, we opted to measure the True Acceptance Rate at 0.001 False Acceptance Rate (TAR @ 0.001 FAR), a standard practice in biometrics.

The experiment is performed independently for each one of the features described in Section 2.4.

2.5.1. Experimental Results

We performed the experimentation with different combinations of *Dataset A* and *B*, and features. Results are shown in Tabs. 2.5-2.7, where *Datasets A* are represented in rows and *Datasets B* in columns. The small-size number present the TAR @ 0.001 FAR performance for TEST 1 and TEST 2. The normal-size number is the drop in performance between the two different testings. Finally, in the last column and the last row, the average drops row-wise and column-wise, respectively.

The first thing we observe is that there is a general drop in performance between TEST 1 and TEST 2. The system has a higher rate of false negatives acceptance at

Table 2.5: LBP. Drop in performance (TAR @ FAR 0.001) of Test 1 respect to Test 2 for each combination of *Dataset A* and *B*.

		Dataset B												
		MobBIO		**FS**		**O2FN**		**FERET**		**LFW**		**IJB-A**		
		Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Av. drop
Dataset A	**MobBIO**			1.0000	0.9600	1.0000	0.9625	1.0000	0.9325	1.0000	0.9750	1.0000	0.9900	-0.0360
				-0.0400		-0.0375		-0.0675		-0.0250		-0.0100		
	FS	0.9583	0.1697			0.9266	0.3242	0.9349	0.1371	0.9905	0.1163	0.9210	0.2814	-0.7405
		-0.7886				-0.6023		-0.7978		-0.8742		-0.6397		
	O2FN	1.0000	0.6891	0.9821	0.6812			0.9857	0.5748	1.0000	0.2774	0.9967	0.7233	-0.4037
		-0.3109		-0.3009				-0.4109		-0.7226		-0.2734		
	FERET	0.9800	0.0495	0.9274	0.1491	0.9769	0.1555			1.0000	0.0276	0.8964	0.2972	-0.8204
		-0.9305		-0.7783		-0.8215				-0.9724		-0.5992		
	LFW	0.9875	0.2162	0.9999	0.1584	1.0000	0.0603	1.0000	0.0496			0.8996	0.6820	-0.7441
		-0.7714		-0.8415		-0.9397		-0.9504				-0.2176		
	IJB-A	0.7959	0.2589	0.9026	0.2796	0.9294	0.1980	0.9256	0.1815	0.8362	0.3751			-0.6193
		-0.5370		-0.6230		-0.7315		-0.7441		-0.4611				
	Av. Drop	-0.5167		-0.3480		-0.6111		-0.6266		-0.6677		-0.3713		

Table 2.6: VGG-Face. Drop in performance (TAR @ FAR 0.001) of Test 1 respect to Test 2 for each combination of *Dataset A* and *B*.

		Dataset B												
		MobBIO		**FS**		**O2FN**		**FERET**		**LFW**		**IJB-A**		Av. drop
		Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	
Dataset A	**MobBIO**			0.9975	0.8582	1.0000	0.7657	1.0000	0.9775	1.0000	1.0000	1.0000	0.9975	-0.0797
				-0.1393		-0.2343		-0.0225		0.0000		-0.0025		
	FS	0.9368	0.6179			0.9368	0.3220	0.9679	0.5666	0.9741	0.6950	0.9406	0.7104	-0.3771
		-0.3598				-0.6148		-0.4014		-0.2791		-0.2302		
	O2FN	1.0000	0.6626	0.9820	0.2312			0.9861	0.8776	0.9807	0.9142	0.9870	0.8634	-0.2773
		-0.3374		-0.7508				-0.1085		-0.0664		-0.1236		
	FERET	1.0000	0.6102	0.9810	0.3653	0.9754	0.3457			0.9852	0.7952	0.9387	0.8407	-0.3847
		-0.3898		-0.6157		-0.6297				-0.1900		-0.0981		
	LFW	0.9973	0.7647	0.9898	0.3608	0.9942	0.3467	0.9685	0.7933			0.9008	0.8496	-0.3471
		-0.2326		-0.6290		-0.6475		-0.1752				-0.0512		
	IJB-A	0.9821	0.4921	0.9051	0.2200	0.9848	0.2248	0.8964	0.5885	0.8550	0.7258			-0.4744
		-0.4899		-0.6851		-0.7600		-0.3079		-0.1292				
	Av. Drop	-0.3619		-0.5640		-0.5773		-0.2031		-0.1329		-0.1011		

Table 2.7: ResNet. Drop in performance (TAR @ FAR 0.001) of Test 1 respect to Test 2 for each combination of *Dataset A* and *B*.

		Dataset B												
		MobBIO		**FS**		**O2FN**		**FERET**		**LFW**		**IJB-A**		
		Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Av. drop
Dataset A	**MobBIO**			1.0000	0.9975	1.0000	0.9850	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	-0.0035
				-0.0025		-0.0150		0.0000		0.0000		0.0000		
	FS	0.9454	0.9496			0.9978	0.7014	0.9821	0.9401	0.9853	0.9524	0.9799	0.9388	-0.0817
		0.0042				-0.2964		-0.0420		-0.0329		-0.0412		
	O2FN	1.0000	0.7090	1.0000	0.7004			0.9959	0.9896	0.9987	0.9855	0.9978	0.9783	-0.1259
		-0.2910		-0.2996				-0.0063		-0.0132		-0.0195		
	FERET	0.9967	0.7047	0.9860	0.6423	0.9853	0.6804			0.9589	0.9268	0.9621	0.9135	-0.2042
		-0.2919		-0.3437		-0.3049				-0.0321		-0.0486		
	LFW	0.9978	0.8703	0.9978	0.8390	1.0000	0.7400	0.9954	0.9531			0.9784	0.9620	-0.1210
		-0.1275		-0.1587		-0.2600		-0.0423				-0.0165		
	IJB-A	0.9696	0.6639	0.9812	0.5945	0.9919	0.5153	0.9380	0.8232	0.8257	0.8815			-0.2456
		-0.3057		-0.3867		-0.4766		-0.1148		0.0558				
	Av. Drop	-0.2024		-0.2382		-0.2706		-0.0410		-0.0044		-0.0251		

the same false-positive rate when testing and training are done on the same dataset. This behaviour reveals that instead of just learning the identity information, the dataset's bias is interfering. This effect corroborates the important influence that dataset bias can have on performance.

Making now a feature-wise comparison, we note that the drop in performance is much more significant with the LBP features, reaching values up to +80% drop. Although a decline in performance is also present with deep features, it is much smaller, especially with the ResNet features.

The drop in performance also depends on the FAR point where the TAR is measured (Fig. 2.4). The influence of the dataset bias in performance correlates with the level of difficulty of the task. So, the higher the level of FAR requirements, the more notorious the effect is. Besides, it is also remarkable that the best results correspond to the case of using MobBIO as *Dataset A*, a dataset with highly controlled conditions acquisition. It seems that its limited amount of intra-class variations makes the verification task easy, even using LBP features.

Comparing the behaviour of the datasets as *Dataset B*, we can remark that datasets oriented for the unconstrained problem of face verification are the ones that behave the best in this role (LFW and IJB-A). On the contrary, the most constrained ones (MobBIO, FS and O2FN) lead to the lowest performances (highest drop). This behaviour is perfectly expected and agrees that for a negative training set, the more general, the better.

In addition, and more specifically, it is remarkable the entanglement between LFW and IJB-A datasets. Using one in the role of *Dataset A*, the other one is the best choice as *Dataset B*. This entanglement highlights the similar acquisition conditions of both datasets and their interchangeability when using them as *Dataset A* or *B*.

A similar but weaker entanglement appears between FERET and IJB-A, possibly due to their wide range of face's poses (we must remember that the pose effect is a bit limited using a quasi-frontal face detector). Faces from the first are taken in a highly controlled environment, whereas the second one is the complete opposite. Despite this fact, when using FERET as *Dataset A*, the lowest drop, in 2 out of 3 cases, is observed with IJB-A in the role of *Dataset B*. This behaviour strengthens the previous statement of the necessity of having impostor data taken in the same conditions as the genuine one.

To sum up, when dealing with specific contexts in which labelled data is scarce, the use of auxiliary datasets as negative samples source will affect real-world performance (where every image has the target domain conditions). In some way, with this setup, we are confusing classifiers. Thus, instead of targeting useful patterns to verify the target identity, we learn to distinguish between datasets.

2.6. Datasets' Feature Space

As aforementioned in Section 2.2.2, this second experiment aims to study the feature space to have some insights about the dataset's sample distribution. For this purpose, we have relied on a Nearest Neighbour search using two different metrics. The premise is that, given a feature vector of a face of a particular identity, the probability of the dataset to which its nearest neighbour belongs (eliminating other images of that same identity) should tend to be uniform for equivalent datasets (non-biased between them).

2.6.1. Building the subsets

To explore the distribution of dataset samples over the feature space, we split for each one of N_D datasets in two sets (Fig. 2.5): the *probe sets* and the *gallery set*. The idea is to search for the nearest neighbours of the elements of the *probe sets* among the elements of the *gallery set*. The creation of these sets consisted of:

- **Probe sets.** A total of N_p different random subsets of n_p faces sampled without replacement from each dataset. N_p will depend on the number of samples of the dataset. The greater the number of samples in a dataset, the higher number of its *probe sets*.

- **Gallery set.** The union of the random subsets of n_g elements sampled without replacement from each dataset, generates the *gallery set* set, with $N_D \cdot n_g$ unique elements.

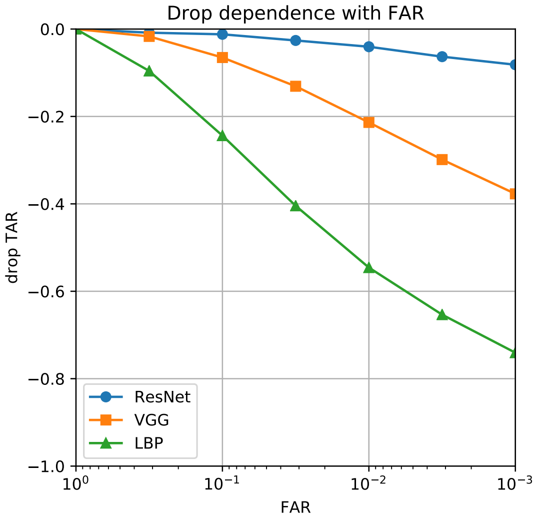


Figure 2.4: Average TAR drop between Test 1 and Test 2 respect to the FAR point in which we perform the measure, using FS just as *Dataset A*.

It is important to note that there will not be any common elements, neither identities, between any generated partitions.

2.6.2. Metrics in the Feature Space

Given a set of feature vectors $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, we denote by $\mathbf{x}_* \in X$ the nearest neighbour of $\mathbf{x}$ if:

$$\min\, d(\mathbf{x}, \mathbf{x_i}) = d(\mathbf{x}, \mathbf{x}_*) \quad i = 1, ..., n \tag{2.1}$$

The function $d(\mathbf{p}, \mathbf{q})$ represents a general metric. For our study, we have used the euclidean distance ($L2-$norm):

$$d_{L2}(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{m} (p_i - q_i)^2} \tag{2.2}$$

and the Manhattan distance ($L1-$norm):

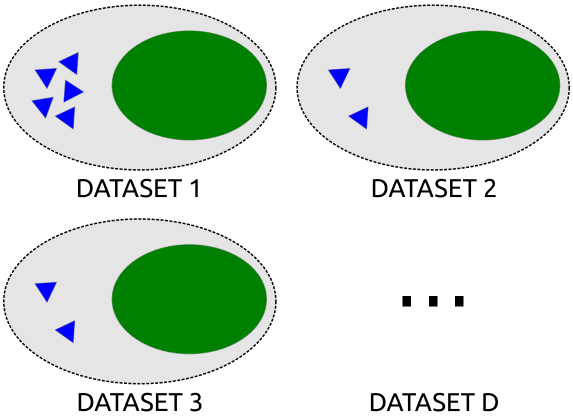


Figure 2.5: Scheme of the subsets generated for each dataset (Dataset 1, Dataset 2, ..., Dataset D_i, ..., Dataset N_D). Blue triangles represent *probe sets* and green ellipses represent the part of the dataset used to generate the *gallery set*. In this case, $N_p = 5$ *probe sets* were generated for the first dataset, and $N_p = 2$ *probe sets* for the rest.

$$d_{L1}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{m} |p_i - q_i| \tag{2.3}$$

Where p and q are two feature vectors in a m-dimensional descriptor space.

2.6.3. Nearest Neighbour Search

For each *probe set*, we have taken each of their elements and drawn without replacement (to avoid taking the same outlier) their nearest neighbours from the *gallery set*. Using the dataset membership of the nearest neighbours, we will generate a histogram for each probe set and average them over the N_p different probe sets of each dataset.

As aforementioned, the premise is that the probability (P) of a sample in the *probe set* finding an element of the *gallery set* gs_j belonging to the D_i dataset, as its nearest neighbour, should tend to be the same for all $i \in \{1, \ldots, N_D\}$:

$$P(x_* = gs_j \in D_i) = \frac{1}{N_D} \tag{2.4}$$

Any distribution different from the uniform may indicate that the datasets do not sample the same distribution, so at least one of them could have some data bias.

In our experiments, we have worked with a total of $N_D = 6$ datasets of different sizes. We have used $N_p = 2$ for O2FN and MobBIO, and $N_p = 5$ for the rest.

As we draw nearest neighbours without replacement, the prior probability change as we remove elements from the *gallery set* in each nearest neighbour search. In order to mitigate this effect, sizes of $n_p = 180$ and $n_g = 1200$ were fixed. This way, the number of elements of the *probe set* will keep low (180) with respect to the number of elements in the *gallery set* (7,200). This makes the effect of drawing nearest neighbours without replacement from the gallery set negligible.

Considering the prior probability (following Eq. 2.4 with $N_D = 6 \Rightarrow P \approx 16,7\%$) in the worst-case scenario where the nearest neighbours always belong to the same dataset, the prior probability of that dataset would decrease down to $\approx 15.2\%$.

2.6.4. Experimental Results

The distributions obtained for each case are shown in Tab. 2.8. Each row of the table contains the distribution (in %) among all the datasets (upper row) of the nearest neighbours to the elements in the *probe set* (leftmost column). Results are very similar for both L1 and L2 metrics.

Next, we analyse the obtained distributions from the two different points view. Finally, we will relate these results to the t-SNE data representation in Section. 2.6.4.

Looking from the Dataset Side

The most evident fact that we can extract from the data is a significant tendency to find the nearest neighbour in the same dataset (Tab. 2.8). Such effect reveals that the initial premise of a uniform distribution, Eq. (2.4), was false. In the case of LBP features, this nearest neighbour search is good enough (up to +90% accuracy) to guess the dataset membership (as done in the *Name that dataset!* challenge, [131]). It is also remarkable that the highest percentage of nearest neighbours is almost always achieved inside the same dataset, whatever the type of feature.

Table 2.8: Distribution of nearest neighbors over each dataset in % using different features.

		L2-norm						*L1-norm*					
LBP (Handcrafted)	Probe\Gallery	**MobBIO**	**FS**	**O2FN**	**FERET**	**LFW**	**IJB-A**	**MobBIO**	**FS**	**O2FN**	**FERET**	**LFW**	**IJB-A**
	MobBIO	**92.78**	0.00	5.00	0.83	0.28	1.11	**94.44**	0.00	2.22	2.50	0.00	0.83
	FS	1.78	**66.22**	19.78	10.44	0.11	1.67	1.44	**67.56**	17.00	12.22	0.00	1.78
	O2FN	1.11	1.11	**95.83**	1.67	0.00	0.28	0.00	2.22	**93.89**	3.89	0.00	0.00
	FERET	0.00	0.67	1.22	**97.33**	0.00	0.78	0.00	0.22	0.44	**99.00**	0.00	0.33
	LFW	0.78	0.44	0.33	0.00	**86.33**	12.11	1.44	0.56	0.44	0.22	**84.89**	12.44
	IJB-A	4.44	1.00	2.56	7.22	41.56	**43.22**	4.44	2.56	2.67	13.33	35.89	**41.11**

		L2-norm						*L1-norm*					
VGG-Face (Learned)	Probe\Gallery	**MobBIO**	**FS**	**O2FN**	**FERET**	**LFW**	**IJB-A**	**MobBIO**	**FS**	**O2FN**	**FERET**	**LFW**	**IJB-A**
	MobBIO	**55.83**	10.83	2.50	11.94	8.89	10.00	**60.28**	8.61	2.22	9.17	11.39	8.33
	FS	10.78	**45.11**	13.89	16.78	4.00	9.44	10.11	**46.67**	14.78	15.44	3.89	9.11
	O2FN	0.56	0.00	**82.50**	13.89	0.56	2.50	0.56	0.28	**82.78**	14.17	1.39	0.83
	FERET	1.11	1.78	15.56	**62.78**	6.67	12.11	1.33	1.78	14.56	**63.11**	7.56	11.67
	LFW	3.11	3.78	5.33	22.89	**34.56**	30.33	3.56	3.22	5.22	24.89	**33.22**	29.89
	IJB-A	1.22	4.44	4.44	20.33	15.78	**53.78**	1.56	4.78	4.00	20.89	20.33	**48.44**

		L2-norm						*L1-norm*					
ResNet (Learned)	Probe\Gallery	**MobBIO**	**FS**	**O2FN**	**FERET**	**LFW**	**IJB-A**	**MobBIO**	**FS**	**O2FN**	**FERET**	**LFW**	**IJB-A**
	MobBIO	**39.17**	38.06	1.11	11.67	4.72	5.28	**36.11**	35.83	1.11	12.50	5.56	8.89
	FS	21.89	**50.22**	1.56	11.89	4.44	10.00	21.56	**50.78**	1.44	12.22	4.22	9.78
	O2FN	1.67	0.00	**88.89**	7.78	1.11	0.56	1.11	0.00	**88.33**	8.89	1.11	0.56
	FERET	3.78	8.67	16.00	**42.22**	8.78	20.56	4.67	8.56	15.78	**41.56**	11.33	18.11
	LFW	4.56	5.67	3.33	19.11	31.11	**36.22**	4.56	6.78	3.00	19.33	31.11	**35.22**
	IJB-A	5.67	5.11	1.67	13.00	22.67	**51.89**	5.22	6.22	1.78	12.67	23.33	**50.78**

Before going any further with the discussion, we can divide datasets into two groups: the ones designed for the unconstrained face recognition problem (LFW and IJB-A) and the ones designed for more specific applications, namely MobBIO, FS and O2FN datasets for mobile applications, and FERET dataset for controlled environments.

There is a certain entanglement between the two unconstrained datasets for every kind of feature and metric. *Probe sets* taken from LFW and IJB-A datasets seem to have the highest proportion of nearest neighbours within one of these two datasets (+60%). This fact suggests that both datasets were drawn from similar distributions. Despite IJB-A being a more advanced dataset, its main contribution is to eliminate the frontal face constraint. Our face detector (see Section 2.3.1) continues to have a tendency of detecting frontal faces. So, the differences in the distribution between LFW and IJB-A seem to dilute. Such entanglement is present for the rest of the datasets.

On the other hand, O2FN is the dataset with the highest mean rate of samples' nearest neighbours in the same dataset; probably, its specific ethnicity is crucial to that outcome. Meanwhile, MobBIO and FERET experienced the most significant changes in the distribution of their nearest neighbours, according to features. For its part, FS dataset is one of the most stable in this respect.

Looking from the Feature Side

We can see that the same dataset pairing behaviour is quite strong in the LBP features. We see a +80% pairing in for 4 of the 6 datasets, meaning that the feature vector is not retaining the target information.

This pairing behaviour seems weaker with deep features. As we can expect, the training process performed to generate the CNN helps the system discard more information not related to the identity.

The main difference in behaviour between the two deep descriptors is that ResNet features break the rule of always having the highest frequency for self-pairing. This behaviour could indicate a certain correlation of the effect we are describing with the performance of the CNN.

Tab. 2.8 shows additional cues about the source of the bias. The first cue is provided by the results obtained for O2FN. As aforementioned, the elements of this dataset are unique to get a +80% paring across the three features. Thus, it does not suffer a solid drop like the other datasets despite being reduced by using learned features. The second cue is given by the results obtained for FERET and MobBIO. In the case of LBP features, these datasets suffer a comparable paring with respect to O2FN. Nevertheless, the drop caused by learned features reduced the paring to a 40-60%, much lower than the case of O2FN.

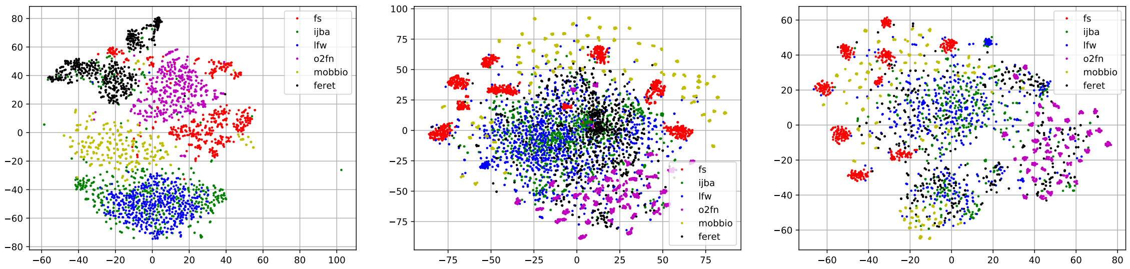
The main characteristic differentiating the O2FN dataset from the others is the prevalent Asian ethnicity of their identities, a bias related to μ. On the other hand, the common characteristic of FERET and MobBIO is the similar controlled-environment condition in which both datasets were generated, a bias related to ϵ.

Consequently, we can state that deep features help deal with bias in data related to ϵ, better than LBP features. But, on the other side, in terms of μ related bias, the effect is more similar between deep and LBP features because this kind of information is retained in both types of feature vectors. We have to recall Section 2.4 to find a theoretical explanation for this fact. The training process creates deep features to retain the identity information (μ) and discard the non-identity one (ϵ). This is something much more challenging to achieve with hand-crafted features. So, this behaviour is an illustrative example of how different feature extractors can hide a bias present in the data.

t-SNE representation of the gallery sets

The high dimension of feature vectors makes the task of directly visualizing data impossible. Therefore, it is necessary to imagine data in a reduced space. One of the most sophisticated options is the t-SNE representation [84], which tries to preserve the local structure of the high-dimensional data as well as some of the more global structure.

We can represent our *gallery sets* using this representation to look for any cue of their distribution (Fig. 2.6). The first thing that can be observed based on the representation is that each dataset distributes differently over the feature space. This fact is especially evident for LBP features since its clusters are the most separable



(a) LBP.

(b) VGG.

(c) ResNet.

Figure 2.6: A t-SNE representation of the 6 *gallery sets* taken from each dataset of study for three feature descriptors. (Red) FS; (Green) IJB-A; (Blue) LFW; (Purple) O2FN; (Olive) MobBIO; (Black) FERET.

and compact.

Finally, it can also be observed how images from the same identity cluster together when using deep features. For the FS dataset where the gallery set has a limited number of identities, we can even easily individually count each of them.

2.7. Conclusions

In this chapter, we have performed a study over the differences between datasets oriented for face verification. The analysis consisted of studying the performance impact and the distribution of the dataset elements. The problem of dataset bias is especially relevant in the development of face verification methods for specific contexts. This study shows the limitations of using public general context datasets as an auxiliary negative sample source.

Dataset bias effects are present across different datasets and feature descriptors. Although the newer deep feature representations help to palliate these harms, dataset bias has a non-negligible impact on verification performance.

A straightforward solution to dataset bias in specific contexts is to build a complete dataset for each application. However, this approach presents evident problems of scalability and accessibility. Another interesting alternative to tackle these issues can be through adaptation. Target domain samples become abundant during real-world operation. An adaptive system could use this data to improve its performance. This solution comes with new challenges, which we will approach in the following chapters.

Chapter 3

An Adaptive Face Verification in Video-surveillance

This chapter is heavily based on the contents of these articles:

E. Lopez-Lopez, C. V. Regueiro, X. M. Pardo, A. Franco, and A. Lumini. Incremental learning techniques within a self-updating approach for face verification in video-surveillance. In *Pattern Recognition and Image Analysis*, pages 25–37. Springer International Publishing, 2019

E. Lopez-Lopez, C. V. Regueiro, X. M. Pardo, A. Franco, and A. Lumini. Towards a self-sufficient face verification system. *Expert Systems with Applications*, 174:114734, 2021

3.1. Introduction

One of the main motivations behind the previous chapter is to study the main handicaps when transitioning face verification systems from a general context to a more specific one. In this regard, video surveillance is a paradigmatic example of these specific contexts. Indeed, faces in this context present many specific built-in characteristics (e.g. camera parameters, poses, illumination, target distance, etc.). Besides, these characteristics can be dependent on the deployment place and vary with time. These aspects make the gathering extensive amount of labelled context-specific data particularly difficult [51] and require for a more scalable a efficient alternative [87, 128, 106].

In more practical means, the standard face verification system is divided [104]: *enrolment* (where samples of the target identity, *genuine*, are registred) and *test/verification* (where the system checks if the identity of an input sample is *genuine* or not, *impostor*). The quality of the acquired samples during each of these phases are affected by these context-specific conditions, affecting the actual performance [44, 43]. For instance, some video-surveillance scenarios allow to perform the *enrolment* phase separately (collaboration is often required) where high-quality photographs or videos are acquired [56, 3, 139, 27, 17]. In these conditions, state-of-the-art systems seem to perform astonishingly well ($\approx$ 90-99% Rank-1 Identification Rates [149, 4]), something that makes this specific problem almost solved. Other cases do not allow this kind of *enrolment* (e.g. criminal watch-list, lost children, disoriented older people, etc.). Thus, the only option is to use data from the video stream as enrolment source [56, 68, 36] lowering system's performance [104, 83, 130, 73, 9]. To alleviate these negative effects, new classical learning strategies are being incorporated into the realm of Deep Learning [58]. Thus, topics as transfer learning [138, 23, 124, 22, 101, 138] reinforcement learning [112, 88] or incremental learning [51, 116, 128, 106], with different supervision degree, are gaining momentum.

Incremental learning consists of performing learning gradually as new data becomes available, without resorting to full retraining of the models. Thus, it is a scalable and efficient approach [87, 128, 106] to tackle very-specific, data-scarce and dynamic contexts where target domain data becomes gradually abundant during the *test/ verification* stage [36, 30, 77]. Recently, the attention of incremental and online learning has been continuously increasing [51, 116]. Despite being possible to perform under different supervision levels, the real value of adaptation arises considering it an unsupervised process [104, 66]. In this direction, the literature proposes semi-supervised incremental learning approaches [68, 36, 135]. However, despite their reduced label requirements, they often require a human operator in the loop to assist with the most challenging samples (by given additional labels).

In this direction [104, 94], self-training [146] is an interesting strategy to reduce labels requirements drastically. This approach follows an incremental learning perspective where the classifier drives the updates using its predictions as pseudo-labels.

Besides, self-training has also been recently used for domain adaptation purposes [153, 60], person Re-ID [150] or object detection [113].

This chapter aims to tackle the problem of progressively computing an efficient classifier for a video-to-video face verification (V2V-FV) setting where labels' availability is minimal. For this purpose, we propose the Dynamic Ensemble of SVM, a method that creates and automatically improves/updates an ensemble of very specific SVM classifiers. This kind of ensembles has proven to achieve remarkable results [86] on static supervised conditions. Here, we provide a novel decision mechanism to incrementally generate the ensemble in a semi-supervised way (i.e. starting from a few labelled data and then autonomously updating) and using only online target domain data. In this regard, the main contributions of the work are:

- The use of the self-updating approach in combination with the current most powerful feature representations as face re-identification system in the context of V2V-FV.
- An extensive comparison between different incremental learning strategies using the self-updating framework.
- The proposition of an ensemble-based adaptive biometric system called Dynamic Ensemble of SVM (De-SVM).

3.2. Related work

Face verification in video-surveillance can be performed under different settings [56]. First, in Still-to-Video face verification (S2V-FV), systems are queried to find an identity over video footage based on a (usually good quality) still image [39, 14]. Second, in a Video-to-Still (V2S-FV), the role of stills and videos is inverted [149, 4, 139]. And finally, in Video-to-Video (V2V-FV), only the verification is performed using just video sequences [68, 36]. As aforementioned in the previous section results achieved on recent databases like COX [56] with high-quality stills (either V2S or S2V conditions) convert the problem into an almost solved one [4, 149]. However, V2V continues to be an open challenge, especially when the labels are scarce.

Incremental Learning. The main goal of incremental (a.k.a template updating in the biometric scene) learning is to learn from data as real-world dynamic sources provide them, usually at a low pace, including noisy samples and, in general, exhibiting non-stationarity. As data distributions change with time, computational systems have to deal with the *stability-plasticity dilemma*, trying to avoid that new knowledge erases old one (*catastrophic forgetting*), while detecting and updating to concept drifts [116].

From a verification perspective, incremental learning has focused on two complementary tasks with different challenges [39, 14]. On the one hand, modifying or adapting a complete model to deal with dynamic environments that can impair performance [30, 68]; and, on the other hand, gradually improving the quality of models created with a small amount of data [146]. In this regard, most works in incremental learning perform adaptation in batches. In other words, they need to accumulate a batch of data before executing the adaptation. Only a few approaches were really designed to tackle the more challenging problem of incremental learning from streaming data (as is the case of video surveillance) [134, 1]. One of its critical difficulties is the infeasibility of complete manual labelling of streaming data in real-world applications. A more realistic approach should only assume that a few instances in data streams are labelled [133].

Self-training or self-updating is the approach used to update identity models in an unsupervised way. Firstly proposed in the scope of natural language processing [146], this approach assumes the classifier itself can do the genuine/impostor labelling, avoiding any supervision [36, 29, 94]. Outside the biometric scope, self-training ideas help to minimise human annotation effort in network traffic classification [32] or to incorporate unlabelled data from auxiliary information sources, like the internet, to improve object detectors [107]. The complicated part of the approach consists of finding balance in the update decision. For example, a too high-confidence threshold could avoid accepting impostor identities, but at the expense of only accepting too redundant information that does not improve performance. On the contrary, a too low-confidence threshold could help to increase diversity in the accepted samples, but at the cost of increasing the risk of accepting more impostors into the model.

Temporal Coherence in videos leads to the assumption that consecutive frames will almost always contain very similar information [6]. The literature often highlights the exploitation of temporal coherence as one of the keys for unsupervised learning [36, 102, 140]. From this assumption, Siamese network architectures have been proposed to train DNNs in an unsupervised way [140, 89, 110]. In video surveillance scenarios, the temporal coherence assumption is perfectly applicable once an identity is verified in a video frame. In the actual implementations, systems often use visual tracking methods to keep the target identified. Thus, visual tracking provides an auxiliary source of supervision and facilitates access to more challenging samples into the model. However, this does not apply to the transition between different sequences, i.e. when a cut takes place.

3.3. Adaptive Biometrics for Face Verification in Video-Surveillance

The objective of this section is two-folded. First, it describes the self-updating general framework in which the classification methods will be wrapped [152]. Second, it describes De-SVM, the novel classification technique specifically designed to work within this framework.

3.3.1. Self-updating Pipeline

From a general perspective, the self-updating approach states the following hypothesis: the use of pseudo-labels given by the model (M_t) at the moment (t) to update helps to improve performance. The considered scenario assumes that initially ($t = 0$) a few video frames of the *genuine* identity (short sequence given by a visual tracker) are selected to create the *template*. The quality of this template can be an important constraint to the performance of the system, as [81] and Sec. 3.5.5 show. Besides, the availability of a group of samples to use as a negative set taken in the operational domain (as it was studied in Chapter 2) is also assumed.

As it is outlined in Algorithm 1 and Figure 3.1, over time ($t = 1, 2, ..., T$), the system is queried with new video sequences (S_t) to verify the identity of the individuals (both *genuine* and *impostor*) appearing in them (Cohort Model, CM

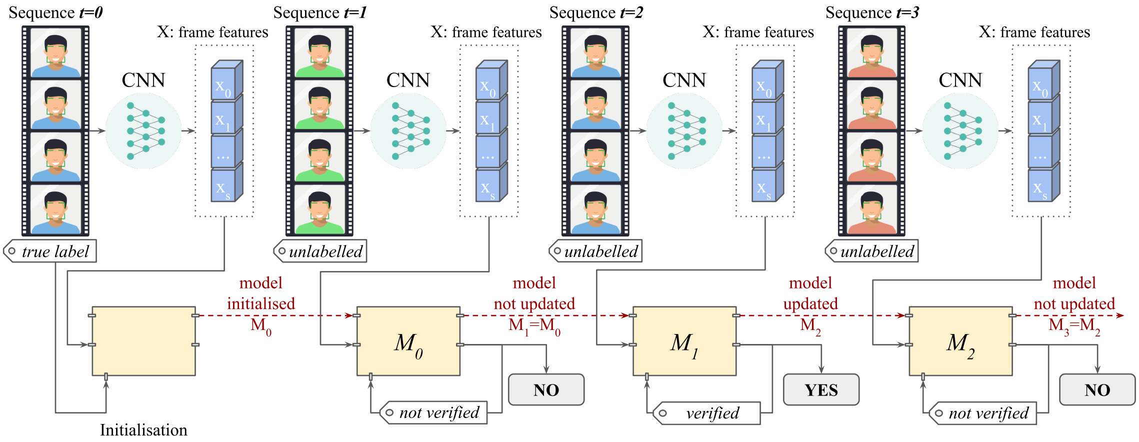


Figure 3.1: The designed pipeline for the implementation of the self-updating strategy with its ideal behaviour. Whenever the target identity (the blue jersey one) appears, the model is updated. Otherwise, the model is maintained.

[68]). Following an *query acceptance* adaptation criterion [104], if M_t accepts the query sequence, the incoming sequence to update M_t into M_{t+1}. In the opposite case, M_{t+1} will remain exactly the same as M_t.

Decision Functions

Given the previously stated pipeline, several decision functions need to be defined during the implementation. These rules control when an identity is verified or not, and, if so, how updates are performed.

- *Frame Decision Function (FDF)* assigns a score to each frame of the query video sequence. It can be the outcome provided by a single classifier (e.g. SVM score, distance in a nearest-neighbour algorithm, or a softmax in a DNN), or the fused output in the case of an ensemble of classifiers.

- *Sequence Decision Function (SDF)* assigns a unique score to the query video sequence based on the FDR individual scores in each frame. Identities will be verified by fixing a certain confidence level to this score, the so-called *operational threshold*. This confidence level will be the one in charge of finding balance in the *stability-plasticity* dilemma.

Algorithm 1 The implementation of the self-updating strategy.

```
Input Query Sequences = $\{S_0, S_1, ..., S_T\}$, negativeSet, type_of_model, $TH$ (operational threshold).
Output Self-updated model, $M_T$
$M_0$ = createModel ( $S_0$, negativeSet, type_of_model )
for t = 1,2,...,T do score = evaluateSequence ( $S_t$, $M_{t-1}$, $SDF$, $FDF$)
    if score > $TH$ then
        $S_t$ assumed to be a genuine sequence
        $M_t$ = updateModel ( $S_t$, $M_{t-1}$, $UF$)
    else
        $S_t$ assumed to be an impostor sequence
        $M_t = M_{t-1}$
    end if
end for
```

- *Update Function (UF)* defines how new information is used to enhance the current model incrementally.

Algorithm 1 and Fig. 3.1 illustrates the role of each decision function in the self-update pipeline. The actual implementation choices for each of the explored models (in the following Sec. 3.3.3) are shown in Tab. 3.1.

3.3.2. The Proposed Dynamic Ensemble of SVM (De-SVM)

The classification power of ensembles of very specific classifiers was first proven by [86]. There, an ensemble of exemplar Support Vector Machines (SVM) classifiers, each of them (exemplar-SVM) trained with just one positive sample and a large number of negative samples, is used in the frame of object detection. The idea behind this strategy is to have an ensemble of very specific classifiers whose combined decision overcomes over-fitting. In the particular case of face verification in video surveillance, [3] uses a similar approach to recognise a target identity among distractors in the case of S2V face recognition. Each exemplar is built during enrolment from a single target sample and multiple distractors' samples to represent the diversity of the same identity appearance due to various perturbation factors. Ensembles of (exemplar) SVMs leverage the intuition according to which a pool of

Table 3.1: Summary of the decision rules with each method.

Method	**Decision rules**	**Method used**
De-SVM	*Frame Decision Function*	Ensemble fusion rule (Median)
	Sequence Decision Function	Median
	Update Function	Add a new classifier with the hardest samples
SVM	*Frame Decision Function*	Raw score
	Sequence Decision Function	Median
	Update Function	Retrain the classifier with the additional information
I-SVM	*Frame Decision Function*	Raw score
	Sequence Decision Function	Median
	Update Function	Partial fit using the query sequence
OS-ELM	*Frame Decision Function*	SoftMax
	Sequence Decision Function	Median
	Update Function	Partial fit using the query sequence

simple classifiers, one for each training sample, can outperform a single and complex one [5]. Besides, another advantage of ensemble-based methods is the extra point of flexibility. One could potentially control how each ensemble member performs, allowing classifiers substitutions or removals whenever needed to keep the ensemble size bounded. The following Chapter 4 particularly explores this possibility.

The proposed De-SVM use these previous ideas as a basis. Besides, it combines them with the self-updating method to create an identity-specific ensemble in an incremental and primarily unsupervised way. Instead of using Exemplar-SVM, the number of positive samples is generalised to n. Still, this number will continue to be relatively low ($n = 5$ in our experiments) to maintain the Exemplar-SVM philosophy. The generalisation need was shown in a previous work [81] where the quality of the initial template is proven to be crucial to deploy the self-updating strategy. The transformation to an incremental classifier consists of adding new classifiers to the ensemble whenever the *genuine* identity is verified. In our case, the number of possible updates keeps relatively bounded, keeping out of the scope a procedure to limit the number of ensemble classifiers. Nevertheless, for an actual application, one should restrict it following either substitution or removing strategies [66] (Chapter 4).

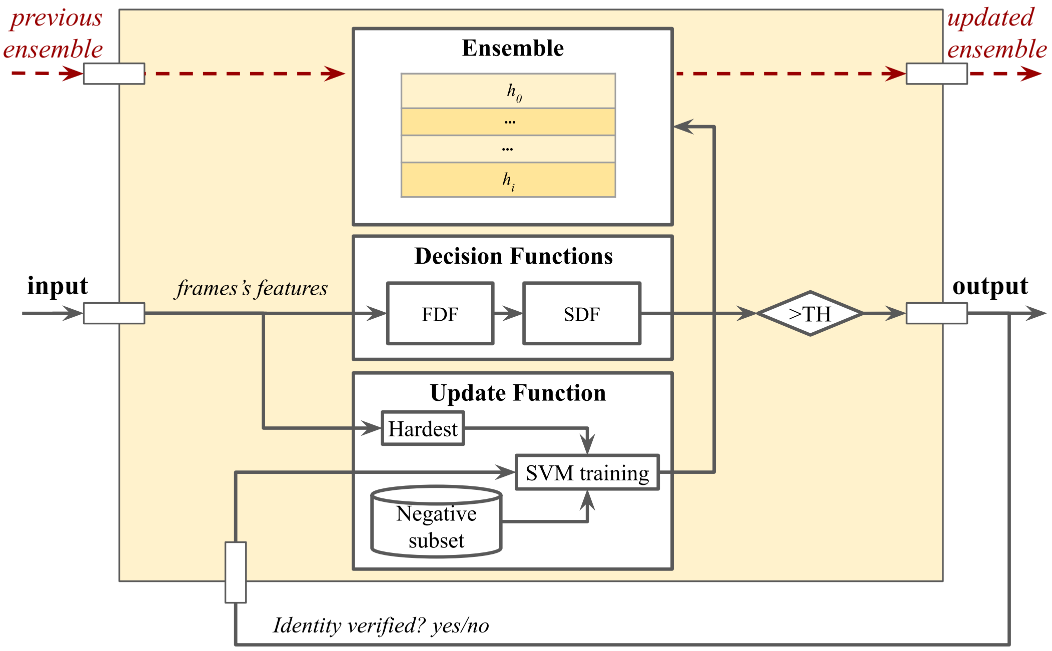


Figure 3.2: The pipeline of De-SVM within the self-updating general strategy. The model (M_t) is updated based on its decision over the query sequences (S_{t+1}) to generate a new model (M_{t+1}). Three different functions are involved in the decision of updating (FDF and SDF) and the way of updating (UF).

Figure 3.2 depicts the pipeline of De-SVM. Following the self-updating paradigm, the ensemble at the moment is the one that decides whether to update or not. First, the median is used as FDF to give a score to each frame (which in practice corresponds to a majority voting). Afterwards, the median of the sequence's frames FDF scores is computed again as SDF. Finally, if the identity is verified (based on the operational threshold), the ensemble adds a new classifier following the UF.

The $n = 5$ samples used to create the new member of the ensemble will be pooled from the query sequence's frames. To enhance diversity within the ensemble, the hardest frames (the ones with the *worst* score obtained using the current model, using the *FDR*) are selected as positive samples to train (against a large number of negative samples) the next classifier of the ensemble.

3.3.3. Other Explored Classification Methods

The self-update approach has a ‘wrapper algorithm’ [152] nature which in practice converts a supervised classification method into an unsupervised and incremental one. Therefore, one could use different classification methods within this same approach. Here, we explore three additional supervised classification methods (either incremental or batch-based):

Linear Soft-Margin Support Vector Machine (SVM)

The Linear Soft-Margin Support Vector Machine (SVM) is a batch-based binary classification method [20] widely used in many applications. Given a set of N labelled training feature vectors $\mathbf{x}_i$ of two classes, the method finds the optimal hyperplane which separates each of classes. Recently, this classification technique has lost a bit of its prominence in favour of CNN’s. Nevertheless, it continues to be used on top of CNN-based features. In the specific context of face recognition, there are numerous examples of SVM-based methods in the recent literature [22, 144, 28].

Incremental SVM (I-SVM)

This is an incremental implementation of the previous method. Here, training data is provided sequentially instead of the batch mode in which all examples are available at once. Thus, new training data is incorporated when it is available, without re-training from scratch. In [63], a simple and computationally efficient algorithm, based on the classical Stochastic Gradient Descent, was developed to update the hyper-plane parameters incrementally.

Online Sequential Extreme Learning Machine (OS-ELM)

The Online Sequential Extreme Learning Machine (OS-ELM) is an incremental implementation of the regular Extreme Learning Machine (ELM) problem. The ELM builds a Single Layer Feed-forward Network (SLFN) with $\tilde{N}$ hidden nodes to approximate a set of N labelled training feature vectors such that:

$$f_{\tilde{N}}(x_j) = \sum_{i=1}^{\tilde{N}} \beta_i G(\mathbf{a}_i, b_i, \mathbf{x}_j) = y_j, \qquad j = 1, .., N \tag{3.1}$$

where $\mathbf{a}_i$ and b_i are the parameters of the hidden nodes activation function G (additive or RBF); and β_i the weight that connects the i-th hidden node with the output. It is showed that (3.1) is satisfied for any randomly assigned values of the node parameters ($\mathbf{a}_i$ and b_i) by analytically computing the weight β_i, as long as $N \geq \tilde{N}$.

In the specific case of OS-ELM, the approach is specifically adapted to compute and update the weight values sequentially as more data is becoming available ('chunk-by-chunk' or one-by-one) [75]. In this case, a sigmoid function is used as an activation function, and the number of hidden nodes is empirically fixed at $\tilde{N} = 80$.

3.4. Methodology

3.4.1. Datasets

COX Face database

COX Face database [56] (COX) was specifically designed for the context of video surveillance. This dataset gathers video frames of 1 000 identities with 3 video sequences each captured by 3 different viewpoints (`cam1`, `cam2` and `cam3`). The subjects were asked to walk over an S-path, and their images were captured under variable lighting, pose, scale conditions, and a considerable amount of blur. Each camera recorded a part of the path without temporal overlapping between them. The number of sequences of each identity is quite limited (3 sequences per subject). To

Table 3.2: YouTube Faces distribution of the amount of videos per person after the face detection phase.

#videos	1	2	3	4	5	6
#people	588	472	305	167	51	8

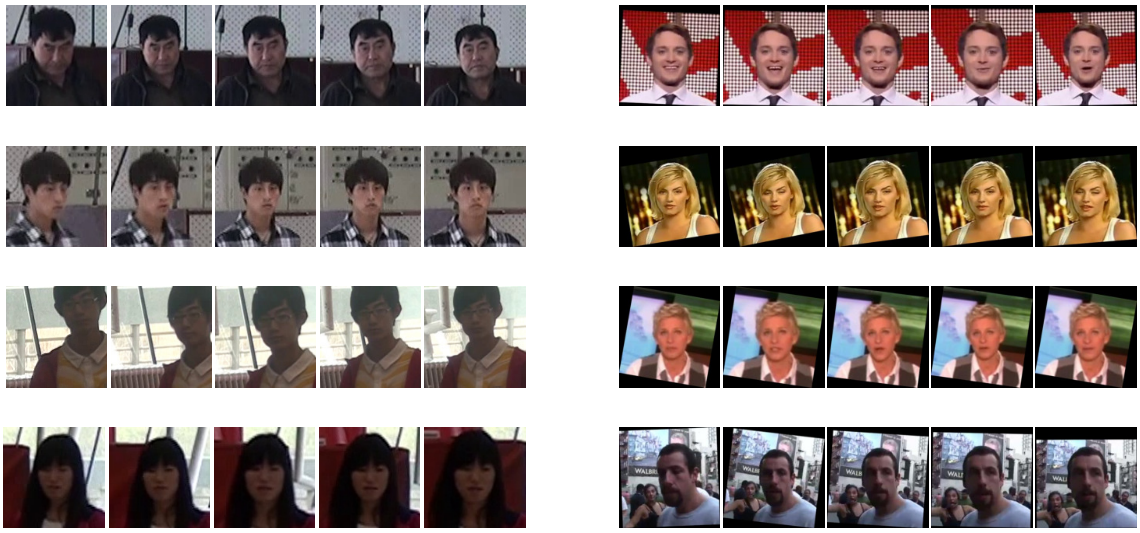


Figure 3.3: Samples of both datasets, COX (left) and YTF (right).

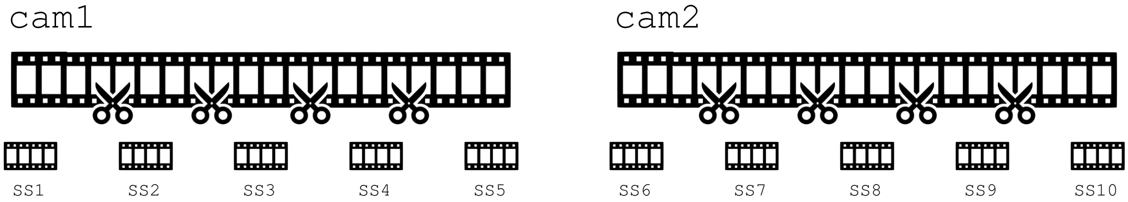


Figure 3.4: Example of division of `cam1` and `cam2` in the COX Face database to generate the query sub-sequences (SS stands for sub-sequence).

mitigate this limitation, we have split each video sequence into several sub-sequences without alteration of the temporal order (Fig. 3.4).

YouTube Faces

YouTube Faces Dataset [143] (YTF) contains a total of 3 425 videos downloaded from the YouTube platform of 1 595 different identities. Each identity appears in between 1 and 6 different videos captured under completely different conditions. Table 3.2 contains the distribution of video frames per identity after face detection. As with the previous dataset, to augment the number of video sequences to query the system, each video sequence has been split into several sub-sequences while maintaining temporal coherence.

3.4.2. Face detection and feature extractor

A face detection technique is applied over every frame to discard the background part of images and for alignment purposes. We selected the tool provided in the Dlib library [61] for this task. After that, we will extract a feature vector of each face using a pre-trained ResNet-34 network [52] with just 29 convolution layers (RN29). The classifier layers have been removed from the network, as provided by Dlib [61], giving a feature vector of 128 dimensions. The network has been trained using a combination of the SCRUB dataset [92] and the VGG-Face dataset [99]. This implementation achieves an accuracy of 99.38% in the LFW dataset (which is comparable to the face verification state-of-the-art) and has shown quite desirable properties in terms of robustness to non-identity related variations (Chapter 2 and [83]).

3.4.3. Testing Protocol

Using the protocol proposed by the COX database as an inspiration, each dataset was divided into three different subsets (See Tab. 3.3):

1. The **train subset** is composed of the face images used as a negative set and as a validation set in the learning process. This negative set will be a random subset of 1 000 samples from the whole train set in the actual implementation.
    - In the case of the COX Face database, this subset comprises the face images of 300 identities taken from each available camera.
    - In the case of YouTube Faces database, this subset contains data from the identities that have less than 4 video-sequence per identity, giving a total of 1 365 identities.
2. The **gallery subset** is composed by the video-frame sequences used to create the initial template as well as the ones used to query the system (from both *genuine* and *impostor* identities).
    - In the case of COX Face database, this set contains the 700 identities taken from `cam1` and `cam2`. Each video was divided into 5 sub-sequences

Table 3.3: How the datasets' identities are divided in order to generate each subset defined in Sec. 3.4.3.

		Genuine				Impostor			
		`still`	`cam1`	`cam2`	`cam3`	`still`	`cam1`	`cam2`	`cam3`
COX	Train	0	0	0	0	300	300	300	300
	Gallery	0	700	700	0	0	0	0	0
	Probe	0	0	0	700	0	0	0	700
		≥ 4 videos		< 4 videos		≥ 4 videos		< 4 videos	
YTF	Train	0		0		0		1365	
	Gallery	226		0		0		0	
	Probe	226		0		226		0	

to augment the number of possible queries, giving a total of 10 sub-sequences.

- In the case of YouTube Faces, this subset contains data from the identities that have equal or more than 4 video sequences per identity, giving a total of 226 identities. It includes all but one video of each identity, which will create the following probe subset. The videos will be divided into a total of 10 sub-sequences without mixing different videos.

3. The **probe subset** contains the video-frames sequences we draw to test the system in each step of the learning phase. The testing is performed after each query of the learning phase. This way, we can measure the evolution of the updating system.

    - In the case of COX Face database, this set contains video sequences captured by `cam3` belonging to the 700 identities in the *gallery subset*. In this case, each video sequence was divided into 10 sub-sequences to have more sequences to test.
    - In the case of YouTube Faces, this subset contains data from the identities that have equal or more than 4 video sequences per identity, giving a total of 226 identities. It includes the remaining video after the creation of the

previous gallery subset. As for the other dataset, each sequence will be divided into 10 sub-sequences too.

It is important to remark that there are no common identities between the *train subset* and the other two subsets. Identities belonging to the *train subset* conform the Universal Model (UM) [68, 33]. On the other hand, *gallery* and *probe subsets* contain different sequences of shared identities. In the experiments, each of these identities will have associated its own Cohort Model (CM) [68, 33]. In practical terms, each identity CM will be conformed by itself and its 10 'most similar' (using SVM as metric [83]) impostors or hard-negatives. Consequently, this kind of testing is quite more demanding than regular random impostor testing.

3.4.4. The *Operational Threshold*

In each verification query, a short sequence of video frames is processed according to the SDF (see Sec. 3.3.1). The SDF gives a score to the video sequence and decides based on a threshold, the so-called *operational threshold*. Its determination is crucial and especially tricky in an incremental learning context.

The strictness/gentleness on the *operational threshold* modulates the self-labelling process's confidence degree. Potentially, aspects as data quality, face characteristics, or the acquisition environment may affect its optimal determination (identity and time dependence). Nevertheless, we have opted to ignore these dependencies (both identity and time) when defining the determination procedure. It seems reasonable as a first step considering the enormous limitations in labelled data of the application's context. Similar assumptions are commonly made in other works [109]. Section 3.5.1 performs a further study of the implications of this assumption.

Thus, the *train subset* (which contains the identities of the UM) will be used as a validation set. Within this set, we will replicate the previous divisions (train, gallery and probe) to build and characterise (compute the ROC curve) sample models using what would be the initial template. As a convention, we have selected the threshold associated with 5% FAR point of the initial model as *operational threshold*. However, we test different operational thresholds (along with other templates sizes) on Sec. 3.5.5.

3.4.5. Metrics

The metrics used to evaluate are the measurement of TAR at a given 1% FAR (TAR@FAR1). In addition, we also provide the Transaction Level performance (TAR and FAR at the *operational threshold*).

Performance is assessed on the *probe subset.* After each query, the system is presented with 10 *genuine* sub-sequences and 1 *impostor* sub-sequence per *impostor* identity (10 in total). As it has been stated in Sec. 3.4.3, both querying and testing is performed using the CM of *genuine* identity. Finally, results are averaged over the total number of identities used as genuine in each verification process.

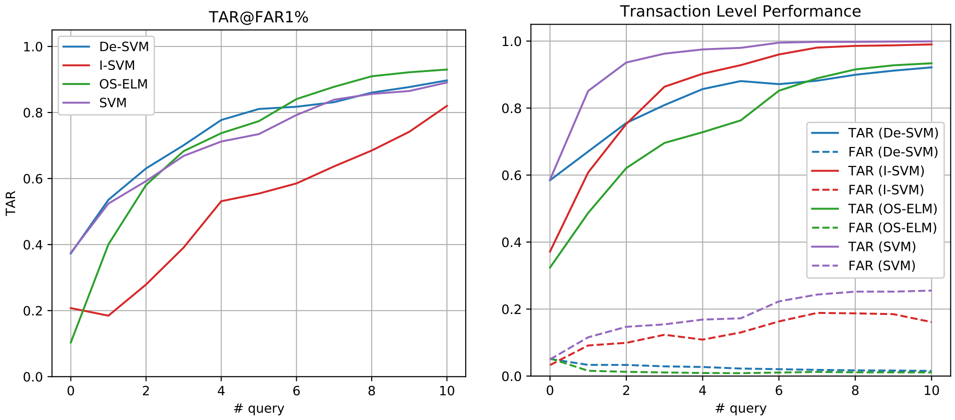
Finally, as mentioned in Sec. 3.4.3, the negative set used for training consists of a random subset of 1000 samples drawn from the *train subset.* This randomness adds uncertainty to the results and needs to be addressed. In order to deal whit it, experiments will be repeated 8 times and averaged.

3.5. Experiments and Results

This section presents the experimental part of the chapter. First, Sec. 3.5.1 explores the supervised adaptation performance of the four different classification techniques. From that, Sec. 3.5.2 explores their ability to build a robust model with a minimum amount of labelling in an environment where both genuine and impostor can potentially query the system. After that, Sec. 3.5.3 tests models' robustness to repeated impostor *attacks.* On Sec. 3.5.5, some additional insights of De-SVM are provided to understand its behaviour and parameters fully. The complete results achieved in the experiments are recalled in Tab. 3.4 to be analysed in a final discussion (Sec. 3.5.4).

3.5.1. Supervised Adaptation

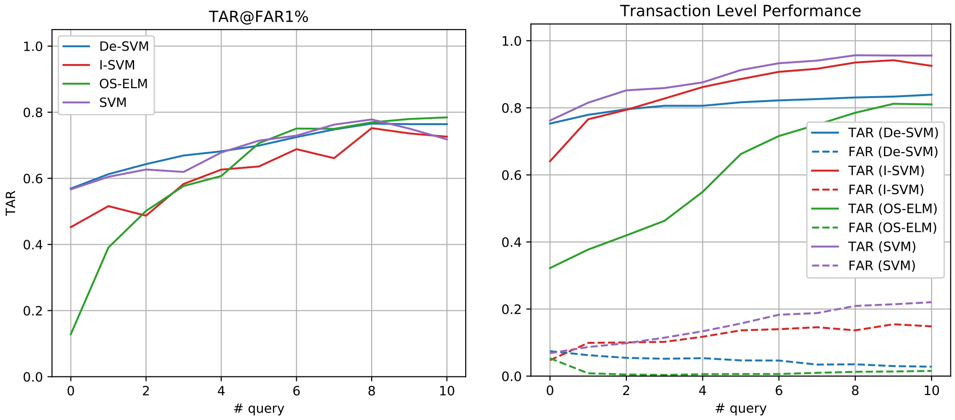
Despite aiming to perform incremental learning in low-labelled contexts, one of the first steps to study each method's performance is to observe its behaviour under supervised incremental learning conditions. The performance obtained in



(a) TAR@FAR1 performance.

(b) TAR and FAR at the operational threshold.

Figure 3.5: COX: Supervised updating performance comparison.



(a) TAR@FAR1 performance.

(b) TAR and FAR at the operational threshold.

Figure 3.6: YouTube Faces: Supervised updating performance comparison.

this experiment represents an upper-bounds since they use entirely supervised labels (perfect self-labelling). Then, the $n = 5$ available labelled frames are used as the initial template. This template is used to create the model M_0 and, after that, the system is queried (and the model consequently updated) with 10 different short video sequences (verification queries) from the genuine identity.

Performance is measured on the *probe set*, using the samples of the CM of each identity. Measurements were done after the creation of the initial model ($t = 0$) and after each query ($t = 1, 2, ..., 10$). Results showcased in this experiment are directly comparable with the following ones on Sec. 3.5.2, when the updates are done in the absence of labels.

Results (Figs. 3.5 and 3.6) show that every method can achieve a remarkable performance, especially when testing on the COX Face database. In this sense, I-SVM is the method that experiences the hardest time during the experiment. Above all, results show the ability of every method to perform supervised incremental learning. Differences in performance are more obvious at the *operational threshold*, as illustrated in Figs. 3.5b and 3.6b. De-SVM achieves the most modest performance in terms of TAR with respect to the other methods. Nevertheless, it is important to remark that higher performances are obtained at the cost of higher (and increasing) values of FAR. Both OS-ELM and De-SVM present a decreasing FAR, which ends up below 5%. These curves suggest a more desirable behaviour when updates will be performed in an unsupervised manner. A high FAR value means that the probability of accepting impostors during the training could also be high, and so the high risk of corrupting the model.

Finally, these last figures give an important clue about the relevance of the *operational threshold* and its crucial influence on the self-updating mechanism. This influence is comprehensively studied in the following section (Sec. 3.5.1).

Time independence of the operational threshold

The *operational threshold* is assumed to be neither time nor identity dependent (Sec. 3.4.4) and is fixated to initially have 5% FAR. Despite not being totally accurate, this assumption is forced by the label's scarceness of the operation's context. In fact, the previous Figs. 3.5b and 3.6b have already suggested that this assumption could be a bit inaccurate. Delving into this analysis, Fig. 3.7 shows how, using a constant *operational threshold*, FAR increases after each update for a supervised batch SVM (described in Section 3.3.3). This section aims to explore further the actual implications of the independence assumption (specifically regarding time).

The experiment conducted consists of representing the temporal evolution of the associated threshold of five different points of the ROC curve (Fig. 3.8) under supervised conditions (same experiment as in before) . According to the previous assumption, the ideal behaviour would correspond to steady curves. Moreover, such behaviour would mean that the FAR point initially chosen would be maintained

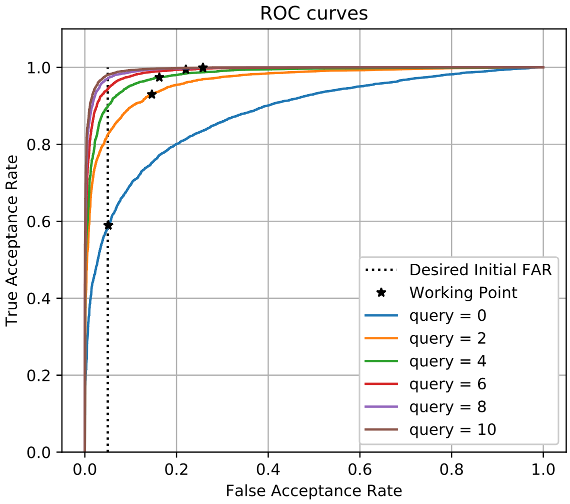


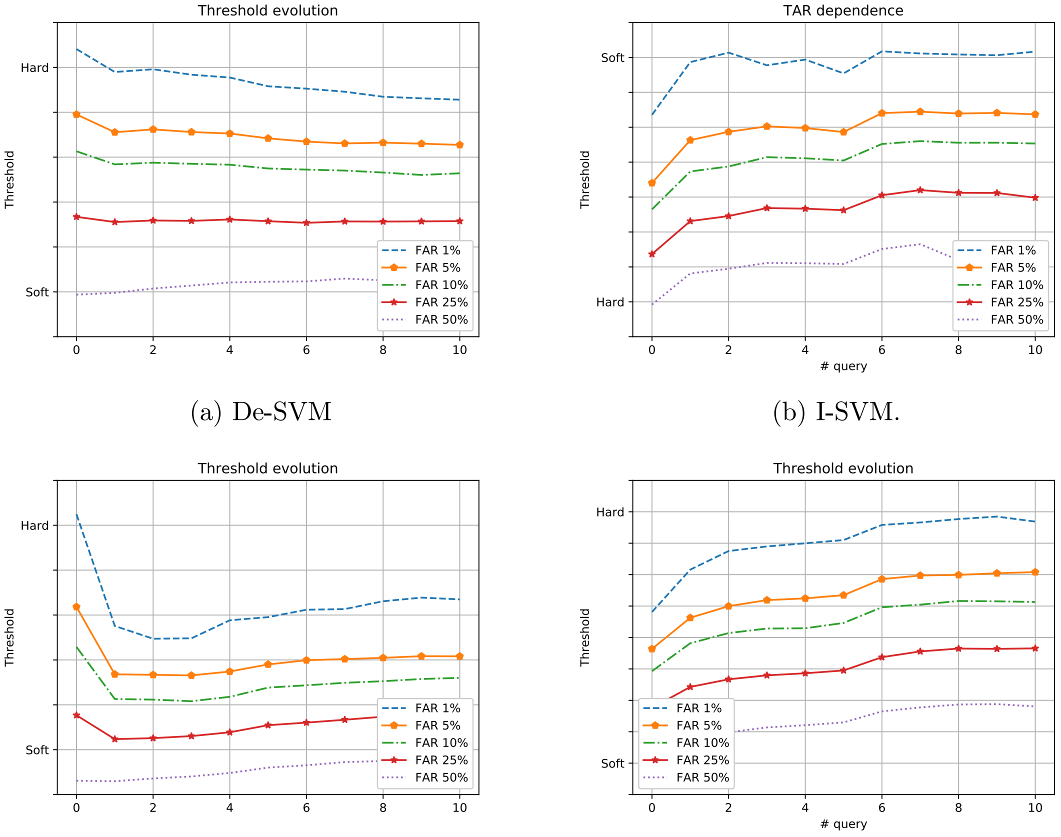
Figure 3.7: Evolution of the ROC curve and the ROC point associated to the *operational threshold* after each query for the supervised case using the SVM classification model.

during the model updates.

Results show different behaviours depending on the classification method used. Overall it is observed the explicit break of the threshold's time independence assumption. De-SVM is the method that presents the most steady behaviour among all the classification techniques during this experiment. For instance, for SVM, thresholds corresponding to 1% FAR (at the beginning) and 25% FAR (at the end) are the same. OS-ELM shows the opposite behaviour. The initial threshold associated with a FAR 5% decreases over time.

This behaviour can be explained in terms of the sample balance (positive/negative) during the classifiers' creation. In the early stages, this balance is compromised. The limited number of positive samples contrasts with a large number of negative ones. The balance is recovered when the model adds more genuine queries. Subsequently, the *operational threshold* selected (at 5% FAR) using the initial model (unbalanced problem) leads to a decision boundary shifted towards the positive sample(s). This boundary does not correspond to the same ROC point of the final model (more balanced problem).

In contrast, De-SVM uses an ensemble of 'unbalanced' SVM. However, each classifier is 'unbalanced' to the same degree, making the decision boundary much more stable during the queries. Besides, De-SVM presents a subtle decrease over



(a) De-SVM

(b) I-SVM.

(c) OS-ELM

(d) SVM.

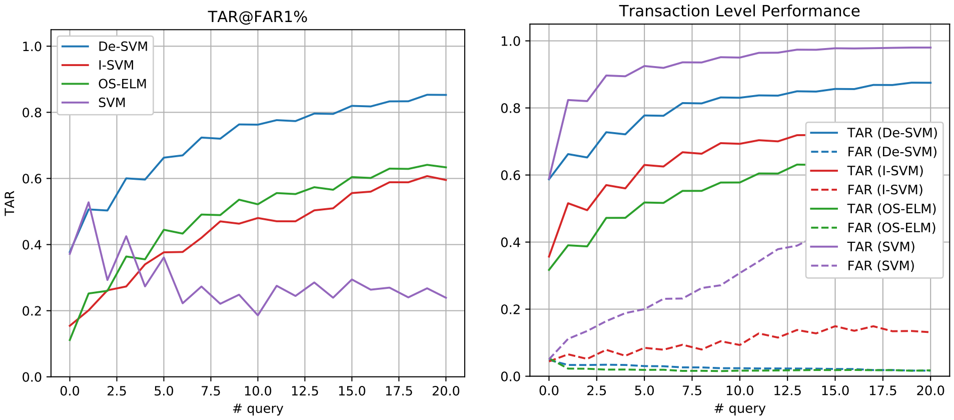
Figure 3.8: COX: Threshold evolution of a same FAR point of the ROC curve (Data for 1%, 5%, 10%, 25% and 50%).

time, indicating that the decision is becoming more strict. OS-ELM even show a more substantial decrease in the initial phases.

These experiments showed that, despite being inaccurate, the time and identity independence assumption is acceptable. Thus, one can start to foreshadow a better performance of both De-SVM and OS-ELM in the following experiments.

3.5.2. Unsupervised Adaptation

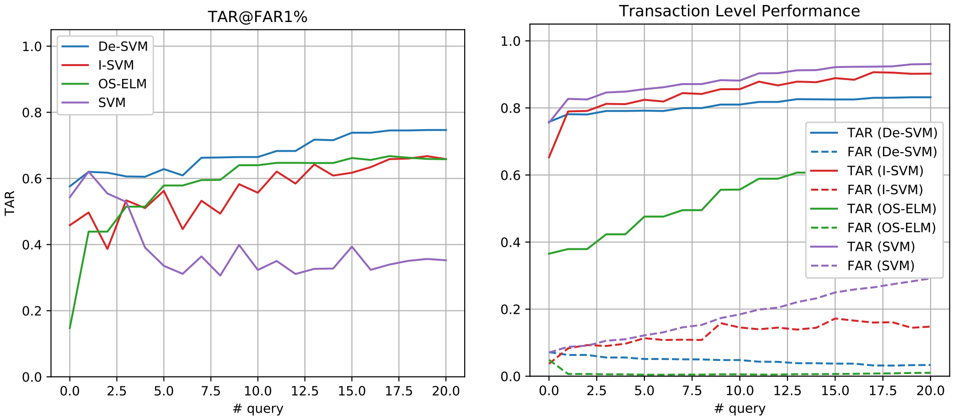
The procedure in this experiment follows the philosophy of the supervised case (see Sec. 3.5.1). However, instead of having access to labels, the SDF must dis-



(a) TAR@FAR1 performance.

(b) TAR and FAR at the operational threshold.

Figure 3.9: FACE COX: Self-supervised experiment using an operational threshold of 5% initial FAR.



(a) TAR@FAR1 performance.

(b) TAR and FAR at the operational threshold.

Figure 3.10: YouTube Faces: Self-supervised experiment using an operational threshold of 5% initial FAR.

tinguish between genuine and impostor queries. Consequently, the first step was to generate M_0 from the initial template (5 frames). After that, the model was queried with 10 genuine sequences, G^s, and 10 different impostor sequences, I_k^s (where s stands for the sub-sequence number and k for the identity of the impostor). All of them belonging to the *gallery subset* (identities of the CM). After each genuine query (odd query, $t = 1, 3, ..., 19$), an impostor query (even query, $t = 2, 4, 20$) was presented. The query order follows this pattern:

$$G^1 \rightarrow I_1^1 \rightarrow \;\; G^2 \rightarrow I_2^1 \rightarrow \;\; ... \;\; \rightarrow \;\; G^s \rightarrow I_k^1 \rightarrow \;\; ... \;\; \rightarrow \;\; G^{10} \rightarrow I_{10}^1$$

Performance measurements are done on the *probe set*, using samples of each identity's CM. As aforementioned, the CM consists of both the genuine identity and its 10 most similar impostors (see Sec. 3.4.3). Performance metrics (see Sec. 3.4.5) measurements are done after each query to the system.

Results (Figs. 3.9 and 3.10) for both COX and YTF, respectively, show the ability to improve the performance of the self-updating approach for every classification method apart from SVM. De-SVM is the one to achieve the best performance scoring at TAR@FAR1 of 86.03±0.48% on COX and 75.5±1.2% on YTF (Sec. 3.4.5 contains a detailed explanation about the uncertainty origin).

A comprehensive analysis of Figs. 3.9b and 3.10b allows to identify two different behaviours. On the one hand, De-SVM and OS-ELM can improve TAR while decreasing/maintaining FAR. Conversely, both SVM and I-SVM are unable to improve TAR without an unacceptable increase of FAR. To explain this behaviour, we need to recall a specific detail.

In Figs. 3.9b and 3.10b, initial and subsequent TAR measurements of I-SVM and SVM are quite high. This means that the model can incorporate much more genuine information (see Fig. 3.11). At the end of the experiment, we have a model that has acquired almost the same genuine information as the supervised case. Moreover, as counter-intuitive as it may seem, FAR seems to increase whenever a genuine sequence is presented, while FAR seems to decrease after an impostor query. This is especially noticeable when testing on COX. However, this behaviour is coherent with the one explored in the 3.5.1. In that experiment, we show the importance of balance stability when using a constant threshold. Therefore, adding this amount of genuine information (high TAR) leans the classification problem to a more balanced one, making the initial threshold obsolete (FAR increases after genuine queries).

Finally, an essential detail to remark is the decreasing FAR observed for De-SVM and OS-ELM. This behaviour continues in the supervised scenario (Sec. 3.5.1). While De-SVM presents a soft, monotonous decrease; OS-ELM presents a sharp decline at the beginning with a slight tendency change at the end.

Relation of Genuine/Impostor Trains

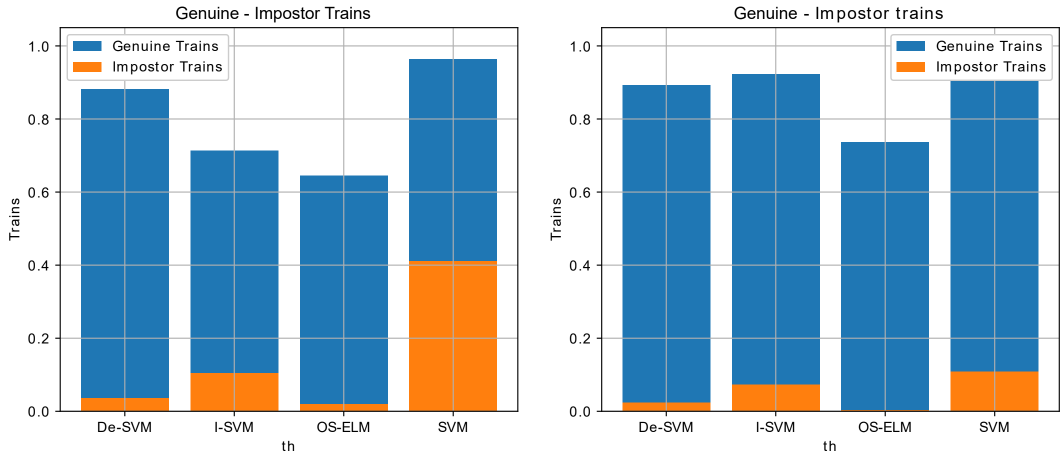
Paying attention to the relation of genuine and impostor update rate (over the total possible updates), we can extract some key conclusions (Fig. 3.11). Firstly, I-SVM and SVM are quite more sensible to model corruption due to false acceptance updates. This is coherent with the behaviour observed using transaction-level performance. This corruption negatively affects, in particular, SVM. The behaviour is even more relevant if we remember that SVM is one of the best-performing methods in the supervised experiments. A possible explanation for this behaviour is further studied on 3.5.1.

Secondly, we can infer from the genuine/impostor updates that the most benefits are obtained from acquiring enough genuine information, even if it is at the cost of including some impostors. This justifies the assumption of setting the operational threshold to a 5% FAR.

3.5.3. Post-robustness impostor testing

In static classification scenarios, regular FAR measures robustness against impostors. However, systems studied here are non-static, making FAR non-static as well. Even more considering that the system's predictions are used as pseudo-labels during the updates, making false acceptances susceptible to feedback. This experiment intends to test each classification technique in the extreme scenario where the system is repeatedly queried with only impostor queries.

The idea is to start from the previous experiment. The initial template of 5 frames is maintained as well as the CM with the genuine identity and the 10 most



(a) FACE COX database.

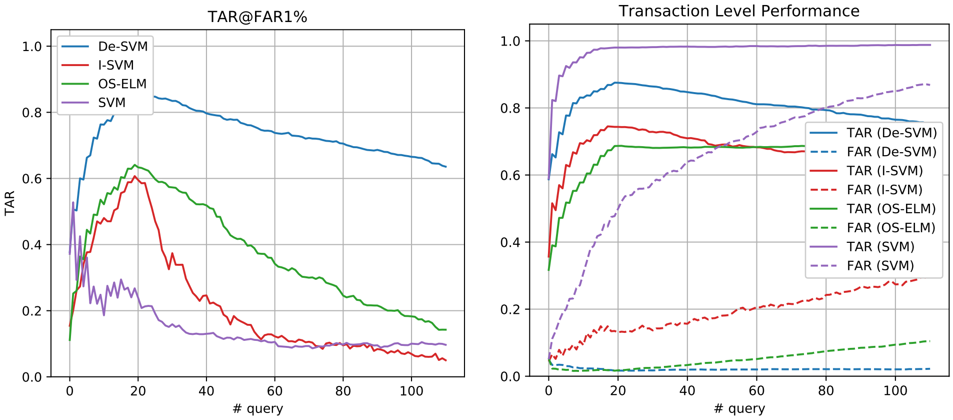
(b) YouTube Faces database.

Figure 3.11: Genuine and impostor updates performed by each classification technique in each database.

similar impostors. So, after the 20th query, we will present a total of 90 impostor queries (every identity from the CM of the genuine target) with additional subsequences of the *gallery subset*. The fact that the CM is maintained (with the 10 most similar impostors) particularly augment the difficulty of the robustness testing. The pattern followed was:

$$Sec.\ 3.5.3 \begin{cases} Sec.\ 3.5.2 \left\{ G^1 \rightarrow I_1^1 \rightarrow \quad \ldots \quad \rightarrow G^{10} \rightarrow I_{10}^1 \rightarrow \right. \\ \qquad\qquad\qquad \left\{ {I_1}^2 \rightarrow \quad \ldots \quad \rightarrow I_{10}^2 \rightarrow \quad \ldots \quad \rightarrow {I_1}^{10} \rightarrow \quad \ldots \quad \rightarrow I_{10}^{10} \right. \end{cases}$$

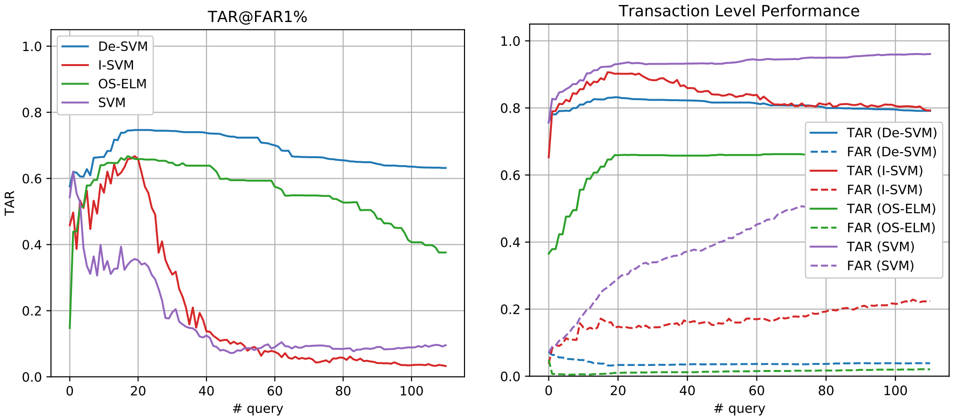
The procedure to measure performance is the same as the previous experiment (using the *probe set* with identities of the CM), using the metrics described in Sec. 3.4.5. Measurements are performed after each of the queries. Results obtained can be seen in Figs. 3.12 and 3.13. Overall, the first thing we observe is that every method suffers a performance loss in this testing. Nevertheless, looking at Fig. 3.12a, we realise De-SVM's outstanding resistance compared to the other methods. TAR@FAR1 moves from 86.03 ± 0.48% and 75.5 ± 1.2% (COX and YTF, respectively) in query 20 to just below 64.4 ± 1.6% and 64.8 ± 1.1% in query 110. The following best performing technique is OS-ELM that goes from 75.0 ± 1.1% and 66.4 ± 1.9% to a final performance of 19.3 ± 2.7% and 32.2 ± 5.2% TAR@FAR1.



(a) TAR@FAR1 performance.

(b) TAR and FAR at the operational threshold.

Figure 3.12: FACE COX: Self-supervised performance comparison fixing the operational threshold at 5% initial FAR, testing robustness.



(a) TAR@FAR1 performance.

(b) TAR and FAR at the operational threshold.

Figure 3.13: YouTube Faces: Self-supervised performance comparison fixing the operational threshold at 5% initial FAR, testing robustness.

Transaction level performance (using the *operational threshold*) shows interesting behaviours again. Intuitively, one could think that a repeated impostor querying would make FAR out of control. Indeed, this behaviour is observed in 3 of the 4 classification techniques that have been tested (mainly in SVM and I-SVM and softer in OS-ELM). This is something that makes sense, given the fact that the only possible mistake is to accept an impostor query as genuine.

However, De-SVM presents just the opposite behaviour. Most damage comes from a decreasing TAR instead of an increasing FAR. To understand this effect, we need to go deeper into ensembles' nature. Ensembles decisions are based on majorities. The majority of accepting genuine identity is built during the initial stages (query 0 to 20). Based on Fig. 3.11, this decision is supported by 9 out of 10 classifiers. It would be necessary to overcome this majority of 9 classifiers to confuse an impostor with a genuine persistently. This is something quite difficult given the fact that FAR is always below 5%.

In other words, impostor classifiers may agree on rejecting genuine identities (TAR decrease), but they cannot agree on accepting another identity as genuine (FAR stability). This is a desirable behaviour for fields like biometric identification, in which the main concern is to avoid impostors entering the system.

3.5.4. Summary and Discussion

As aforementioned (Sec. 3.1 and Sec. 3.2), the limited amount of previous literature framed in the particular conditions of our problem makes it difficult to establish a direct comparison with other previously used techniques. In this regard, [86] uses ensembles of very-specific SVM classifiers like De-SVM. Nevertheless, the fact that we are focusing on face related problems (in which their proposed calibration phase is problematic) prevent us from using their work as a fair and acceptable baseline. Thus, we opted to establish a benchmark with static methods trained with the available labelled data ($n = 5$ frames).

The use of deep feature embedding in combination with traditional classifiers is quite common in the literature and has proven to achieve comparable performance to end-to-end deep learning methods [22]. Thus, the baseline is established us-

Table 3.4: Summary of TAR@FAR1% performances values obtained (values in %). Uncertainty is not represented in previous graphs for the sake of clarity. SU stands for self-updating.

COX				
Model	Initial	Superv. Adapt.	Unsuperv. Adapt.	Robustness
RN29 + SVM	37.19 ± 0.63	-	-	-
RN50-AF + SVM	**51.6 ± 1.0**	-	-	-
RN29 + SVM + SU	37.19 ± 0.63	88.89 ± 0.74	24.5 ± 1.0	10.04 ± 0.48
RN29 + I-SVM + SU	17.7 ± 3.6	79.8 ± 1.2	61.3 ± 1.5	5.51 ± 0.49
RN29 + OS-ELM + SU	10.27 ± 0.51	**92.35 ± 0.45**	75.0 ± 1.1	19.3 ± 2.7
RN29 + De-SVM (**Ours**)	37.17 ± 0.58	89.47 ± 0.24	**85.45 ± 0.25**	**64.4 ± 1.6**
YouTube Faces				
Model	Initial	Superv. Adapt.	Unsuperv. Adapt.	Robustness
RN29 + SVM	55.9 ± 1.3	-	-	-
RN50-AF + SVM	**81.33 ± 0.58**	-	-	-
RN29 + SVM + SU	55.9 ± 1.3	**88.68 ± 0.42**	34.7 ± 1.5	9.0 ± 1.1
RN29 + I-SVM + SU	42.0 ± 2.2	73.9 ± 2.2	65.5 ± 2.5	2.26 ± 0.67
RN29 + OS-ELM + SU	13.8 ± 1.8	76.8 ± 2.0	66.4 ± 1.9	32.2 ± 5.2
RN29 + De-SVM (**Ours**)	56.91 ± 0.59	76.26 ± 0.75	**75.5 ± 1.2**	**64.8 ± 1.1**

ing the ResNet-29 (RN29+SVM) feature representation described in Sec. 3.4.2 and ResNet50-AF (RN50-AF+SVM) feature representation [26] (which tops the state-of-the-art on LFW verification benchmark).

Tab. 3.4 shows the complete set of results of this chapter. The first two rows contain the baselines we have just mentioned. The rest of the rows contain the four classification techniques compared in this chapter. The table shows, in each column, their initial performance, final performance after supervised adaptation (Sec. 3.5.1), final performance after unsupervised adaptation (Sec. 3.5.2) and performance after the robustness test (Sec. 3.5.3).

Overall, it is evident that self-updating is an interesting approach to perform unsupervised adaptation. It can improve initial performance in 3 of the 4 tested methods (including De-SVM). This improvement is enough to overcome the state-of-the-art performance with the available data when using the COX database. This result is remarkable given the short number of gallery samples (5 low-quality video frames). Nevertheless, in the case of YTF, AF+RN50 [26] stills beats the methods that include self-updating. A possible explanation for this is two-folded. First, unlike RN29, this feature representation is designed with the YTF test in mind [26] acquiring state-of-the-art performance in this dataset too. Secondly, YTF is not a database designed for the specific problem of video surveillance. Consequently, most of the specific built-in characteristics (e.g. variable scale and light conditions, blurriness, etc.) are less present in video-frames (Fig. 3.3). This fact makes easier the transition between stills and videos.

Comparing to supervised adaptation (Sec. 3.5.1), every model experiences a predictable drop in performance when updating phase is done under unsupervised conditions. In this regard, De-SVM is the method that experiences the smallest of all. Thus, De-SVM can achieve comparable performance using less than a tenth of labels. On the other hand, SVM is the method that experiences the highest drop in this comparison.

Finally, in terms of robustness, De-SVM presents impressive characteristics in comparison to other classification methods. Moreover, its behaviour is even more remarkable, given that the only labelling used is the one to create the initial template. Besides, the fact that the primary performance damage is caused by lowering TAR and not increasing FAR represents a promising quality for any biometric application.

3.5.5. Further testing on De-SVM

Using other central tendency measures in the FDF and SDF

This experiment shows the effect of using different central tendency measures in the FDF and SDF. Results on Tab. 3.5 show that the differences in performance are not significant. Thus, the choice of the median was merely conventional. Conceptually, the median is equivalent to perform a majority vote after obtaining a binary

Table 3.5: COX: Comparison on the function used in the FDF and SDF.

	TAR@FAR1		TAR		FAR	
Function	Initial	Final	Initial	Final	Initial	Final
Median	37.17±0.58	85.45±0.25	59.01±0.37	88.14±0.24	5.01±0.13	1.714±0.052
Mean	36.56±0.56	84.27±0.48	59.00±0.26	87.39±0.32	5.01±0.12	1.681±0.040

classifier response (using the *operational threshold*).

Effects on performance of different operational thresholds and template sizes

In this section, we explore the effects of having different operational thresholds and templates sizes on De-SVM. To do that, we repeat the previous experiment on the COX database varying these parameters. Results can be seen on Tabs. 3.6 and 3.7.

Analysing the operational threshold dependence (Tab. 3.6), we observe TAR@FAR1 increases as the operational threshold becomes less strict. However, the selected 5% FAR is an inflexion point from which the gain begins to be more subtle. Looking at TAR and FAR performance, we observe that the increase of final TAR is done at FAR expenses. In this regard, we can assume then the 5% operational threshold as an acceptable compromise.

Turning now to the template size effect (Tab. 3.7), we observe that most of the performance improvement appears initially. Again, considering the final performance, there is an inflexion point at the selected size of 5 frames. Since we want to address data scarceness, we want to extract the maximum power from the minimum amount of labelled information. Therefore, based on this behaviour and the one observed in [81], the election of a 5 frames template seems quite reasonable.

Table 3.6: COX: Study on FAR point of operational threshold De-SVM (template size = 5 frames).

Operational	TAR@FAR1		TAR		FAR	
threshold	Initial	Final	Initial	Final	Initial	Final
1%	37.17±0.58	74.08±0.67	30.92±0.32	57.1±1.1	0.568±0.061	0.118±0.031
3%	37.17±0.58	82.61±0.68	48.41±0.21	79.08±0.52	2.407± 0.048	0.650±0.024
5%	37.17±0.58	85.45±0.25	59.01±0.37	88.14±0.24	5.01±0.13	1.714±0.052
7%	37.17±0.58	86.03±0.48	62.39±0.31	89.80±0.25	6.23±0.16	2.53±0.11
10%	37.17±0.58	86.69±0.40	68.90±0.22	93.01±0.43	9.34±0.10	4.90±0.10

Table 3.7: COX: Study on template size De-SVM at operational threshold 5%.

Template	TAR@FAR1		TAR		FAR	
size	Initial	Final	Initial	Final	Initial	Final
1	20.30±0.64	52.42±0.41	32.00±0.28	53.68±0.38	3.36±0.10	1.178±0.053
3	30.19±0.76	77.91±0.52	45.47±0.50	78.09±0.30	3.213±0.080	1.036±0.076
5	37.17±0.58	85.45±0.25	59.01±0.37	88.14±0.24	5.01±0.13	1.714±0.052
7	40.74±0.49	86.69±0.65	60.90±0.41	87.32±0.49	3.938±0.096	1.138±0.073
10	45.21±0.77	88.87±0.24	65.60±0.23	89.02±0.21	4.098±0.095	1.024±0.043

3.6. Conclusions

This chapter has tackled the problem of V2V-FV in a context with no collaborative manual enrolment. In combination with state-of-the-art CNN features, self-updating has proven to be an interesting approach as pseudo-labelling for incremental learning purposes during the operational phase. In this regard, De-SVM uses all of these ideas to incorporate new relevant unsupervised data while maintaining a relatively low FAR.

De-SVM behaviour becomes even more promising because the full power of ensemble-based approaches is not fully exploited yet. Indeed, one of the main benefits of these type of methods is that they allow isolating each of the updates, making them reversible. Moreover, as the updating process consists of adding new classifiers (*learn*), by providing a way of removing classifiers, the ability to *forget* could be easily implemented. These ideas will be explored in the following Chapter 4 when extending De-SVM from verification to general recognition.

Chapter 4

Adaptive Face Recognition in Video-surveillance

This chapter is heavily based on the contents of these articles:

E. Lopez-Lopez, C. V. Regueiro, and X. M. Pardo. An adaptive video-to-video face identification system based on self-training. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2590–2596, 2021

E. Lopez-Lopez, C. V. Regueiro, and X. M. Pardo. Incremental learning from low-labelled stream data in open-set video face recognition, 2020

4.1. Introduction

In the previous chapter, De-SVM has proven to be an interesting approach to implement unsupervised incremental learning for face verification in video surveillance contexts. Here, the aim is to extend the proposed system to work to the more general task of face recognition. Contrary to what one could expect in the first place, this extension is far more tricky than having n-verification systems in parallel [47].

As aforementioned, incremental learning is the ability of a classifier to evolve by continuously integrating information from new instances and/or new classes without resorting to complete retraining [58]. In the context of multi-class classification (as is the case of face recognition), most efforts have been focused on extending the class-set of a classifier considering only labels from the new classes while avoiding the problem of catastrophic forgetting [128, 106]. This is particularly important when the computational power prevents full retraining or privacy issues impede the new access to previous samples. In contrast, less progress has been made in continual learning of a set of non-stationary classes, mainly when applied to tasks involving unsupervised streaming data.

Video-to-video face recognition (V2V-FR) in video surveillance is a paradigmatic example of the application of incremental learning due to the specific context conditions mentioned in Chapter 3 (e.g. high pose variations, resolution or illumination). This variability in the conditions often excess the diversity available in datasets used to train deep networks (generally focused on web extracted images) [46, 126] and requires new alternatives.

When extending a face verification system to a recognition scenario, another relevant aspect is the consideration of closed-set (only enrolled identities query the system during operation) or open-set (non-enrolled identities can query the system) classification [119]. The decision of distinguishing between enrolled and non-enrolled identities (*known* and *unknown*, respectively) has proven to be quite challenging [119, 47, 10]. Still, for real-world applications of V2V-FR, the open-set consideration should be a must. Take, for example, a practical case of an airport video-surveillance aimed to track some *individuals of interest (IoI)* who have not been collaboratively enrolled in the system, e.g. those exhibiting suspicious behaviours, among a larger number of non-target identities, that should be identified as *unknown* identities.

This chapter proposes the Open-Set Dynamic Ensembles of SVM (OSDe-SVM), an extension of the De-SVM (see Chapter 3) to the more complex Open-set V2V-FR problem. The architecture of OSDe-SVM (Fig. 4.1) follows the same philosophy of De-SVM: from deep feature embeddings, the system acts as an incremental learning module, fed with stream data, which simultaneously predict and update classifiers using the self-training strategy [146]. Like De-SVM, the approach is based on dynamic ensembles of very specific SVM classifiers that combine their responses using Extreme Value Theory to work in the open-set. Additionally, exploiting the modular nature of ensembles, adaptations consist of either adding or removing classifiers.

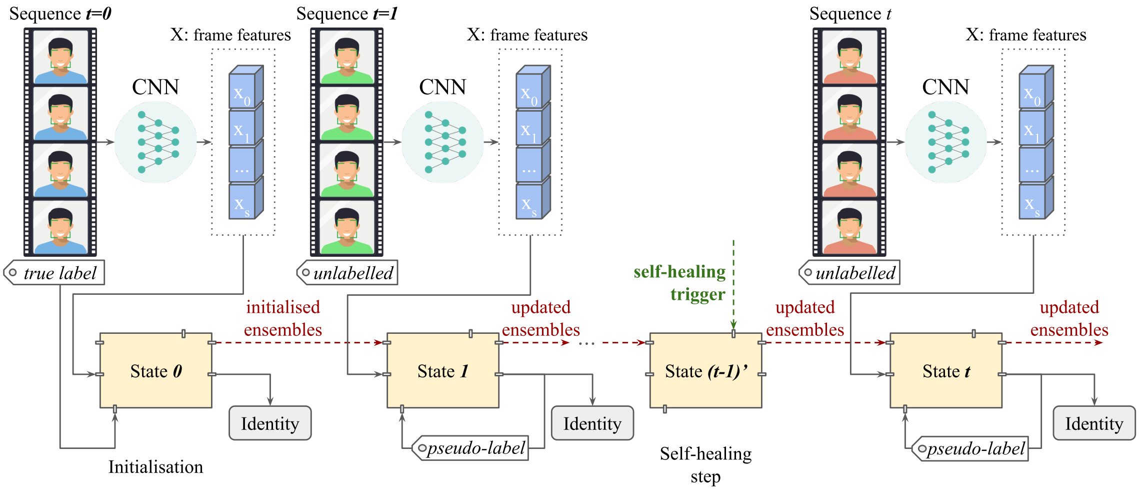


Figure 4.1: Open-Set Dynamic Ensembles of SVM (OSDe-SVM) is able to incorporate new knowledge and correcting wrong updates by adding and removing classifiers in an unsupervised way. The system is designed to work under open-set recognition conditions.

The main contributions of this chapter are:

- An approach to unsupervised incremental face recognition designed to operate online with stream data. During its operation, predictions also play the role of pseudo-labels.
- A strategy to deal with both catastrophic forgetting issues and the effect of mistaken pseudo-labels.
- An approach to instance-incremental learning in the open-set, which could be extended to cope with the class-incremental problem.
- A method for person re-identification based on face, which is not directly based on a reservoir of face images.

4.2. Related Work

Open-set Recognition. In open set recognition, training is performed on a dataset with samples of some known classes, while samples of both known and unknown classes are presented for testing. Therefore, classifiers should appropriately deal with all of them [38]. Within this approach, closer to real-world applications, decision boundaries not only separate instances of different known classes, but they separate the known from the unknown as well [34, 119, 8, 95]. A recent survey [38] distinguishes between discriminative and generative approaches to open set recognition. Discriminative classifiers are trained to discriminate between the known classes and then, given the most likely class label, to decide whether a test sample was in fact drawn from the distribution of known class samples or not [47]. Meanwhile, generative methods try to provide explicit probability estimation over unknown categories, most of them based on deep networks [37, 100]. Plenty of methods in both sets of approaches leverage Extreme Value Theory (EVT) to tackle the unknown [18, 38]. EVT is a branch of statistics aimed to assess the probability of observing an event more extreme than any previously observed and has been widely used for outlier detection [123], attribute fusion [120] and in open-set recognition [114].

In face recognition, the most realistic scenario corresponds to an open-set setting (e.g. criminal watch-lists, restricted areas access control, smart-homes, etc.) [47]. In this domain, apart from EVT based methods, solutions based on siamese networks have been proposed to address the open-set as they are metric learning methods, and their similarity scores can be thresholded to perform recognition [117]. Although they do not fit the data stream context, they could be used as a baseline for comparison purposes [121].

Incremental Learning for Multi-class Classification. The main goal of incremental (a.k.a lifelong, continuous or continual) learning is to learn from data as real-world dynamic sources provide them, usually at a low pace, including noisy samples and, in general, exhibiting non-stationary properties. As data distributions change with time, computational systems have to deal with the *stability-plasticity dilemma*, trying to avoid new knowledge to erase old one (*catastrophic forgetting*), while detecting and adapting to concept drifts [105, 116].

From a multi-class classification perspective, deep continual learning methods have been focused on learning new tasks/classes, more than on enhancing the performance of classifiers (fixed number of classes) as new instances arrive [87, 98]. Among common strategies are the exploitation of, at least, partial rehearsal (looping over old data) [58, 50, 1], dynamic changes in architectures (retraining after pruning/increasing the number of neurons, filters or layers), and regularisation (updating weights in order not to forget previous knowledge) [62]. Among the last are usually also included a wide range of knowledge distillation methods, in which a teacher network transfer knowledge to a student network [74, 13, 147]. However, the drawback of distillation is that it generally needs to retain big past memories [7, 128]. Notwithstanding the progress made in supervised incremental learning in recent years, there is still a substantial gap between the performance of batch offline learners on stationary data and the performance of the incremental learners that deal with non-stationary data [50, 58].

Most of the semi-supervised methods leverage unlabelled examples by making some assumptions, using label propagation or generating pseudo-labels during the learning process [72]. Some approaches are based on keeping a set of dynamic clusters to summarise class distributions and model their evolution over time [133]. Others use a few labelled data to initialise a set of models, which are afterwards sequentially updated based on pseudo labelled data [68, 104, 94]. In the specific case of video recognition, weak labels can be provided by the temporal tracking [36, 102], but also co-training or predictions of the own classifiers can provide pseudo-labels (self-training).

In this regard, ensemble methods have been acknowledged as powerful tools to overcome *catastrophic forgetting* [105, 19, 58], when dealing with data streams [66, 68, 80]. Moreover, ensemble algorithms can be integrated with drift detection algorithms and incorporate dynamic updates, such as selective removal or addition of classifiers [41]. In the semi-supervised scenario, it must be considered that any kind of weak labelling or pseudo labelling is prone to error. So, dynamic updates can also be useful to healing from the effect of mislabelling. Unlike other incremental learning approaches (either classic [63, 75] or DL-based [51]), ensembles provide simple way to isolate updates and, consequently, make changes reversible. And not only that, since decisions are based on majorities, ensembles are robust to outliers.

Ensembles of deep networks have been proposed to encourage networks to cooperate and take advantage of their prediction diversity in the context of few-shot classification [31]. Besides, to deal with tasks where training data are inadequate, the training of a collection of incrementally fine-tuned CNN models and their combination using an ensemble was presented [151]. In [45], the authors propose an ensemble learning framework based on multiple CNN classifiers. The CNN acts as a feature extractor for the posterior use of different ensemble frameworks to classify its content. Recently, already in the context of incremental learning, an approach based on ensembles, which is close to ours, was proposed for tackling the problem of mechanical fault diagnosis [137].

4.3. Open-Set Dynamic Ensembles of SVM (OSDe-SVM)

This chapter presents the pipeline of *Open-Set Dynamic Ensemble of SVM (OSDe-SVM)* for the problem of V2V-FR in the open-set (Fig. 4.1). Like De-SVM, this method takes advantage of transfer learning from large labelled datasets to get discriminant feature embeddings that feed an instance-incremental learning module.

In the case of OSDe-SVM, for the encoder part, the choice was to use deep embeddings taken after the last pair of convolutional and batch normalisation layers of the ResNet100-ArcFace (RN100-AF) network trained on MS1MV2 dataset [26]. This is one of the top-performing CNN in the face recognition state-of-the-art. ArcFace is a loss-function specifically designed to enhance the discriminative power of face recognition models, being the deepest networks, as ResNet-100, the ones that take the most advantage of it [26]. The encoding transforms a 112x112 face crops into a 512-D feature embedding.

The general structure of OSDe-SVM is depicted in Fig. 4.2 follows a very similar structure compared to the verification case (Sec. 3.3.2), but specially adapted and fine-tuned for the present case. Each individual of interest (IoI), k, has an associated ensemble, e^k, composed by a set of SVM classifiers, h_i^k. This ensemble is updated whenever the system is queried. The updates are driven by the responses of the Ensemble's Decision Functions (Sec. 4.3.1), following the self-training paradigm (Sec. 4.3.2). Besides, OSDe-SVM can remove classifiers when the maximum number of classifiers is reached (Limitation Module, Sec. 4.3.3) or when a possible mistake is detected (Self-healing, Sec. 4.3.4).

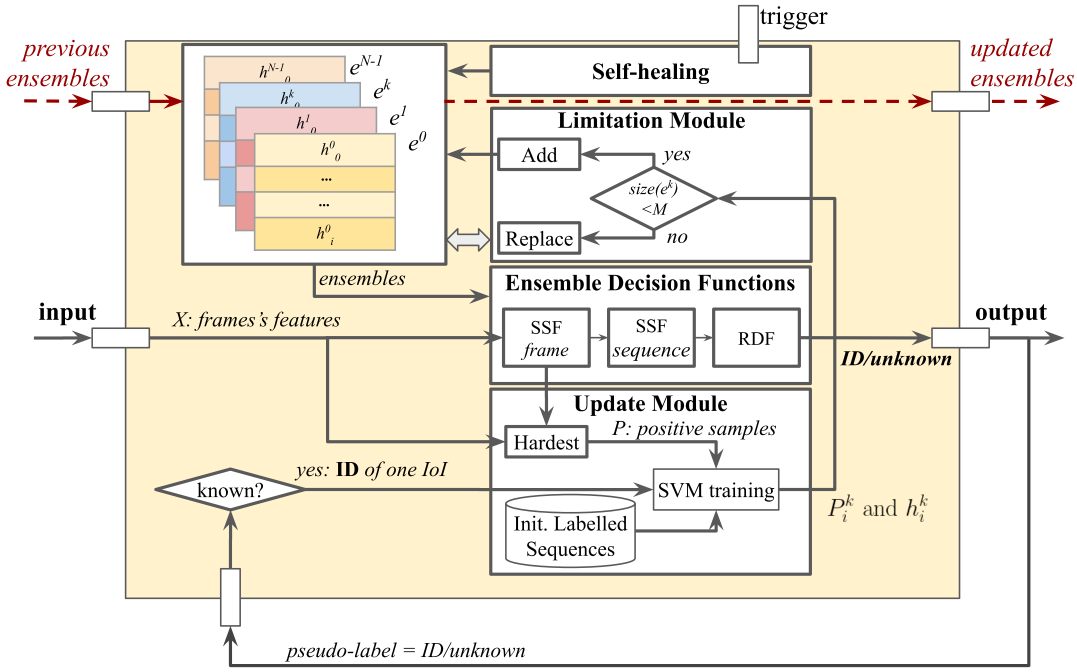


Figure 4.2: Pipeline of OSDe-SVM.

4.3.1. Ensemble's Decision Functions

OSDe-SVM ensembles make their decisions in a two-step process. First, the *Sequence Scoring Function* assigns a particular score to the query sequence. And second, the *Recognition Decision Function* uses these scores to assign an identity label (either as one of the IoI or as an unknown).

Sequence Scoring Function (SSF)

When making decisions, it is convenient that each ensemble gives a unique score to each incoming sequence. Nevertheless, both (sequences and ensembles) are composed elements. Being n_F the number of sequence's frames and M^k the number of classifiers of ensemble k, we would have a total of $n_F \times M^k$ different responses. All of these responses are combined into a unique score by the use of the *Sequence Scoring Function (SSF)*. This process consists of two levels:

- At *frame level*, the equivalent of the FDF in De-SVM (Sec. 3.3.1), the responses of the ensemble's classifiers are combined to give a unique score to each frame.

The function used here is the median of the individual ensemble's SVM scores. In practice, this corresponds to a majority voting.

- At *sequence level*, the equivalent of the SDF in De-SVM (Sec. 3.3.1), the temporal coherence assumption allows to assign a unique identity to the whole input sequence, combining the frame's scores into a unique one. The function used here is the median.

Recognition Decision Function (RDF) based on Extreme Value Theory

Once every ensemble delivers its prediction score about an input query, the next step is to combine all the predictions to decide the underlying identity. The identity assignment based on the best score is the usual procedure in a closed-set scenario [26, 76, 79]. That is because input sequences always belong to a known IoI. In an open-set scenario, assigning identities becomes trickier because *non-match responses*, (corresponding to unknown identities) need to be also expected. To tackle these scenarios, OSDe-SVM was endowed with a *Recognition Decision Function (RDF)* based on the Extreme Value Theory (EVT), which also allows to deal with the non-calibrated outputs provided by SVM's.

When applying EVT, we follow an approach similar to [118]. As any input sequence belongs to a unique identity, the ensembles associated with other identities should deliver *non-match* outputs. According to the Fisher-Tippet-Gnedenko Theorem of EVT [18], the distribution of these *non-match* scores is modelled by some particular functions.

In this case, for left bounded positive samples, the distribution of minima, $G(z)$, is given by the Reversed Weibull distribution. We can perform a simple transformation to the ensemble's scores (given by the SSF) to satisfy these conditions and be able to fit a Weibull distribution to the tail of the distribution, as is depicted in Alg. 2. Then, to discriminate between *unknown* and *known* identities, the best ensemble response (the best score) can be checked whether it comes from this *non-match* distribution (Fig. 4.3) or not.

More importantly, Alg. 2 also provides a way of distinguishing the *known* from the *unknown* by thresholding the Weibull distribution (T_W), instead of the actual

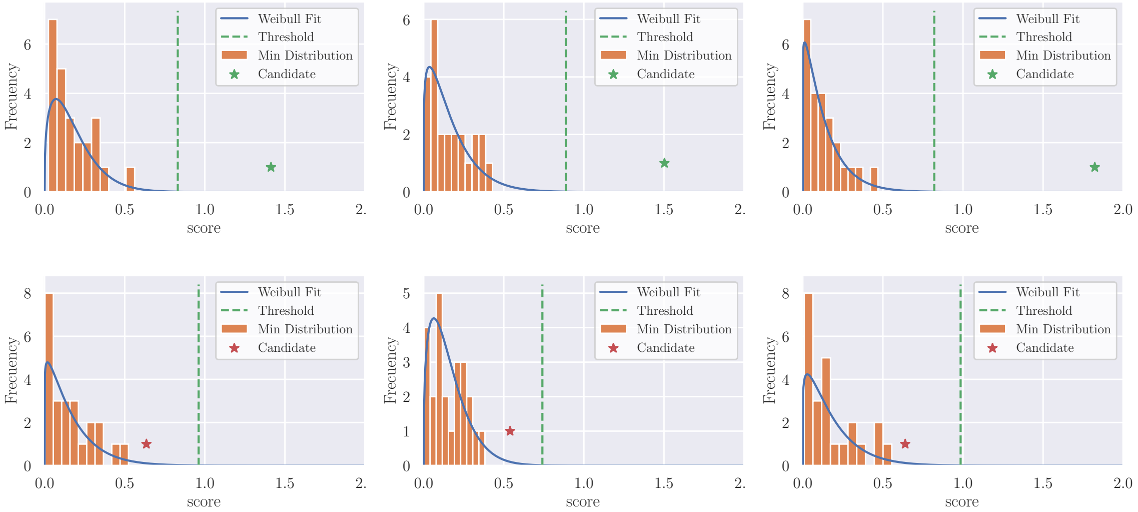


Figure 4.3: Examples of Extreme Values distributions and application of RDF. A Weibull function is fitted to the min distribution to see whether the candidate belongs or not to this distribution. First row illustrates examples of *known* identities and second row does the same with the *unknwon* ones.

scores. Since the fitted function is different depending on the input sequence, we are implicitly personalising the threshold to each input sequence, as is depicted in Fig. 4.3.

4.3.2. Update Module: Incremental learning based on Self-Updating

This module is OSDe-SVM equivalent of the De-SVM Update Function (Sec. 3.3.1). As OSDe-SVM is conceived to operate in the context of a shortage of labelled data. Only the first classifier of each ensemble is trained with a very short labelled sequence extracted from the input. The first five frames have proven to be the bare minimum for both OSDe-SVM and De-SVM [81]. From that point on, incremental learning is exclusively based on pseudo-labels (Fig. 4.1).

After an ensemble is initialised, the method decides whether a new classifier must be added up to enhance future performance, each time a sample of the same identity is identified. OSDe-SVM follows a self-updating strategy based on pseudo-labels provided by EDF to input sequences. Whenever an identity, k, is identified

Algorithm 2 Recognition Decision Function (RDF) based on EVT.

```
$S$ is the input sequence, $T_W$ is the threshold in the Weibull function
$E = \left\{e^0, e^1, \ldots, e^{N-1}\right\}$ set of ensembles associated to *known* identities
$R = \{\varnothing\}$ set of scores given by each ensemble to a candidate
**for** $e^i$ in $E$ **do**
    $R \leftarrow SSF(e^i, S)$
**end for**
$c = \min(R)$; $m = median(R \setminus c)$
$V = \{\|x - m\| \mid x \in (R \setminus c) \wedge (x < m)\}$
Fit V to a Weibull function, $W$
**if** $W(\|c - m\|) < T_W$ **then**
    ID $= arg(c)$
**else**
    ID $=$*unknown*
**end if**
```

in an input sequence, a new SVM is created using as (pseudo-labelled) positive samples, P_j^k, the 5 **hardest** frames of the sequence, namely those which got the lowest scores returned by the SSF (see Fig. 4.2). This way, diversity within each ensemble is encouraged.

4.3.3. Limitation Module

In a self-updating context where each ensemble is initialised with only one classifier trained with a few labelled frames, further updates can only occur when close samples of the same identity query the system. If they are almost identical, there is nothing to be learnt. However, if they are very different, there is a danger of not being identified. So, the model can only learn from samples in the borderline, i.e. samples that can still be recognised by the ensemble of the corresponding identity and include some level of novelty in their features. However, ensembles' size should not grow indefinitely whenever the *EDF* recognised their target identities in input sequences. As ensembles' performance relies on diversity, we have chosen a solution inspired in [71, 42] to decide which classifiers need to be removed once the maximum size is reached.

Within each ensemble, classifiers are compared against each other to obtain a measurement of their relative relevance, the *diversity score* $D(\cdot)$. Given an ensemble, e^k, composed by M_k SVM classifiers, $\left\{h_0^k, h_1^k, ..., h_{M_k-1}^k\right\}$, $D(h_i^k)$, is computed from the binary response of each of the classifiers of the ensemble over a certain set of video frame features $\{x_0, x_1, ..., x_{Q-1}\}$:

$$D(h_i^k) = \sum_{j=0;j\neq i}^{N-1} d\left(h_i^k, h_j^k\right) \tag{4.1}$$

$$d\left(h_i^k, h_j^k\right) = -\frac{1}{Q}\sum_{q=0}^{Q-1} sgn\left(h_i^k(x_q)\right) \cdot sgn(h_j^k(x_q)), \tag{4.2}$$

where $h_i^k\left(x_q\right)$ is the response of the SVM classifier h_i^k to the frame feature x_q, and $sgn(\cdot)$ is the sign function.

Whenever an ensemble e^k reaches the maximum size, the classifier h_*^k with the lowest diversity will be removed.

4.3.4. Self-healing: Correcting Wrong Updates

Since the whole adaptation process performs **without supervision**, wrong updates, provoked by errors in pseudo-labelling, should be expected. This behaviour may affect re-identification performance, mainly in the long term. The *self-healing* procedure is designed to mitigate this problem.

Self-healing relies on the fact that the ensembles build its decisions based in majorities. Therefore, if an ensemble reaches a relatively high accuracy in the first classifications, it should be difficult for wrong classifiers to take over very soon. This fact opens the possibility of detecting wrong updates before it becomes irreversible. We expect that, with a limited amount of wrong updates, ensembles are still able to recognise their target identity. Consequently, the future detection of the target identity can build a stronger majority capable of detecting the previous wrong update.

To implement these ideas, along with each SVM classifier, h_i^k, we store the positive samples used to create it, P_i^k, which, in practice, can be considered a sequence.

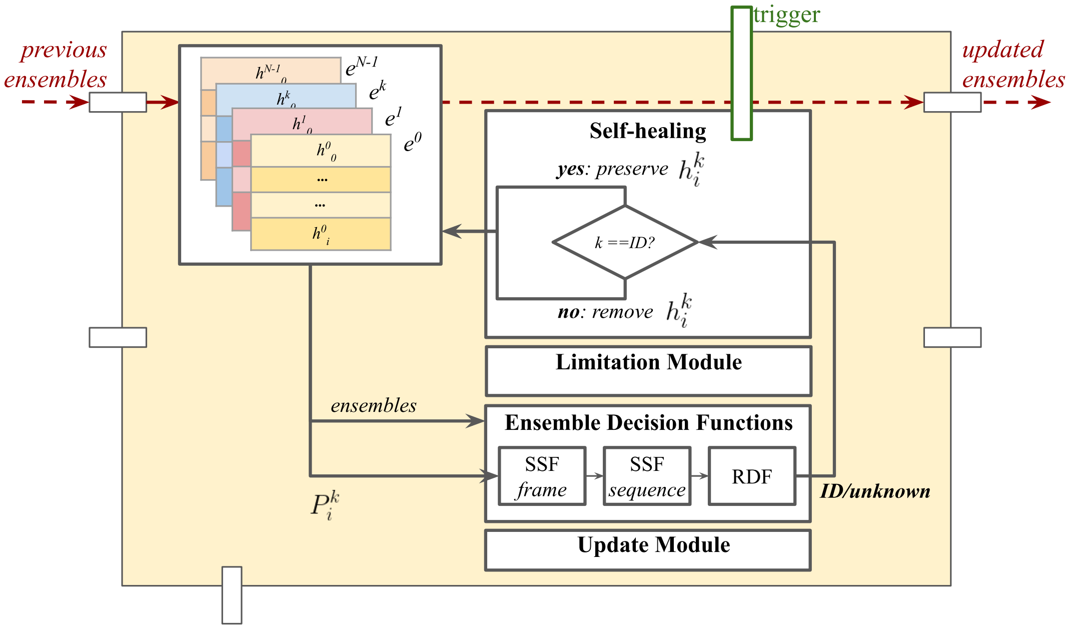


Figure 4.4: Pipeline of OSDe-SVM when self-healing is performed.

Therefore, we can pass every set (for all k and i) again through the *EDF* for a re-evaluation. If the system assigns the same identity as before, the classifier is maintained. Otherwise, the classifier is removed (Figs. 4.4). The self-healing module triggers after a certain period which is adjustable (see Fig. 4.1).

4.4. Methodology

4.4.1. Database Selection

As it has been explored in Chapter 2, image quality (especially in terms of resolution) has a profound impact on the performance of face identification methods [44, 43, 54]. Video surveillance datasets are also affected by this fact. While some of them, as in the case of CMU FiA [40], allow deep learning methods trained with general face datasets to achieve a pretty good performance, others, as the COX Face dataset [56], possess highly marked characteristics that reduce performance. In this section, these two datasets are presented to illustrate this problem.

COX Face Database

The details of COX Face database [56] were already comprehensively described in the last Chapter (Section 3.4.1). Here, the face detection module used here is the one described in [148] due to its integration with the feature encoder module. Its primary purpose will be to fine-tune the background removal of the face tracker and for proper alignment.

Like the verification case (Chapter 3), it was necessary to perform some adjustments to adapt the data provided by the COX database to how we operate. Similarly, the first adaptation consists of dividing each of the available sequences into 3 sub-sequences. However, despite following the same philosophy, the organisation in the different sets would be slightly different to simulate to the recognition conditions properly:

- The **initially labelled sequences** are labelled video-frames of target identities used to create the first classifier of each ensemble (the sets positive samples, P_i^k). They consist of the first 5 frames from `cam1` from the 1 000 individuals.
- The **operational sequences** simulate input sequences which would be received in the operational phase. They consist of the three sub-sequences of `cam1` and `cam2`, and the first two sub-sequence of `cam3` of the same 1 000 individuals of the *initially labelled sequences.*
- The **testing sequences** are used to assess performance. They correspond to the last sub-sequence of `cam3` of the 1 000 individuals.

Hereafter, each sequence will be noted by S_t^k, where t refers to temporal order and k refers to the identity. Following this notation, $t = 0$ corresponds to the 5-frame sequences of the *initially labelled sequences* used in the initialisation, $t = 1, 2, \ldots, 8$ correspond to the streaming of sequences (*operational sequences*), and $t = 9$ corresponds to a sequence for performance assessment (*testing sequences*).

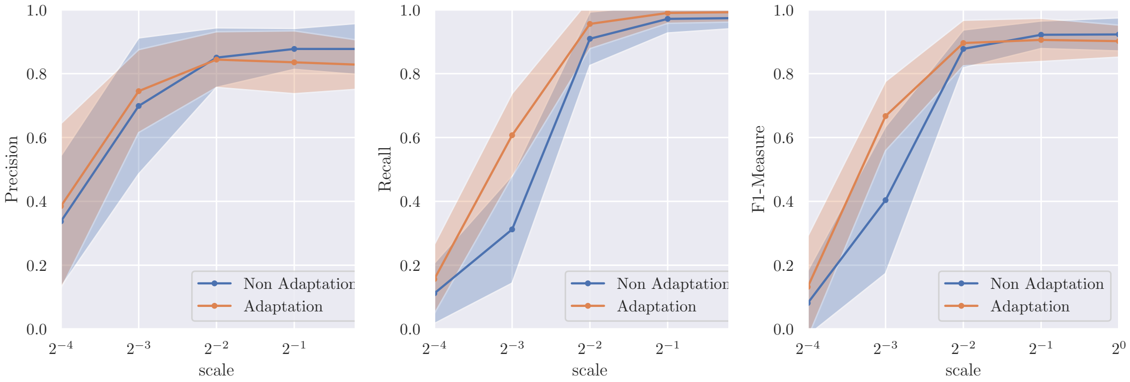


Figure 4.5: Performance versus image resolution, scales 1, 1/2, 1/4, 1/8 and 1/16 (that is, 112x112, 56x56, 28x28, 14x14, and 7x7 pixel image sizes).

CMU Face in Action (FiA) database

The CMU FiA database contains 20-second videos of more than 200 different individuals simulating a passport checking scenario in both indoor and outdoor environments [40]. Six synchronised cameras acquired data from 3 different angles, 2 focal lengths per angle, in 3 different sessions (3-months span between each pair of sessions). FiA video-frames present a considerable high quality, specifically in terms of resolution, since they were captured in a relatively controlled scenario. This dataset has been used to assess other adaptive methods, like the one in [68]. In our experiments, we have used the videos provided by the smaller focal length of the frontal camera, both indoor and outdoor, and only considered the 70 identities present in all sessions.

Since this dataset is only used to illustrate the current state-of-the-art performance in datasets where adaptation was required in the past [68], the experiment's results are briefly stated here. Following the same protocol as the one described for COX Face Database for the case of 35 IoI in a universe of 70 identities, OSDe-SVM can achieve +92% in F1-score without adaptation (Fig. 4.5). This performance widely surpasses the one presented in [68]. This behaviour can be attributed to the higher frame quality, which better matches the general context. To emulate more realistic conditions, video frames are down-sampled before entering the feature encoder. In Fig. 4.5 performance results for 5 different down-scaling ratios are shown. Without having the possibility of averaging performance under different universes, we decided to randomly draw 20 different sets of 35 IoI for average and deviation computations. We measure OSDe-SVM performance before (corresponding to the raw performance of the network) and after adaptation. Results in Fig. 4.5 show the performance degradation as the resolution decrease, which OSDe-SVM alleviates with its unsupervised adaptation. Considering that a 1/16 downscale gives face crops of size 7x7 (for an original size of 112x112); such low resolutions are not too realistic and make identification almost impossible.

4.4.2. Experimental Set-up

The experimental setup tries to simulate the stream data scenario of V2V-FR. First, initial models (one-classifier ensembles) of the IoIs are created. This classifier is created using samples from the *initially labelled sequences* (S_0^k): 5 frames of the actual identity as a positive set and 100 frames from other IoI as negative set (randomly drawn for each classifier). The size of the negative set is maintained for future classifier additions to have the same balance in each of the ensemble's classifiers. After this initialisation process, unlabelled sequences repeatedly query the system. Since we are working in an open-set scenario, these input sequences can belong to one of the IoI or not.

Experiments are organised in *adaptation steps*, after which performance is measured. An *adaptation step* corresponds to either the initialisation, a complete iteration over the k available identities with the same t, or a process of self-healing (See Fig 4.6). Additionally, we fully iterate over $t = \{1, 2, ..., 8\}$ a total of **3** times (laps), always preserving the temporal order. This way, we can increase the number of possible updates and study the system's behaviour with redundant data of both IoI and *unknowns*. Self-healing was performed at *adaptation steps* multiples of **5**, and the maximum number of classifiers per ensemble, M, was fixed to **10**. This gives us a total of **31** *adaptation steps* per experiment. Alg. 3 outlines the whole procedure.

Both the size of the identity universe and the number of IoIs vary with the experiment. For universe sizes smaller than 1000, the experiment is repeated for different splits of identities (following Alg. 4) to compute an average performance. A partial overlapping between splits was considered to get a more comprehensive sampling. For the case of 1000 identities, we repeat the experiment **5** times to

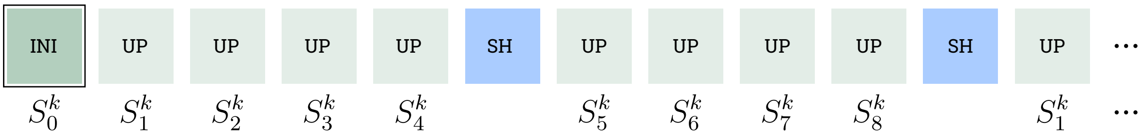


Figure 4.6: Adaptation steps performed during the experiments. INI stands for initialisation, UP for update and SH for self-healing. The last step corresponds to the beginning of the second lap.

Algorithm 3 Experimental procedure and testing protocol.

```
$S_t^k$ is the sequence $t$ of the identity $k$, $L$ is the number of laps
$f$ number of different sub-sequences per identity
$N$ = number of IoI, $N_U$ = number identities in the universe
for each split do
    for $k = 0$ to $N - 1$ do
        Initialise ensemble $k$ using $S_0^k$
    end for
    Perform testing using the set of $S_f^{k=\{0,1,\ldots,N-1\}}$
    for $lap = 0$ to $L - 1$ do
        for $t = 1$ to $f - 1$ do
            for $k = 0$ to $N_U - 1$ do
                Perform adaptation using $S_t^k$.
            end for
            Perform testing using the set of $S_f^{k=\{0,1,\ldots,N-1\}}$
        end for
    end for
end for
```

address the variations provoked by the random set of negatives. For example, for the case of a universe with 100 identities, we would have a total of 19 different splits. As for metrics, we measure precision, recall and F1-measure, using a T_W fixed to 0.01.

4.5. Experiments and Results

The experimental part of this chapter begins with a study of the dependence of performance against the size of the universe while maintaining the ratio with respect to the number of IoI constant (Sec. 4.5.1). After that, the temporal evolution is

Algorithm 4 Algorithm to create the splits

```
$N_U$ is the number of identities in each experiment universe
$N_D$ is the number of identities in the dataset
i = 0
**while** $i + N_U < N_D$ **do**
    *splits* $\leftarrow$ Samples with ID $\in \left[\frac{i}{2}, \frac{i}{2} + N_U\right]$
    $i{+}= N_U$
**end while**
```

further explained in a more comprehensive analysis (Sec. 4.5.2). Then, OSDe-SVM is compared against other state-of-the-art recognition methods (Sec. 4.5.3). Finally, the effect of openness is assessed (Sec. 4.5.4).

4.5.1. Performance vs. Universe Size

This experiment shows the performance behaviour of OSDe-SVM under different universe sizes (N_U) while keeping openness in a ≈18% (see Sec. 4.5.4 for further details), which corresponds to the case of having 1 IoI out of 2 identities in the universe. Results are shown in Fig. 4.7 and Tab. 4.1. We measure initial (non-adaptation) and final (after adaptation) performance of OSDe-SVM, using the previously described experimental set-up (Sec. 4.4.2). It is important to remark that non-adaptation means that ensembles do not incorporate new SVMs apart from the initial one. Thus, performance is quite similar to the one provided by the original network [26].

From the experimental results, the benefits provided by the adaptive nature of the OSDe-SVM are patent. F1-scores increase in all cases (2-20% improvement), mainly due to the impact on recall (9-30% improvement). OSDe-SVM helps to enhance and enrich the existent face models, being able to recognise what previously were unrecognisable. This improvement is even more remarkable, accounting for the challenging experimental conditions. First, only 5 low-quality frames are provided with true labels to create the initial models. After that, no additional labelling is provided. Second, we use the same identities (both *known* and *unknown*) to perform the queries in each adaptation step. Therefore, confusions between identities could reinforce the impostor and eventually provoke a complete identity theft.

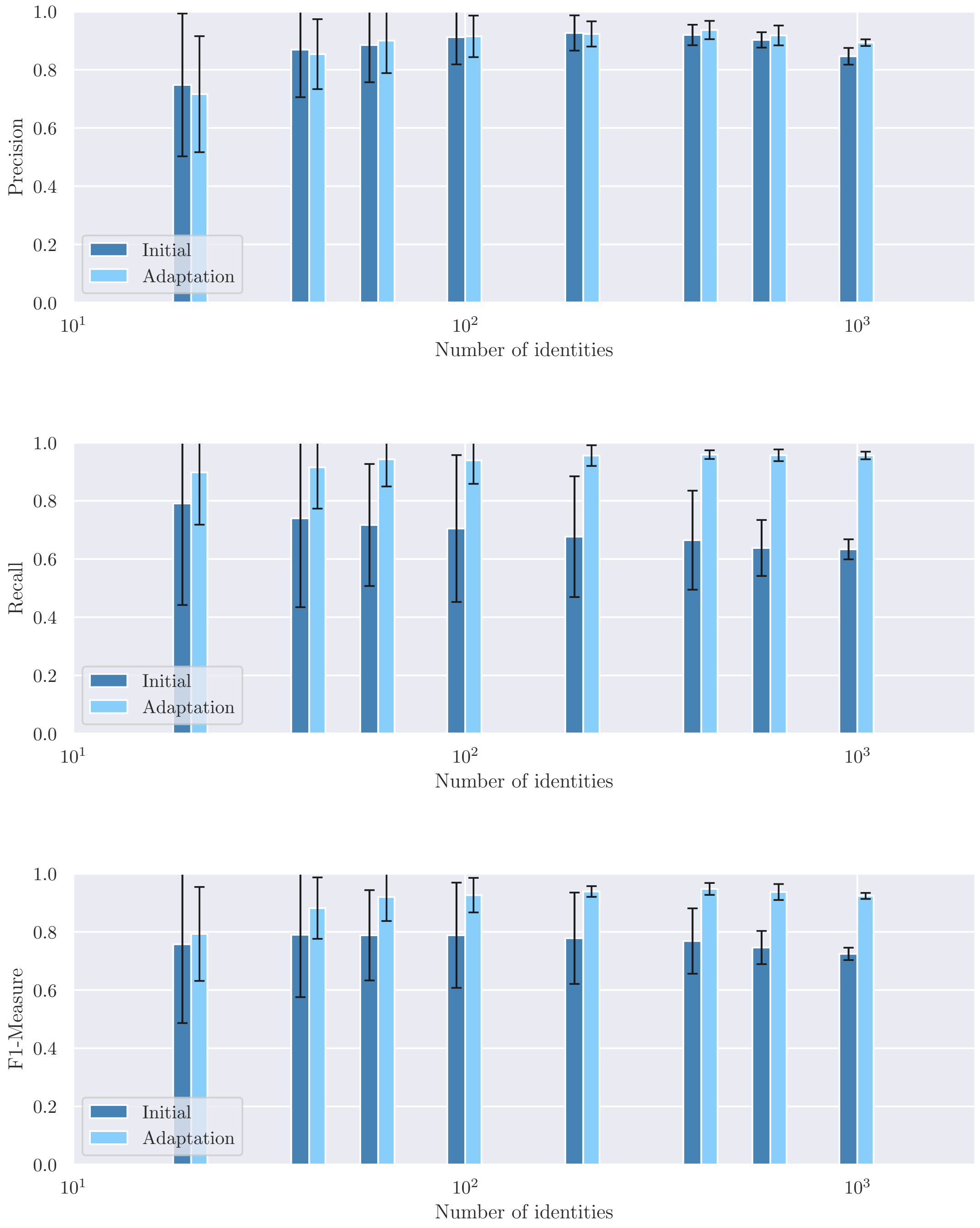


Figure 4.7: Performance under different universe sizes (20, 40, 60, 100, 200, 400, 600, and 1000) and same ratio between number of IoI and universe size (1:2).

Table 4.1: Performance over different universe sizes, while preserving the ratio with the number of IoI. Values are expressed as $\mu(\sigma)$, where μ stands for mean and σ for standard deviation.

		Precision		Recall		F1	
N	N_U	Initial	Final	Initial	Final	Initial	Final
10	20	75 (12)	71 (12)	79 (17)	88 (11)	76 (14)	78.5 (9.7)
20	40	86.9 (8.2)	85.1 (6.3)	74 (15)	91.1 (6.7)	79 (11)	87.8 (5.4)
30	60	88.5 (6.4)	89.7 (5.5)	72 (10)	94.3 (3.7)	78.8 (7.8)	91.8 (3.7)
50	100	91.2 (4.7)	91.9 (3.8)	70 (13)	94.2 (4.1)	78.8 (9.0)	93.0 (3.2)
100	200	92.6 (3.0)	92.6 (2.1)	68 (10)	95.1 (1.9)	77.8 (7.8)	93.79 (0.97)
200	400	92.0 (1.8)	93.5 (1.6)	66.5 (8.5)	95.7 (1.0)	76.8 (5.6)	94.6 (1.1)
300	600	90.3 (1.3)	91.9 (1.3)	63.8 (4.8)	95.6 (1.0)	74.6 (2.9)	93.8 (1.1)
500	1000	84.6 (1.4)	89.33 (0.76)	63.3 (1.7)	95.33 (0.59)	72.4 (1.1)	92.23 (0.64)

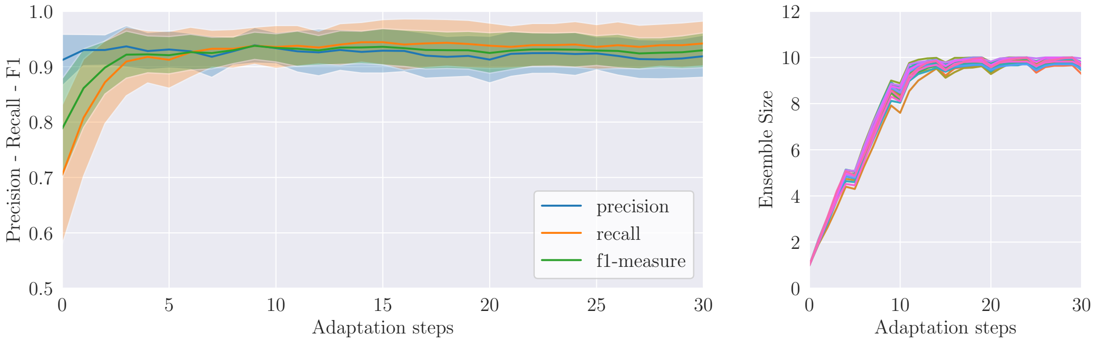
Although overall, the behaviour observed is stable, the greatest improvement in performance corresponds to larger universes. This behaviour can be explained by how we use the EVT. The quality of the Weibull fit in *RDF* (Section 4.3.1) increases as the number of samples to fit do so. For instance, since just half of the data is used in this process (those greater than the median, L8 in Alg. 2), when the IoI is 10 the Weibull fit is done with only 5 points.

4.5.2. Temporal Evolution

This experiment was aimed at performing a detailed study of the temporal evolution of the OSDe-SVM performance for one of the previous cases (50 IoI in a universe of 100). Results are shown in Fig. 4.8.

The first thing we can extract from the experiments (Fig. 4.8a) is that the performance's improvement is higher in the first steps. This is something which could be expected as adding individual classifiers has a higher impact when the size of the ensemble is lower. Besides, this behaviour shows the system's robustness against repeated unknown queries.

These figures also allow observing in a more detailed manner the remarkable recall improvement provided by OSDe-SVM. Precision is also improved but to a lesser extent. Besides, Fig. 4.8b shows the evolution of the average ensemble size for each of the splits (Alg. 4). We can see the effect of self-healing (every 5 steps)



(a) Performance evolution.

(b) Ensemble size evolution.

Figure 4.8: Evolution of OSDe-SVM for the case of 50 IoI over an universe of 100 identities.

and the limitation module. First, drops in size correspond to the triggering of the self-healing process. Second, the size of each ensemble, M^k, is effectively restricted by the limitation module to 10 SVM classifiers.

4.5.3. Comparison against state-of-the-art face recognition models.

Here, we compare the performance of OSDeSVM against two other well-known methods for face recognition (Tab. 4.2). In these two methods, the focus was on obtaining the most widely separated classes in feature space to make the classification as easy as possible. On the one hand, FaceNet [121] feature embedding is designed to distinguish faces by computing the euclidean distance between two features (99.6% accuracy on LFW). On the other hand, ArcFace [26] embeddings are designed to distinguish features by using cosine similarity. All of this makes them

Table 4.2: Comparison against state-of-the-art face recognition models: FaceNet [121] and RN100-AF (ArcFace) [26], (for the case of 50 IoI in an universe of 100).

Method	Precision	Recall	F1-measure
FaceNet +Euclidean+TH	38.7 (7.4)	71.3 (9.4)	49.0 (8.7)
RN100-AF +Cosine+TH	77.3 (9.9)	86 (11)	80.7 (6.5)
RN100-AF +OSDe-SVM, Initial	91.2 (4.7)	70 (13)	77.8 (7.8)
RN100-AF +OSDe-SVM, After Adapt.	91.9 (3.8)	94.2 (4.1)	93.0 (3.2)

suitable for application in any face related task (either verification, identification or general recognition) or, as in our case, to use as a basis for the development of an adaptive method.

Both euclidean distance and cosine similarity are used to compare two single features. Since here we work with the features of all the frames in each query sequence, the centre of this cluster of features is computed as proposed in the original paper [26], to obtain a unique feature per sequence. Besides, the thresholds were tuned offline to get the best F1 scores, which are used as baselines. This would be impossible to do in-stream learning conditions.

Results on Tab. 4.2 allow us to gain insights into the issues addressed in this paper. First, the performance of FaceNet shows the difficult endeavour of transitioning to real-world problems (low-quality, open-set considerations, etc.). Second, our initialisation of OSDe-SVM with RN100-AF embeddings preserves most of the discrimination power of the original decision function (cosine similarity). Finally, the enhanced performance provided by OSDe-SVM is put into perspective against other state-of-art static face recognition models. This improvement translates into a 15% higher F1-score.

4.5.4. Performance vs. Openness

The goal of this experiment is to study how the behaviour of OSDe-SVM changes with the *openness* ratio (4.4), that is the ratio of *known* to *unknown* identities [119]. This measure goes from 0% openness (closed-set recognition) to, theoretically, 100%:

$$\text{openness} = 1 - \sqrt{2 \cdot \frac{N_{\text{training}}}{N_{\text{target}} + N_{\text{testing}}}}, \tag{4.3}$$

where N_{training} is the number of identities used on training (in our case, N), N_{target} is the number of identities to recognise (in our case, N as well) and N_{testing} are the number of identities used on testing (in our case, N_U). Thus, Eq. 4.3 simplifies to:

$$\text{openness} = 1 - \sqrt{2 \cdot \frac{N}{N + N_U}}. \tag{4.4}$$

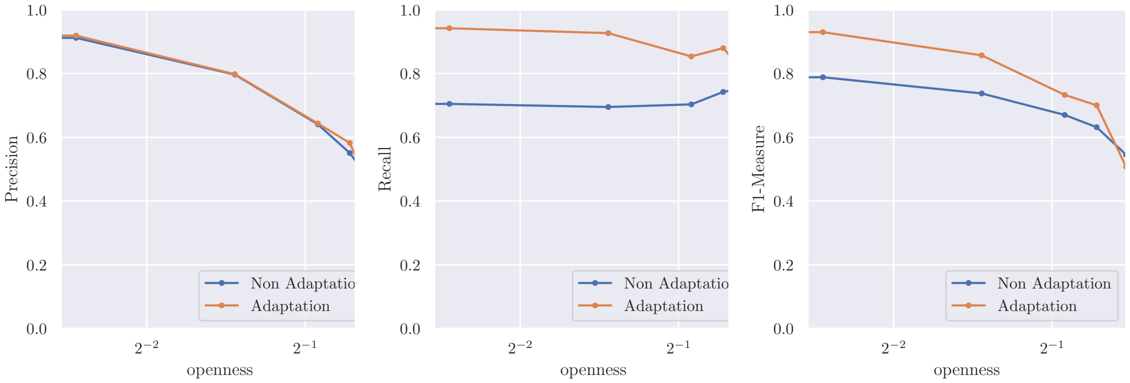


Figure 4.9: Performance openness dependence with fixed IoI (50), and universe ∈ {50, 100, 200, 400, 600, 1000}.

To have a wide range of openness values, we selected a relatively low number of IoI (50) to then vary the size of the universe from 50 identities (0% *openness*, i.e. closed-set) to 1000 identities (≈70%). Fig. 4.9 shows the experimental results, where performance is represented in terms of precision, recall and F1-scores.

The performance graphs show a clear decay of F1 performance as *openness* increases because of the loss of precision. It must be noted that openness affects both the unsupervised adaptation and the testing process. An increase in openness provokes a decay in precision, making more mistakes during the self-adaptation. Accordingly, the drop in precision leads to a decay in recall after the adaptation process. Against all odds, the system proves its robustness until almost 60% of *openness*.

4.6. Conclusions

In this chapter, the adaptation ideas of De-SVM are fully deployed into the creation of OSDe-SVM. This way, this instance-incremental learning approach is able to tackle the V2V-FR problem in both closed and open-set face conditions. As De-SVM, the method is able to operate en real-world non-stationary environments with almost no label requirements. Once initialised (using 5 labelled frames per IoI), the proposed method creates and updates an ensemble of SVM classifiers using samples directly taken from the input sequence, which effectively deals with catastrophic forgetting. These updates are performed following the self-training paradigm in which OSDe-SVM predictions are used as pseudo-labels to incorporate new knowledge without additional supervision. Open-set decisions are rooted in EVT, providing for a way of distinguishing between known and unknown identities. Finally, OSDe-SVM also exploits the fact that updates are encapsulated into individual SVM to achieve fully update reversibility.

Experiments are mainly performed on the COX Face Database, one of the most challenging video surveillance database available. Guided by real-world necessities, the experimental set-up simulates open set recognition conditions. Results showed up to a 15% F1-measure (achieving up to a $\approx 94\%$ F1-measure, depending on the amount of IoI to recognise) increase with respect to the closest static state-of-the-art (ResNet100+AF) face recognition model. Furthermore, the proposed system's performance is tested under different *openness*, proving to be reliable up to $+60\%$ *openness* (50 IoI in a universe of 1000 identities), where *unknown* identities appear as many times as IoI.

Chapter 5

Conclusions and future work

This chapter summarises the the main contributions of the Thesis and provides for some insights on future research directions.

Conclusions

In the development of face recognition systems designed to operate in the real world, one of the most influential aspects is the actual context of operation. Recently, there has been a tremendous effort to collect a large amount of labelled data to meet the demands of the most advanced neural networks. However, the task of collecting (and labelling) data in every possible specific context is an endless endeavour. The consequences of this lack of coverage appear when transitioning face recognition systems from general contexts to more specific ones. In this regard, one of the central themes of this Thesis was to prove that, despite the recent advances, there is still a non-negligible performance tax to pay during this transition (with both hand-crafted and deep learning representations), in a phenomenon often referred to as *dataset bias*.

As a more efficient and scalable solution to these challenges, this Thesis has been inspired by one of the core parts of biological intelligence: adaptation. This idea has crystallised into the creation of De-SVM adaptive face verification system and OSDe-SVM, its extension to the more general open-set face recognition problem.

Up to our knowledge, the proposed method represents one of the few methods that can operate with such label limitations (just 5 anotated frames) and in an online manner.

Its architecture uses state-of-the-art deep feature embeddings as a basis (without any requirements in terms of the network type) to then combine the use of ensembles of very-specific SVM classifiers with the self-training paradigm to perform unsupervised incremental learning and so implement these adaptation capabilities. This way, we retain the deep learning discrimination power, while ensembles provide for additional beneficial properties like modularity, scalability, reversibility, generalisation and robustness.

Following the self-training paradigm then, during the system's operation, predictions of the incoming samples also play the role of pseudo-labels to improve the models' discrimination power. This way, the proposed method can perform a self-sufficient online adaptation in specific contexts without requiring a huge amount of data (either labelled or unlabelled) or the storage of the collected samples. The absence of these two requirements is especially interesting for biometry-related applications since they avoid any privacy-related concern.

For experimental purposes, the video-surveillance context was chosen as a paradigmatic example of a specific context in which face recognition is often required. Thus, we have simulated an environment where we aim to recognise some individuals using only a few (5 or less) video-frames (video-to-video). As the performance obtained with this limited data is relatively low, the proposed system needs to unsupervisedly use additional information from the stream to enhance the recognition power.

As aforementioned, De-SVM was the system proposed for the case of face verification. Experiments are conducted using both COX Face Database and YouTube Faces (YTF), showing in both cases an important unsupervised improvement of the initial performance (from 37% to 85% TAR@FAR1% on COX and from 57% to 75% on YTF) in comparison to other incremental learning methods wrapped under the same self-training strategy. Besides, the ensemble architecture of De-SVM has shown remarkable robustness to repeated impostor queries. This robustness also allows setting a less cautious operational threshold to incorporate diverse information without fearing false inclusions as much as with other methods.

For the case of open-set face recognition, the ideas of De-SVM are extended for the creation of OSDe-SVM. Here, the Extreme Value Theory is used to assist in the decision of distinguishing between *known* and *unknown* identities. Besides, since we encapsulate each update into individual classifiers, we provide a method for achieving update reversibility (to correct possible errors and effectively limit the size of the ensembles). In this case, the experiments are restricted to the COX Face Database, obtaining up to a 15% F1-measure improvement (achieving up to a $\approx 94\%$ F1-measure, depending on the number of individuals) and improving the results obtained by other non-adaptive state-of-the-art face recognition networks.

Besides, OSDe-SVM also showed a remarkable performance in terms of *openness* robustness, proven to be reliable in a scenario where the*Individuals of Interest (IoI)* represent just a 5% of total identities querying the system (60% openness, 50 IoI in a universe of 1000 identities). In any case, the performance decays slowly with the openness increase and maintains its performance when the proportion of IoI and *unknowns* is maintained to 1:1 (from 50 to 500 IoI).

Future work

As the performance improvements of fully supervised end-to-end deep learning approaches begin to moderate, the research community turns its attention to other pivotal topics of machine learning, as self-supervised learning or continual learning. This is the frame in which this Thesis makes its contributions. However, as with any incipient topic, there is still plenty of opportunities for additional contributions. Here we state some of the future research lines opened from this work.

First, one of the most relevant handicaps found during the research development was the scarcity of proper datasets mimicking a video surveillance scenario. In this regard, COX Face Database was one of the best in its class. Notwithstanding, despite containing a relatively high number of identities, the available video sequences are limited to just 3 per individual (all of them acquired in the same day). This fact limits the presence of visual changes over time and the possibility to perform adaptations on a greater time scale, indispensable conditions to *lifelong learning.* For all of these reasons, an important future contribution could be the development of a testing dataset with the same visual quality (and challenges) as COX Face database samples and increasing both the number of sequences available per identity and the time scale of the data acquisition.

Second, using the self-training approach in combination with ensembles and deep feature embeddings has proven to be an interesting and successful approach to implement unsupervised incremental learning. Nevertheless, even maintaining these core ideas, there is still further research to perform. For instance, one of the more challenging parts of working with ensembles in unsupervised conditions is the definition of intelligent rules to eliminate the ensemble's classifiers or weigh its decisions. Intuitively, either wrong update classifiers or too redundant classifiers can be harmful to the ensemble. The last chapter proposes two modules, self-healing and the limiting module, to respectively tackle these two issues. Presented as a first proposition, we believe that this problem can require extensive research to find an optimal solution. In this same frame, another interesting line to follow could be to find a way of taking profit of negative responses of the ensemble to enhance its decisions instead of just being discarded. And, finally, another interesting follow up of the self-training research line would be to extend its capabilities to the class-incremental problem.

And third, despite being the core of our experimental procedure, we do not need to forget that face recognition in video surveillance is just considered a proof-of-concept task. We believe that this Thesis's solutions could be easily translated to other video-related contexts as object detection from mobile robots, person Re-ID or other detection applications. In this regard, applications that allow extracting multi-modal data could be especially interesting since they could assist with the pseudo-labelling of the self-training (in a more co-training fashion).

Bibliography

[1] M. Acharya, T. Hayes, and C. Kanan. Rodeo: Replay for online object detection. In *British Machine Vision Conference (BMVC)*, 2020.

[2] S. Banerjee, J. Brogan, J. Krizaj, et al. To frontalize or not to frontalize: Do we really need elaborate pre-processing to improve face recognition? In *Winter Conference on Applications of Computer Vision (WACV)*, pages 20–29, March 2018.

[3] S. Bashbaghi, E. Granger, R. Sabourin, and G.-A. Bilodeau. Dynamic ensembles of exemplar-SVMs for still-to-video face recognition. *Pattern Recognition*, 69:61 – 81, 2017.

[4] S. Bashbaghi, E. Granger, R. Sabourin, and M. Parchami. *Deep Learning Architectures for Face Recognition in Video Surveillance*, pages 133–154. Springer Singapore, Singapore, 2019.

[5] F. Becattini, L. Seidenari, and A. Del Bimbo. Indexing quantized ensembles of exemplar-SVMs with rejecting taxonomies. *Multimedia Tools and Applications*, 76(21):22647–22668, Nov 2017.

[6] S. Becker. Implicit learning in 3d object recognition: The importance of temporal context. *Neural Computation*, 11(2):347–374, 1999.

[7] E. Belouadah and A. Popescu. Scail: Classifier weights scaling for class incremental learning. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1255–1264, March 2020.

[8] A. Bendale and T. E. Boult. Towards open set deep networks. In *2016 IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 1563–1572, June 2016.

[9] S. Bianco. Large age-gap face verification by feature injection in deep networks. *Pattern Recognition Letters*, 90:36 – 42, 2017.

[10] T. E. Boult, S. Cruz, A. R. Dhamija, M. Gunther, J. Henrydoss, and W. J. Scheirer. Learning and the unknown: Surveying steps toward open world recognition. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 9801–9807, 2019.

[11] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, volume 81, pages 77–91. PMLR, 2018.

[12] F. Casado, C. Regueiro, R. Iglesias, X. Pardo, and E. López. Automatic selection of user samples for a non-collaborative face verification system. In *ROBOT 2017: Third Iberian Robotics Conference*, pages 555–566. Springer, 2018.

[13] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari. End-to-end incremental learning. In *European Conference on Computer Vision (ECCV)*, pages 241–257, 2018.

[14] A. Chefrour. Incremental supervised learning: algorithms and applications in pattern recognition. *Evolutionary Intelligence*, 12(2):97–112, Jun 2019.

[15] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision – ECCV 2012*, pages 566–579, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[16] D. Chen, X. Cao, D. Wipf, F. Wen, and J. Sun. An efficient joint formulation for bayesian face verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):32–46, Jan 2017.

[17] X. Chen, C. Wang, B. Xiao, and X. Cai. Scenario oriented discriminant analysis for still-to-video face recognition. In *IEEE International Conference on Image Processing (ICIP)*, pages 738–742, 2014.

[18] S. Coles. *Classical Extreme Value Theory and Models*, pages 45–73. Springer London, London, 2001.

[19] R. Coop, A. Mishtal, and I. Arel. Ensemble learning in fixed expansion layer networks for mitigating catastrophic forgetting. *IEEE Transactions on Neural Networks and Learning Systems*, 24(10):1623–1634, Oct 2013.

[20] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[21] D. Crispell, O. Biris, N. Crosswhite, J. Byrne, and J. Mundy. Dataset augmentation for pose and lighting invariant face recognition. *CoRR*, abs/1704.04326, 2017.

[22] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman. Template adaptation for face verification and identification. *Image and Vision Computing*, 79:35 – 48, 2018.

[23] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4109–4118, June 2018.

[24] E. Davies. Chapter 1 - vision, the challenge. In E. Davies, editor, *Computer Vision (Fifth Edition)*, pages 1–15. Academic Press, fifth edition edition, 2018.

[25] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[26] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, June 2019.

[27] M. A. A. Dewan, E. Granger, G.-L. Marcialis, R. Sabourin, and F. Roli. Adaptive appearance model tracking for still-to-video face recognition. *Pattern Recognition*, 49:129 – 151, 2016.

[28] T. I. Dhamecha, A. Noore, R. Singh, and M. Vatsa. Between-subclass piecewise linear solutions in large scale kernel svm learning. *Pattern Recognition*, 95:173 – 190, 2019.

[29] L. Didaci, G. L. Marcialis, and F. Roli. Analysis of unsupervised template update in biometric recognition systems. *Pattern Recognition Letters*, 37:151 – 160, 2014.

[30] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar. Learning in nonstationary environments: A survey. *IEEE Computational Intelligence Magazine*, 10(4):12–25, Nov 2015.

[31] N. Dvornik, J. Mairal, and C. Schmid. Diversity with cooperation: Ensemble methods for few-shot classification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3722–3730, Oct 2019.

[32] A. Fahad, A. Almalawi, Z. Tari, K. Alharthi, F. S. A. Qahtani, and M. Cheriet. Semtra: A semi-supervised approach to traffic flow labeling with minimal human effort. *Pattern Recognition*, 91:1 – 12, 2019.

[33] Fayin Li and H. Wechsler. Open set face recognition using transduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1686–1697, Nov 2005.

[34] Fayin Li and H. Wechsler. Open set face recognition using transduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1686–1697, Nov 2005.

[35] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *International Conference on Computer Vision (ICCV)*, pages 2960–2967, 2013.

[36] A. Franco, D. Maio, and D. Maltoni. Incremental template updating for face recognition in home environments. *Pattern Recognition*, 43(8):2891 – 2903, 2010.

[37] Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi. Generative openmax for multiclass open set classification. In *British Machine Vision Conference Proceedings (BMVC)*, 2017.

[38] C. Geng, S. Huang, and S. Chen. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[39] A. Gepperth and B. Hammer. Incremental learning algorithms and applications. In *European Symposium on Artificial Neural Networks (ESANN)*, pages 27–29, Bruges, Belgium, 2016.

[40] R. Goh, L. Liu, X. Liu, and T. Chen. The CMU Face In Action (FIA) Database. In W. Zhao, S. Gong, and X. Tang, editors, *Analysis and Modelling of Faces and Gestures*, pages 255–263. Springer, 2005.

[41] H. M. Gomes, J. P. Barddal, F. Enembreck, and A. Bifet. A survey on ensemble learning for data stream classification. *ACM Comput. Surv.*, 50, 2017.

[42] H. M. Gomes, J. P. Barddal, F. Enembreck, and A. Bifet. A survey on ensemble learning for data stream classification. *ACM Comput. Surv.*, 50(2), 2017.

[43] G. Guo and N. Zhang. What is the challenge for deep learning in unconstrained face recognition? In *IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 436–442, May 2018.

[44] G. Guo and N. Zhang. A survey on deep learning based face recognition. *Computer Vision and Image Understanding*, 189:102805, 2019.

[45] Y. Guo, X. Wang, P. Xiao, and X. Xu. An ensemble learning framework for convolutional neural network based on multiple classifiers. *Soft Computing*, 24, 06 2020.

[46] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In *European Conference on Computer Vision (ECCV)*, pages 87–102, 2016.

[47] M. Günther, S. Cruz, E. M. Rudd, and T. E. Boult. Toward open-set face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 573–582, July 2017.

[48] T. Hassner. Viewing real-world faces in 3d. In *International Conference on Computer Vision (ICCV)*, pages 3607–3614, Dec 2013.

[49] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4295–4304, June 2015.

[50] T. L. Hayes and C. Kanan. Lifelong machine learning with deep streaming linear discriminant analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 887–896, June 2020.

[51] J. He, R. Mao, Z. Shao, and F. Zhu. Incremental learning in online scenario. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13923–13932, 2020.

[52] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[53] G. Hu, X. Peng, Y. Yang, T. M. Hospedales, and J. Verbeek. Frankenstein: Learning deep face representations using small data. *IEEE Transactions on Image Processing*, 27(1):293–303, Jan 2018.

[54] P. Hu and D. Ramanan. Finding tiny faces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1522–1530, July 2017.

[55] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[56] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, and X. Chen. A benchmark and comparative study of video-based face recognition on cox face database. *IEEE Transactions on Image Processing*, 24(12):5967–5981, Dec 2015.

[57] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1867–1874, 2014.

[58] R. Kemker, A. Abitino, M. McClure, and C. Kanan. Measuring catastrophic forgetting in neural networks. In *AAAI*, 2018.

[59] A. Khosla, T. Zhou, T. Malisiewicz, A. Efros, and A. Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision (ECCV)*, pages 158–171. Springer, 2012.

[60] S. Kim, J. Choi, T. Kim, and C. Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[61] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[62] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[63] J. Kivinen, A. J. Smola, and R. C. Williamson. Online learning with kernels. *IEEE Transactions on Signal Processing*, 52(8):2165–2176, 2004.

[64] B. F. Klare, B. Klein, E. Taborsky, et al. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939, 2015.

[65] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter. Empirically analyzing the effect of dataset biases on deep face recognition systems. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2174–217409, June 2018.

[66] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak. Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37:132 – 156, 2017.

[67] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017.

[68] M. D. la Torre, E. Granger, P. V. Radtke, R. Sabourin, and D. O. Gorodnichy. Partially-supervised learning from facial trajectories for face recognition in video surveillance. *Information Fusion*, 24:31 – 53, 2015.

[69] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[70] Z. Lei, M. Pietikäinen, and S. Z. Li. Learning discriminant face descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):289–302, Feb 2014.

[71] N. Li, Y. Yu, and Z.-H. Zhou. Diversity regularized ensemble pruning. In *Machine Learning and Knowledge Discovery in Databases*, pages 330–345, 2012.

[72] Y. Li, Y. Wang, Q. Liu, C. Bi, X. Jiang, and S. Sun. Incremental semi-supervised learning on streaming data. *Pattern Recognition*, 88:383 – 396, 2019.

[73] Y. Li, F. Yang, Y. Liu, Y. Yeh, X. Du, and Y. F. Wang. Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 285–2856, June 2018.

[74] Z. Li and D. Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, Dec 2018.

[75] N. Liang, G. Huang, P. Saratchandran, and N. Sundararajan. A fast and accurate online sequential learning algorithm for feedforward networks. *IEEE Transactions on Neural Networks*, 17(6):1411–1423, 2006.

[76] H. Liu, X. Zhu, Z. Lei, and S. Z. Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11939–11948, June 2019.

[77] V. Lomonaco and D. Maltoni. Comparing incremental learning strategies for convolutional neural networks. In F. Schwenker, H. M. Abbas, N. El Gayar, and E. Trentin, editors, *Artificial Neural Networks in Pattern Recognition*, pages 175–184, 2016.

[78] E. Lopez-Lopez, C. V. Regueiro, and X. M. Pardo. Incremental learning from low-labelled stream data in open-set video face recognition, 2020.

[79] E. Lopez-Lopez, C. V. Regueiro, and X. M. Pardo. An adaptive video-to-video face identification system based on self-training. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2590–2596, 2021.

[80] E. Lopez-Lopez, C. V. Regueiro, X. M. Pardo, A. Franco, and A. Lumini. Towards a self-sufficient face verification system. *Expert Systems with Applications*, 174:114734, 2021.

[81] E. Lopez-Lopez, C. V. Regueiro, X. M. Pardo, A. Franco, and A. Lumini. Incremental learning techniques within a self-updating approach for face verification in video-surveillance. In *Pattern Recognition and Image Analysis*, pages 25–37. Springer International Publishing, 2019.

[82] Z. Luo, J. Hu, W. Deng, and H. Shen. Deep unsupervised domain adaptation for face recognition. In *International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 453–457, May 2018.

[83] E. López-López, X. M. Pardo, C. V. Regueiro, R. Iglesias, and F. E. Casado. Dataset bias exposed in face verification. *IET Biometrics*, 8(4):249–258, 2019.

[84] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[85] U. Mahbub, S. Sarkar, V. M. Patel, and R. Chellappa. Active user authentication for smartphones: A challenge data set and benchmark results. In *International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, Sept 2016.

[86] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-SVMs for object detection and beyond. In *IEEE International Conference on Computer Vision (ICCV)*, pages 89–96, Nov 2011.

[87] D. Maltoni and V. Lomonaco. Continuous learning in single-incremental-task scenarios. *Neural Networks*, 116:56 – 73, 2019.

[88] S. Mathe, A. Pirinen, and C. Sminchisescu. Reinforcement learning for visual object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2894–2902, 2016.

[89] I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *European Conference On Computer Vision (ECCV)*, pages 527–544, 2016.

[90] N. C. Mithun, R. Panda, and A. K. Roy-Chowdhury. Generating diverse image datasets with limited labeling. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pages 566–570, New York, NY, USA, 2016. ACM.

[91] H. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *International Conference on Image Processing (ICIP)*, pages 343–347, Oct 2014.

[92] H. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 343–347, Oct 2014.

[93] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[94] G. Orrù, G. L. Marcialis, and F. Roli. A novel classification-selection approach for the self updating of template-based face recognition systems. *Pattern Recognition*, 100:107–121, 2020.

[95] P. Oza and V. M. Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2302–2311, June 2019.

[96] S. A. Papert. The summer vision project. *MIT. Vision Memo.*, 1966.

[97] M. Parchami, S. Bashbaghi, and E. Granger. Video-based face recognition using ensemble of haar-like deep convolutional neural networks. In *International Joint Conference on Neural Networks (IJCNN)*, pages 4625–4632, 2017.

[98] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54 – 71, 2019.

[99] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015.

[100] P. Perera, V. I. Morariu, R. Jain, V. Manjunatha, C. Wigington, V. Ordonez, and V. M. Patel. Generative-discriminative feature representations for open-set recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11811–11820, 2020.

[101] F. Pernici, F. Bartoli, M. Bruni, and A. Del Bimbo. Memory based online learning of deep representations from video streams. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2324–2334, 2018.

[102] F. Pernici and A. D. Bimbo. Unsupervised incremental learning of deep descriptors from video streams. In *IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 477–482, July 2017.

[103] P. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.

[104] P. H. Pisani, A. Mhenni, R. Giot, E. Cherrier, N. Poh, A. C. P. d. L. Ferreira de Carvalho, C. Rosenberger, and N. E. B. Amara. Adaptive biometric systems: Review and perspectives. *ACM Comput. Surv.*, 52(5):102:1–102:38, Sept. 2019.

[105] R. Polikar, L. Upda, S. S. Upda, and V. Honavar. Learn++: an incremental learning algorithm for supervised neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 31(4):497–508, Nov 2001.

[106] J. M. Pérez-Rúa, X. Zhu, T. M. Hospedales, and T. Xiang. Incremental few-shot object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13843–13852, June 2020.

[107] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He. Data distillation: Towards omni-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4119–4128, June 2018.

[108] A. Ratner, H. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Ré. Learning to compose domain-specific transformations for data augmentation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 3236–3246. Curran Associates, Inc., 2017.

[109] A. Rattani, G. L. Marcialis, and F. Roli. Biometric system adaptation by self-update and graph-based techniques. *Journal of Visual Languages & Computing*, 24(1):1 – 9, 2013.

[110] C. Redondo-Cabrera and R. Lopez-Sastre. Unsupervised learning from videos using temporal coherency deep networks. *Computer Vision and Image Understanding*, 179:79 – 89, 2019.

[111] J. Ren, X. Jiang, and J. Yuan. A complete and fully automated face verification system on mobile devices. *Pattern Recognition*, 46(1):45 – 56, 2013.

[112] L. Ren, X. Yuan, J. Lu, M. Yang, and J. Zhou. Deep reinforcement learning with iterative shift for visual tracking. In *European Conference on Computer Vision (ECCV)*, 2018.

[113] A. Roychowdhury, P. Chakrabarty, A. Singh, S. Jin, H. Jiang, L. Cao, and E. Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 780–790, June 2019.

[114] E. M. Rudd, L. P. Jain, W. J. Scheirer, and T. E. Boult. The extreme value machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):762–768, March 2018.

[115] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision (ECCV)*, pages 213–226. Springer, 2010.

[116] D. Sahoo, Q. Pham, J. Lu, and S. C. H. Hoi. Online deep learning: Learning deep neural networks on the fly. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2660–2666, 7 2018.

[117] G. Salomon, A. Britto, R. H. Vareto, W. R. Schwartz, and D. Menotti. Open-set face recognition for small galleries using siamese networks. In *International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 161–166, July 2020.

[118] W. Scheirer, A. Rocha, R. Micheals, and T. Boult. Robust fusion: Extreme value theory for recognition score normalization. In *European Conference on Computer Vision (ECCV)*, pages 481–495, 2010.

[119] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, July 2013.

[120] W. J. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2933–2940, June 2012.

[121] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015.

[122] A. F. Sequeira, J. C. Monteiro, A. Rebelo, and H. P. Oliveira. Mobbio: A multimodal database captured with a portable handheld device. In *Conference on Computer Vision Theory and Applications (VISAPP)*, volume 3, pages 133–139, Jan 2014.

[123] A. Siffer, P.-A. Fouque, A. Termier, and C. Largouet. Anomaly detection in streams with extreme value theory. In *23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1067–1075, 2017.

[124] K. Sohn, S. Liu, G. Zhong, X. Yu, M. Yang, and M. Chandraker. Unsupervised domain adaptation for face recognition in unlabeled videos. In *International Conference on Computer Vision (ICCV)*, pages 5917–5925, Oct 2017.

[125] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 443–450, Cham, 2016. Springer International Publishing.

[126] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1891–1898, June 2014.

[127] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2892–2900, June 2015.

[128] X. Tao, X. Hong, X. Chang, S. Dong, X. Wei, and Y. Gong. Few-shot class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12180–12189, June 2020.

[129] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars. A deeper look at dataset bias. In *Domain Adaptation in Computer Vision Applications*, pages 37–55. Springer, 2017.

[130] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars. *A Deeper Look at Dataset Bias*, pages 37–55. Springer International Publishing, Cham, 2017.

[131] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1521–1528, June 2011.

[132] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, July 2017.

[133] S. Ud Din, J. Shao, J. Kumar, W. Ali, J. Liu, and Y. Ye. Online reliable semi-supervised learning on evolving data streams. *Information Sciences*, 525:153 – 171, 2020.

[134] G. M. van de Ven, H. T. Siegelmann, and A. S. Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, 11(1):4069, Aug 2020.

[135] M. Villamizar, A. Sanfeliu, and F. Moreno-Noguer. Online learning and detection of faces with low human supervision. *The Visual Computer*, 35(3):349–370, 2019.

[136] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–I. IEEE, 2001.

[137] J. Wang, Z. Mo, H. Zhang, and Q. Miao. Ensemble diagnosis method based on transfer learning and incremental learning towards mechanical big data. *Measurement*, 155:107517, 2020.

[138] M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135 – 153, 2018.

[139] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[140] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, December 2015.

[141] G. Wen, H. Chen, D. Cai, and X. He. Improving face recognition with domain adaptation. *Neurocomputing*, 287(C):45–51, 2018.

[142] C. Whitelam, E. Taborsky, A. Blanton, et al. Iarpa janus benchmark-b face dataset. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 592–600, July 2017.

[143] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 529–534, June 2011.

[144] X. Wu, W. Zuo, L. Lin, W. Jia, and D. Zhang. F-svm: Combination of feature transformation and svm learning via convex relaxation. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11):5185–5199, Nov 2018.

[145] H. Xu, J. Zheng, A. Alavi, and R. Chellappa. Cross-domain visual recognition via domain adaptive dictionary learning. *CoRR*, abs/1804.04687, 2018.

[146] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Annual Meeting on Association for Computational Linguistics (ACL)*, pages 189–196, 1995.

[147] J. Zhang, J. Zhang, S. Ghosh, D. Li, S. Tasci, L. Heck, H. Zhang, and C. . Jay Kuo. Class-incremental learning via deep model consolidation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1120–1129, March 2020.

[148] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.

[149] M. Zhang, R. Liu, H. Nada, H. Uchida, T. Matsunami, and N. Abe. A pairwise learning strategy for video-based face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[150] X. Zhang, J. Cao, C. Shen, and M. You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[151] X. Zhang, F. Yan, Y. Zhuang, H. Hu, and C. Bu. Using an ensemble of incrementally fine-tuned cnns for cross-domain object category recognition. *IEEE Access*, 7:33822–33833, 2019.

[152] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, University of Wisconsin-Madison, Dept. Computer Science,, 2005.

[153] Y. Zou, Z. Yu, X. Liu, B. V. Kumar, and J. Wang. Confidence regularized self-training. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

Appendix A

Resumen extendido en galego

Motivación

Os humanos adquiren información do mundo exterior a través de cinco sentidos – vista, oído, gusto, tacto e olfacto – dos cales a vista sobresae como un dos máis importantes. Proba disto son as estimacións que sitúan a cantidade de información recibida polos ollos nunhas dúas ordes de magnitude superiores ao resto de sentidos xuntos. Para que esta cantidade de información sexa abarcable, necesitamos comprimila, clasificala e estruturala. Polo que, o conxunto de mecanismos que forman o que chamamos *visión* non se limitan a meros receptores luminosos (os ollos), senón que involucran toda unha rexión do cerebro chamada o córtex visual. Por se esta non fora unha complexidade suficiente, a visión humana adulta non é unha habilidade completamente estática co paso do tempo. É certo que parte da visión involucra procesos relativamente estables fixados durante o desenvolvemento dos nenos ou pola evolución natural. Non obstante, existe tamén outra parte esencial da visión que é *modulada e adaptada co seu uso activo.*

Nesta tesitura, a visión artificial nace coa tarefa de replicar todo este complexo sistema humano. Aínda sen entender completamente os mecanismos naturais desta habilidade, o enorme potencial en termos de automatización abriuse paso. A investigación da visión artificial involucra un gran rango de expertos interdisciplinarios (dende enxeñeiros, psicólogos, biólogos, ata incluso filósofos, entre outros). Dende o seu nacemento como un proxecto de verán [96] nos 60, os avances foron relativamente graduais ata a última década. No 2012, o uso dun tipo específico de redes neuronais [67] (as redes neuronais de convolución ou CNN) en conxunción coas cada vez máis potentes GPUs e a dispoñibilidade de conxuntos de datos etiquetados cada vez de maior escala [25] provocaron un salto no rendemento moi notorio. De repente e grazas a estas novas técnicas, algunhas das tarefas máis complicadas da visión artificial (e.g. asistencia á condución, recuperación de imaxes, filtros de cambio de cara, etc.) comezaban a ser accesibles incluso a nivel comercial.

Retomando a perspectiva biolóxica, as CNNs poden asociarse con esa parte máis estática do sistema visual humano. Porén, na actualidade, aínda non proporcionan a capacidade de *adaptarse co uso activo*. Ao intentar adaptar unha rede neuronal pre-adestrada usando unha pequena mostra de datos, esta tende a esquecer as súas capacidades iniciais nun problema denominado esquecemento catastrófico (*catastrophical forgetting*). Este comportamento é problemático tanto dende unha perspectiva máis teórica, a capacidade de adaptación debería ser algo irrenunciable nun sistema verdadeiramente intelixente; como dende unha perspectiva máis práctica, que imos analizar a continuación.

O poder das CNN está intimamente relacionado coa calidade e canto de completo é o conxunto de datos etiquetados utilizados durante o seu adestramento. Neste sentido, incluso o máis exhaustivo dos conxuntos de datos non pode garantir un bo rendemento en cada un dos contextos específicos posibles. Mentres que unha solución inmediata a este problema sería a recolección (e etiquetado) de datos adicionais nese contexto específico, a verdade é que existen casos nos que esta acción é particularmente difícil (debido a temas de privacidade, limitación de recursos, problemas de custos, etc.). Polo tanto, a dotación dos sistemas con capacidades de adaptación emerxe como unha solución práctica e eficiente para abordar este problema. Na literatura, esta capacidade de adaptación adóitase a implementar seguindo varias liñas de investigación inter-relacionadas: *transfer learning*, *lifelong learning*, *domain adaptation*, *continual learning* or *incremental learning* [38].

Os retos desta tese

Esta tese apunta cas súas contribucións precisamente na liña de desenvolver unha forma de implementar a capacidade de adaptación na visión artificial. Debido á gran variedade de aplicacións posibles e como proba de concepto, o campo de estudo desta tese restrinxiuse a tarefa do recoñecemento facial.

O recoñecemento facial é unha tarefa da visión artificial que recibiu unha atención significativa por parte da comunidade científica de forma persistente. A grandes trazos, consiste na asignación de identidades usando imaxes de facianas. Na práctica, o certo é que o recoñecemento facial inclúe varios sub-problemas (e.g. a verificación de caras ou a identificación de caras) así como as ferramentas auxiliares involucradas na tarefa (e.g. detección de caras, seguimento de caras ou representación de características). Nesta liña, é importante tamén remarcar que o recoñecemento de caras, dado que trata de distinguir entre obxectos (caras) moi similares entre si, é un campo no que o contexto de adquisición particularmente relevante. Intuitivamente, as diferenzas na aparencia dunha persoa nunha foto de carné respecto á que pode presentar nun contexto de vídeo-vixilancia, poden ser incluso máis notorias que as que as distinguen doutras persoas. Estamos, polo tanto, ante un escenario no que a capacidade de adaptación pode ser especialmente beneficiosa.

Neste marco entón, pasamos a enumerar os principais retos aos que trata de dar resposta esta tese:

O **primeiro reto** consiste en definir conceptualmente como imos reter o poder de discriminación das CNN evitando o *esquecemento catastrófico* ao mesmo tempo. Nesta dirección, optamos por fixar as capas convolucionais (sen a última capa) de redes neuronais do estado da arte. Desta maneira, retomamos a separación clásica entre as partes de extracción de características e clasificación; para así implementar a adaptación mellorando esta última.

O **segundo reto** fai referencia ao grado de robustez existente a variacións de contexto nos métodos do estado da arte. Polo tanto, o primeiro paso será estudar os efectos no rendemento de mesturar mostras de dous contextos distintos para o adestramento dun clasificador. Dende o punto de vista do recoñecemento facial, intentamos entender o impacto de usar conxuntos de datos auxiliares (normalmente tomados en contextos diferentes) ao despregar un sistema de verificación de caras nun contexto específico no que a dispoñibilidade de datos etiquetados e escasa (i.e. biometría en móbiles, vídeo-vixilancia, etc.).

Nesta liña, realizamos un estudo exhaustivo do *nesgo do conxunto de datos* usando algunhas das bases de datos de caras máis coñecidas e facendo un énfase especial nos nesgos derivados do contexto de adquisición.

O **terceiro reto** comprende o propio de deseño do clasificador para implementar a adaptación. A proposta consiste nun sistema de aprendizaxe incremental baseado en comités de clasificadores que se construirá sobre os extractores de características das redes neuronais máis avanzadas. A elección da rede específica que se usará non é relevante, e irá variando ao longo da tese segundo o estado da arte avance. O método é un comité dinámico de SVM (*Dynamic Ensemble of SVM, De-SVM*) que irá engadindo paulatinamente novos clasificadores para adaptarse a contextos específicos e mellorar o rendemento. A literatura científica sempre se referiu aos comités como unha solución eficiente, robusta e escalable para os problemas de clasificación en comparación coas capas completamente conectadas (*fully conected layers*) das CNN. Ademais, ao permitir illar cada actualización en clasificadores individuais, os comités tamén proporcionan reversibilidade da ditas actualizacións asía como unha mellor explicabilidade. Isto é especialmente interesante para operar en contextos nos que existe escaseza de etiquetas, como veremos a continuación.

O **cuarto dos retos**, entón, relaciónase coa ausencia de supervisión ao que diriximos as nosas propostas. Para isto, o método proposto, o *De-SVM*, usa o chamado enfoque de auto-adestramento ou *self-training* no que as predicións do comité en cada momento se usarán como pseudo-etiquetas para engadir ou non novos clasificadores. Desta maneira, o proceso de adestramento intégrase coa operación do sistema sen necesitar que se produza en fases separadas. Gracias este enfoque, despois da inicialización (usando soamente 5 mostras etiquetadas), o sistema será capaz de realizar toda a aprendizaxe incremental de forma non supervisada.

As decisións dentro do comité usan a saída directa de cada SVM para logo fusionalas nunha única saída coa utilización da función mediana. Finalmente, as identidades serán asignadas ou non en base a un *limiar*. A fixación deste *limiar* é un dos aspectos máis desafiantes de todos e fai referencia ao problema do *dilema* *entre estabilidade e plasticidade.*

Toda a parte experimental está centrada no caso específico da verificación de caras vídeo a vídeo nun contexto de vídeo-vixilancia. Este é un claro exemplo de contexto específico no que as necesidades de adaptación poden ser determinantes.A nosa proposta, De-SVM, destaca como o método cun mellor rendemento respecto outras técnicas de aprendizaxe incremental e outros métodos non adaptativos.

O *quinto reto* fai referencia as condicións máis complexas do recoñecemento de cara en conxuntos abertos. Neste sentido, necesitaremos estender o *De-SVM* inicial para que sexa capaz de funcionar nestas condicións nun proceso que estará dividido en dúas partes:

- A adaptación do mecanismo de decisión para o problema de clasificación multiclase de conxunto aberto. Nesta configuración, os sistemas deben estar preparados para recibir tanto mostras de identidades enroladas como mostras adicionais de identidades non coñecidas. Utilizaremos o poder da Teoría dos Valores Extremos (*Extreme Value Theory*) para discriminar entre mostras *coñecidas* como *descoñecidas.*
- A creación de dous módulos adicionais deseñados para eliminar clasificadores do comité, xa sexa por manter o seu tamaño acoutado (módulo de limitación) como para tentar corrixir posibles actualizacións erróneas (módulo de *self-healing*). Desta maneira, aproveitaremos ao máximo o potencial dos enfoques baseados comités.

Contribucións principais

As contribucións principais derivadas desta tese son as seguintes:

- Sobre a extracción de características e as necesidades de adaptación en contornas específicas:
    - Un estudo do impacto do sesgo do conxunto de datos no rendemento da verificación de caras, ao combinar distintas bases de datos de caras durante a fase de adestramento.

  - Un novo enfoque á observación do nesgo do conxunto de datos facendo unha análise da propia distribución das mostras no espazo de características utilizando un perspectiva xeométrica.
  - Unha comparación do comportamento de distintos descritores de características, tanto deseñados manualmente como xerados mediante aprendizaxe, en termos de robustez ante o nesgo do conxunto de datos. Os resultados mostran que incluso os baseados en *deep learning* son susceptibles ao nesgo do conxunto de datos.

- Sobre a verificación de caras en contextos de vídeo-vixilancia:
  - A proposta dun sistema biométrico baseado en comités de clasificadores chamado *Dynamic Ensemble of SVM (De-SVM)*. Este é capaz de reter o poder discriminatorio das CNN mentres que ao mesmo tempo proporciona os numerosos beneficios dos comités relacionados coa modularidade, reversibiliade, xeralización e robustez.
  - O uso do enfoque de auto-adestramento (*self-training*) para usar as propias predicións do clasificador como pseudo-etiquetas para aprender e operar simultaneamente.

- Sobre o recoñecemento de caras en contextos de vídeo-vixilancia:
  - A extensión dos enfoques previstos ao problema do recoñecemento de caras mediante aprendizaxe incremental non supervisado.
  - Unha estratexia para combater con tanto o *esquecemento catastrófico* como o efecto das pseudo-etiquetas incorrectas, aproveitando desta forma o potencial pleno dos comités.
  - Un enfoque á aprendizaxe incremental utilizando mostras individuais que podería ser estendido ao caso de ter un conxunto de clases variable de forma sinxela.
  - Un método para o recoñecemento de caras que non está directamente baseado na recolección e almacenamento de imaxes de caras, e que soamente require 5 fotogramas de vídeo etiquetados para a inicialización.

Conclusións

No desenvolvemento de sistemas deseñados para operar no mundo real, un dos aspectos máis importantes a considerar é o contexto obxectivo. Recentemente, a comunidade científica realizou un enorme esforzo para colleitar grandes conxuntos de datos etiquetados para así poder satisfacer as necesidades das redes neuronais máis avanzadas. Non obstante, debido a gran variabilidade presente no mundo visual o certo é que posuír conxuntos o suficientemente completos de cada contexto específico posible é case algo interminable. As consecuencias desta falta de cobertura aparecen cando trasladamos os sistemas de recoñecemento facial de contextos xerais a outros máis específicos. Nesta liña, un dos temas principais desta tese consistiu en demostrar que, a pesar dos avances producidos recentemente, aínda hai unha non importante peaxe a pagar en termos de rendemento durante esta transición nun fenómeno que se adoita a denominar *nesgo do conxunto de datos*. Esta peaxe foi observable tanto con descritores deseñados manualmente como os baseados en *deep learning*.

Como unha proposta máis eficiente e escalable a estes retos, esta tese inspírase dunha das partes centrais da intelixencia biolóxica: a adaptación. Esta idea cristalizou na creación de *De-SVM*, como un sistema de verificación de caras adaptativo, así como *OSDe-SVM*, a súa extensión ao problema máis xeral do recoñecemento de caras en conxuntos abertos. Ata onde sabemos, esta proposta é un dous poucos métodos que poden operar con tales limitacións na dispoñibilidade de etiquetas (soamente 5 fotogramas etiquetados) e de forma *online*.

A súa arquitectura usa representacións de características baseadas en *deep learning* do estado da arte (sen ningún requirimento especial en canto ao tipo de rede neuronal) para logo combinalas co uso conxunto de comités de clasificadores SVM moi específicos e do enfoque de auto-adestramento (*self-training*) para realizad a aprendizaxe incremental non supervisada e así implmentar a capacidade de adaptación. Desta maneira, retemos o pode discriminatorio das CNN, mentres que aproveitamos os beneficios dos comités en termos de modularidade, escalabilidade, reversibilidade, xeneralización e robustez.

Seguindo entón o marcado polo estratexia de auto-adestramento, as predicións do comité nun determinado momento úsanse como pseudo-etiquetas para realizar (ou non) actualizacións que melloren o rendemento. Así, a nosa proposta pode realizar unha adaptación *online* en contextos específicos sen requirir un gran cantidade de mostras para crear os modelos (xa sexa etiquetados ou non). Tampouco será necesario que as mostras colleitadas sexan almacenadas en memoria. A ausencia destes dous requirimentos é de especial interese para aplicacións relacionadas coa biometría debido a que eliminan calquera problema relacionado coa privacidade.

Durante o experimentos, eliximos o campo da vídeo-vixilancia como un exemplo paradigmático de contexto específico no que o recoñecemento facial pode ser de gran utilidade. Polo tanto, simulamos unha contorna no que nos propoñíamos recoñecer varios individuos usando soamente uns poucos fotogramas anotados. Como o rendemento inicial obtido con este conxunto de datos tan limitado é relativamente baixo, o sistema deberá ser capaz de adquirir nova información de forma non supervisada para incrementar o poder de recoñecemento.

Como se mencionou anteriormente, *De-SVM* foi o sistema proposto para o caso de verificación de caras. Os experimentos for realizados usando tanto a COX Face Database como a YouTube Faces (YTF), mostrando en ambos caso un importante incremento do rendemento inicial (dende un 37% a un 85% TAR@FAR1% en COX e dende un 57% a un 75% en YTF) en comparación con outras técnicas de aprendizaxe incremental baixo o mesmo enfoque de auto-adestramento. Ademais, a arquitectura baseada en comités de *De-SVM* mostrou unha gran robustez respecto ao repetido intento de identificación de secuencias de impostores. Esta robustez permite ademais o fixado dun limiar de verificación un pouco máis ambicioso debido a que a inclusión de impostores no modelo dunha determinada identidade non afecta tanto ao rendemento como con outras técnicas.

Para o caso do recoñecemento facial en conxuntos abertos, as ideas de *De-SVM* son estendidas para a creación de *OS-DeSVM*). Neste caso, a Teoría dos Valores Extremos (*Extreme Value Theory*) úsase para ser capaces de distinguir entre identidades *coñecidas* e *descoñecidas*. Ademais, grazas a encapsular cada actualización nun clasificador individual, proporcionamos un método para reverter as actualización (xa sexa debido a actualización erróneas como para limitar o tamaño dos comités). Neste caso, os experimentos restrínxense á COX Face Database, onde obtemos ata un 15% de mellora no F1-measure (acadando ata un 94% de F1-measure, dependendo do número de individuos) e mellorando o rendemento que obterían outros métodos de recoñecemento facial non adaptativos do estado da arte.

Finalmente, *OSDe-SVM* tamén mostra un rendemento bastante salientable respecto ao grado de apertura (*openness*) do conxunto de individuos. Demostramos que o seu rendemento continua a ser fiable cando os individuos de interese representan tan só un 5% do total de identidades que poden cuestionar ao sistema (60% *openness*, 50 individuos de interese nun total de 1000 identidades). En termos xerais, o rendemento decae paulatinamente ao aumentar a *openness* pero mantén o seu rendemento cando a proporción de individuos de interese (*coñecidos*) e os demais (*descoñecidos*) é 1:1.

Traballo futuro

A medida que as melloras no rendemento proporcionadas polos métodos totalmente baseados *deep learning* supervisado comezan a saturar, a comunidade científica está comezando a pivotar cara outros temas da aprendizaxe automática, como a aprendizaxe auto-supervisada (*self-supervised* learning) ou a aprendizaxe continua (*continual learning*). Esta é a tesitura na que esta tese fai as súas contribucións. Non obstante, como calquera liña de investigación incipiente, existen aínda unha gran cantidade de oportunidades para contribución adicionais. Nesta sección, recollemos algunhas das máis importantes.

En primeiro lugar, debemos destacar que un dos hándicaps máis relevantes que atopamos durante o desenvolvemento da tese foi a escaseza de de bases de datos que simulen un contexto real de vídeo-vixilancia. Neste sentido, o COX Face Database é un dos mellores da súa clase. Aínda así, a pesares de conter unha gran cantidade de individuos distintos, as secuencias dispoñibles por individuo estaban limitadas a 3 (todas elas gravadas durante o mesmo día). Este feito limita bastante a presencia de grande cambios temporais na aparencia das caras en cuestións e imposibilita realizar probas de adaptación nunha escala de tempo maior, algo indispensable para a aprendizaxe ao longo da vida (*lifelong learning*). Por todos estes motivos, a creación dun conxunto de datos realmente completo representaría unha importante contribución para o avance deste tipo de enfoques incrementais.

En segundo lugar, o enfoque de auto-adestramento (*self-training*) en combinación co uso de comités e descritores baseados en *deep learning* demostraron ser unha solución tanto interesante como efectiva para implementar a aprendizaxe incremental non supervisada. Porén, aínda mantendo estas ideas centrais, existe aínda moita investigación que realizar. Por exemplo, unha das partes máis complexas de traballar con comités de clasificadores en condicións de pouca supervisión é a definición de regras intelixentes que eliminen clasificadores ou que ponderen as súas decisións. Intuitivamente, tanto as actualizacións incorrectas como as actualizacións redundantes poden ser daniñas para o comité. Con *OSDe-SVM* propoñemos dous módulos (o de *self-healing* e o de limitación) para abordar, respectivamente, cada problema. Presentados ambos como unha primeira aproximación ao problema, cremos que este tema requiriría unha investigación máis exhaustiva para poder atopar unha solución óptima. Nesta mesmo marco, unha liña de investigación interesante sería a de aproveitar as respostas negativas do comité para mellorar as decisión dalgunha maneira en troques de simplemente descartalas. Finalmente, outra alternativa interesante sería a de estender o enfoque de auto-adestramento ao caso no que sexa posible incorporar novas clases (identidades) ao conxunto de individuos de interese a recoñecer.

E en terceiro lugar, a pesar de ser o núcleo do noso procedemento experimental, non debemos esquecer que o recoñecemento de caras en vídeo-vixilancia é utilizado soamente como unha tarefa de proba de concepto. Cremos que as propostas desta tese poderían ser facilmente trasladadas a outros contextos como por exemplo o recoñecemento de obxectos en robots, o recoñecemento de persoas ou calquera outra aplicación na que se requira detectar algo. Neste sentido, as aplicacións que permitan a extracción de datos multi-modais poderían ser especialmente relevantes debido a que axudarían ao proceso de auto-etiquetado, seguindo un enfoque máis de co-adestramento (*co-training*).