Electronics Laboratory (ELLAB)

Physics Department

University of Patras

Neural Networks: Deep learning strategies for problems with limited data

Dissertation

A thesis submitted for the degree of
Philosophiae Doctor (**PhD**)

Dimitrios Tsourounis

Physicist
Master in Electronics and Information processing

Patras, 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

Supervisor: George Economou

Professor Emeritus

Board of Examiners

- Economou George, Professor Emeritus, Department of Physics, University of Patras (Supervisor)

- Zois Elias, Associate Professor, Department of Electrical and Electronic Engineering, University of West Attica (member of the Supervisory Committee)

- Anastassopoulos Vassilis, Professor, Department of Physics, University of Patras (member of the Supervisory Committee)

- Skodras Athanassios, Professor, Department of Electrical and Computer Engineering, University of Patras

- Dermatas Evangelos, Associate Professor, Department of Computer Engineering & Informatics, University of Patras

- Bakalis Dimitris, Assistant Professor, Department of Physics, University of Patras

- Theodorakopoulos Ilias, Assistant Professor, Department of Electrical and Computer Engineering, Democritus University of Thrace

Το έργο συγχρηματοδοτείται από την Ελλάδα και την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) μέσω του Επιχειρησιακού Προγράμματος «Ανάπτυξη Ανθρώπινου Δυναμικού, Εκπαίδευση και Διά Βίου Μάθηση», στο πλαίσιο της Πράξης «Ενίσχυση του ανθρώπινου ερευνητικού δυναμικού μέσω της υλοποίησης διδακτορικής έρευνας – 2ος Κύκλος» (MIS-5000432), που υλοποιεί το Ίδρυμα Κρατικών Υποτροφιών (ΙΚΥ).

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

Επιχειρησιακό Πρόγραμμα
Ανάπτυξη Ανθρώπινου Δυναμικού,
Εκπαίδευση και Διά Βίου Μάθηση
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the project "Strengthening Human Resources Research Potential via Doctorate Research – 2nd Cycle" (MIS-5000432), implemented by the State Scholarships Foundation (ΙΚΥ).

Ευρωπαϊκή Ένωση
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Abstract

Small sample size learning (SSSL) problem arises when the available training data are limited, making it challenging for machine learning models to capture meaningful patterns and provide accurate predictions. In computer vision applications, constraints on training data are common due to data collection difficulties or high annotation costs. This PhD thesis focuses on exploring deep learning strategies tailored for addressing the SSSL problem, with a specific emphasis on developing efficient training methods for convolutional neural networks (CNNs) when only a limited amount of data are available. Different approaches exist based on the space being considered: data augmentation techniques in the input space, approximating target functions with regularization and pretraining in the model space and encoding relationships between data points within a latent feature space. In this dissertation we propose methods that attack SSSL in one or multiple spaces simultaneously. The applications studied in this thesis include biometric verification in the offline signature verification (OffSV) problem, which currently lacks a large available offline signature dataset, and the biomedical problem of human epithelial type-2 (Hep-2) cell classification through indirect immunofluorescence (IIF) microscopy images, involving a challenging annotation process.

Initially, shallow representation learning approaches, utilizing traditional computer vision techniques, are studied as a baseline scenario of approaching SSSL. This enabled us to gain valuable insights into the intrinsic characteristics of the studied problems and enhances the interpretability of the results. Subsequently, a hybrid scheme combining hand-crafted descriptors with a CNN model is proposed. Hand crafted features can create representations with desired invariance characteristics, hence when used as input to a CNN, can provide a more effective starting point for training with limited samples size.

A different path to address the SSSL problem studied in this dissertation involves utilizing external data from a similar domain with data abundance. These data can serve as information carriers within a sophisticated training procedure, aimed at enhancing performance in the target problem that suffers data limitations. Such methods were developed in the context of OffSV, where auxiliary handwritten text data were utilized during the training of CNNs in the writer identification task, managing to learn effective encodings of signature images by employing domain adaptation techniques, achieving comparable performance or even surpassing models trained on thousands of signature images.

The first such approach proposed in this thesis is explicit domain adaptation, which encompasses metric learning using an additional transformation layer trained via contrastive loss, used to transform the outputs of a pretrained CNN model. The second proposed technique is implicit domain adaptation, implemented through teacher supervision in the Feature-based Knowledge Distillation (FKD) scheme. This method leverages both local and global information from intermediate representations of the teacher to facilitate efficient knowledge transfer. Results demonstrate that the proposed approaches effectively address the SSSL problem in the OffSV domain, operating in either the feature space or the model space, by utilizing auxiliary data in the input space to overcome the challenges posed by the data limitations.

Περίληψη

Το πρόβλημα εκμάθησης με μικρό πλήθος δεδομένων προκύπτει όταν τα διαθέσιμα δεδομένα εκπαίδευσης είναι περιορισμένα, κάτι που καθιστά δύσκολη την αποτελεσματική εκπαίδευση μοντέλων μηχανικής εκμάθησης και την πρόβλεψη ακριβών αποτελεσμάτων. Η διαθεσιμότητα μεγάλου όγκου δεδομένων συχνά αποτελεί πρόκληση, τόσο σε επίπεδο αποθήκευσης και επεξεργασίας, όσο και σε επίπεδο συλλογής, ελέγχου και χειρωνακτικής επισήμανσης των δεδομένων, ιδιαίτερα σε προβλήματα επιβλεπόμενης εκμάθησης. Η παρούσα Διδακτορική Διατριβή εστιάζει στην ανάπτυξη αποδοτικών και πρωτοποριακών τεχνικών εκμάθησης που καθιστούν εφικτή την αξιοποίηση τεχνικών αιχμής από το χώρο των βαθιών Συνελικτικών Νευρωνικών Δικτύων (ΣΝΔ) σε προβλήματα με εγγενείς περιορισμούς στα διαθέσιμα δεδομένα εκπαίδευσης, όπως π.χ. είναι οι βιομετρικές και βιο-ιατρικές εφαρμογές.

Μια απλή ταξινόμηση των προσεγγίσεων για την επίλυση του προβλήματος εκμάθησης με μικρό πλήθος δεδομένων μπορεί να πραγματοποιηθεί με βάση το πεδίο που εφαρμόζονται οι διάφορες τεχνικές, αξιοποιώντας το χώρο εισόδου των δεδομένων με κυρίαρχες τις μεθόδους επαύξησης των δεδομένων, το πεδίο του μοντέλου αναζητώντας τη βέλτιστη συνάρτηση για την αποδοτική κωδικοποίηση της πληροφορίας, και την ανάπτυξη σχέσεων (αν)ομοιότητας στα εξαγόμενα αποτελέσματα του μοντέλου. Αρχικά μελετώνται μοντέλα εκμάθησης με χρήση ρηχών μεθόδων αναπαράστασης εικόνων, χρησιμοποιώντας κλασσικές τεχνικές υπολογιστικής όρασης ως βάση αναφοράς στο πρόβλημα με μικρό πλήθος δεδομένων. Αυτό είναι βοηθητικό για την κατανόηση των χαρακτηριστικών των σημάτων που μελετώνται αλλά και την καλύτερη εξήγηση των αποτελεσμάτων. Στη συνέχεια, προτείνεται μια υβριδική μέθοδος που συνδυάζει κλασσικούς περιγραφείς εικόνων με ένα ΣΝΔ. Ο κλασσικός τρόπος κωδικοποίησης της πληροφορίας εφοδιάζει τις προκύπτουσες αναπαραστάσεις της εικόνας με επιθυμητά χαρακτηριστικά, και όταν χρησιμοποιούνται ως είσοδος σε ένα ΣΝΔ, μπορούν να παρέχουν ένα πιο αποτελεσματικό σημείο εκκίνησης για την εκπαίδευση του δικτύου με περιορισμένο αριθμό δειγμάτων. Σε μια διαφορετική κατεύθυνση αντιμετώπισης του προβλήματος εκμάθησης με μικρό πλήθος δεδομένων, αξιοποιήθηκε η χρήση εξωτερικών δεδομένων από έναν παρόμοιο πρόβλημα με πληθώρα δεδομένων. Αυτά τα δεδομένα εξάχθηκαν έτσι ώστε να εξυπηρετούν ως φορείς πληροφορίας μια ειδικά σχεδιασμένη διαδικασία εκπαίδευσης, με στόχο να βελτιωθεί η απόδοση στο πρόβλημα που υποφέρει από περιορισμούς δεδομένων. Σε αυτή την περίπτωση, η αποτελεσματική προσαρμογή των δύο προβλημάτων, του προβλήματος με επάρκεια δεδομένων και του προβλήματος ενδιαφέροντος με περιορισμένα δεδομένα, πραγματοποιείται σχεδιάζοντας τη διαδικασία εκπαίδευσης τόσο άμεσα με την εκμάθηση αποστάσεων μέσω ενός πρόσθετου επιπέδου μετασχηματισμού που χρησιμοποιείται για να μετασχηματίσει τις εξόδους ενός προ-εκπαιδευμένου μοντέλου ΣΝΔ στη βάση της ομοιότητας των σημάτων που μελετώνται όσο και έμμεσα μέσω ενός σχήματος απόσταξης γνώσης μεταξύ δύο ΣΝΔ, όπου το ένα δίκτυο έχει το ρόλο του μαθητευόμενου και το άλλο του επιβλέποντος, σχηματίζοντας καινοτόμες συναρτήσεις ομοιότητας μεταξύ των ενδιάμεσων αναπαραστάσεων των δύο μοντέλων για την αποτελεσματική μεταφορά της πληροφορίας από το δίκτυο επιβλέποντα κατά τη διάρκεια της εκπαίδευσης του δικτύου μαθητευόμενου. Επομένως, στα πλαίσια αυτής της διδακτορικής διατριβής, σχεδιάστηκαν πρωτοποριακές προσεγγίσεις επίλυσης του προβλήματος εκμάθησης με περιορισμένα δεδομένα αναπτύσσοντας τεχνικές σε πολλαπλά πεδία του προβλήματος αλλά και δοκιμάζοντας διαφορετικές εφαρμογές ενδιαφέροντος.

Acknowledgements

During my years of PhD research, I had the privilege of interacting and collaborating with numerous individuals, many of whom played pivotal roles in the completion of my dissertation. Therefore, I would like to take this opportunity to express my sincere gratitude to all those who generously contributed to the successful realization of my academic endeavor.

First and foremost, I am deeply grateful to my supervisor, Professor George Economou, for sharing his wealth of experience in both scientific and personal aspects throughout these years. His constructive feedback and unwavering support were instrumental in completing my PhD thesis. I also extend my appreciation to Professor Elias Zois, a member of the three-member advisory committee, for our countless insightful discussions and his guidance in conducting high-level research. The progress we achieved together is truly invaluable, and I am thankful for all that I have learned from them.

A special acknowledgment goes to Professor Spiros Fotopoulos, a former member of the three-member advisory committee, who believed in me from the very beginning. His encouragement has been a driving force in my academic journey. Furthermore, I am grateful to Professor Vassilios Anastassopoulos, member of the three-member advisory committee, for his valuable advice and support throughout all these years. I would also like to express my thanks to Professors Thanos Skodras, Dimitris Bakalis, and Vangelis Dermatas for accepting to be part of the jury and evaluating my work.

I am particularly thankful to Ilias Theodorakopoulos, who served as a Post Doc colleague during the early stages of my PhD research and later became a member of the seven-member examination committee as a professor. His extensive experience and deep understanding of the scientific subject have been instrumental in pushing the boundaries of my research. I am deeply indebted and fortunate to have collaborated with him, and I look forward to continuing our partnership in the future. Additionally, I express my gratitude to Dr. Dimitris Kastaniotis for our excellent collaboration in various projects over the years, as well as to Dr. Christos Theoharatos for his belief in my abilities and our fruitful collaboration.

I would also like to extend my thanks to the faculty members of the Electronic Laboratory for their collaboration and support.

In conclusion, I extend a heartfelt thank you to my friends and family, whose support has been crucial in maintaining balance, especially during the challenging moments of this entire journey.

Dimitrios Tsourounis,

Patra, July 2023

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Small sample size learning problem

Deep learning has revolutionized our world in the last decade, achieving remarkable performance in various applications, such as computer vision, speech recognition, natural language processing, and recommendation engines [1]–[3]. Among the most popular deep learning models are convolutional neural networks (CNNs) due to their efficiency in many tasks [4]–[6]. The advancements in parallel processing hardware along with the availability of large amount of annotated data and the emergence of new theoretical tools and techniques in the field of deep learning have made it possible to train CNNs with deep hierarchical representations [7], [8]. CNNs are compositional learning models that use multiple layers to learn features. Higher-level learned features are formed by combining lower-level features in a hierarchical way. Automatically learning features at multiple levels of abstraction enables CNNs to learn complex functions, mapping the input to the output directly from data. These layers of features are not designed by human engineers, but they are learned from data using a data-driven learning procedure through end-to-end training. Training a CNN involves optimizing the learnable parameters across all layers of the model. Since the number of model's learnable parameter (weights) is usually very large, training typically requires a large dataset. Ultimately, the effectiveness of deep learning models heavily relies on the availability of abundant and high-quality training data.

In many real-world scenarios, collecting large number of samples and annotating them is not always practical. The small sample size learning problem (SSSL) arises when the available training data are limited in size, posing a challenge for machine learning models to extract meaningful patterns and make accurate predictions. The scarcity of training data poses difficulties in capturing the underlying patterns and variability in the dataset accurately, leading to overfitting. As a result, the machine learning models perform well on the training data but fail to generalize to unseen data. The SSSL problem is particularly evident in fields such as biometrics, where privacy concerns hinder the collection of a large amount of labeled data, while operational conditions of such systems rely on a small number of available reference samples for each user. Similar challenges are also encountered in the realm of biomedical applications, where manual annotation (diagnosis) requires high expertise and entails significant costs. These limitations trammel the use of deep learning models in a wide range of problems, despite the potential benefits they offer.

In a general formulation, the learning problem can be expressed as a minimization problem:

$$\min_{W,b} \ell(\varphi, Y) \qquad \textit{eq. 1.1}$$

and the process of mapping data using a learning model can be mathematically represented as:

$$F = \varphi(WX + b) \quad \textit{eq. 1.2}$$

where φ is the model with weights W and bias b, X is the input data, F is the output feature, Y is the desired output, and ℓ is the loss function. To address the SSSL problem, the proposed solutions investigate the input space, model space, and feature space depending on whether they operate on X, φ, or F [9]. Methods that operate on input space aim to augment the data by generating additional samples or transforming existing samples to optimize the feature space. The methods that focus on the model space approximate target functions to map inputs to the output while the feature space optimization aims to efficiently encode relationships among data points on the embedded space.

In the input space, there are several techniques that can be employed to address the SSSL problem by increasing the number of samples or making the data representation more informative. When working with images, these methods involve generating additional image representations to artificially expand the dataset [10], [11]. One approach is to introduce variations in certain qualities of the data, such as through geometric and color space transformations [12]. By creating new samples from the existing ones, the size and diversity of the training set can be increased. To further enhance the information extraction, a more advanced approach is to apply encoding techniques that transform existing images into new representations. This encoding process captures specific image characteristics and provides valuable inputs to deep learning models. By leveraging these techniques, the models can generalize better or faster and effectively handle the SSSL problem [13], [14]. Additionally, the utilization of synthetic data is another strategy to address the SSSL problem [15]. Synthetic data can be generated even without relying on the original dataset by using generative methods that generate new images resembling the target data distribution. One significant advantage of synthetic data generation is the elimination of manual data labeling, as it becomes possible to generate synthetic samples that are tailor-made with desired characteristics in the first place.

However, it is important to ensure the quality and realism of the augmented data for its effectiveness in deep learning applications. Therefore, it is crucial to provide augmented data that captures relevant image properties and aligns with the specific objectives of the problem at hand since there are domains, where large-consistent data generation or transformations are not always feasible, like the biometric problems.

By working on the model space, the SSSL problem can be effectively addressed through various approaches. Initialization tricks and transfer learning methods are commonly utilized to mitigate this challenge [16]–[18]. One approach involves adapting pretrained CNN models or predefined CNN layers to the target problem, enabling finetuning with a small amount of data [19], [20]. Initializing the model's weights based on a similar problem serves as a good initial baseline for further re-training. Also, incorporating handcrafted filters in CNN layers, either as a starting point or by freezing coefficients, enhances the model's ability to generalize with limited training data, following layer-wise or end-to-end training settings [21], [22]. To prevent overfitting, custom loss functions tailored to the available data and regularization techniques are often utilized [23]. Introducing a penalty term to the loss function during training discourages the CNN from becoming too complex or having large parameter values and incorporates specific characteristics into the model. Another valuable approach is knowledge distillation, where a student network is trained with feedback from a teacher network [24]. This knowledge transfer process enables the student model to benefit from the insights captured by the expert teacher model, resulting in better generalization. It is valuable to note that designing these methods requires empirical exploration since there is no definitive rule-of-thumb that specifies the commonalities between the first and target tasks in each application of interest. Therefore, it is essential to experiment and adapt these techniques based on the specific characteristics and requirements of the problem.

Deep metric learning is a popular approach for addressing the SSSL problem in the feature space. Deep metric learning is focused on learning representations that encode similarity and dissimilarity between samples [25]. By employing loss functions that encourage similar samples to have closer embeddings in the latent space and dissimilar samples to have larger distances, deep metric learning facilitates the extraction of discriminative features even with limited training data. Exploring pairwise relationships between data instances results in effectively expanding the number of available training samples. This approach shares similarities with few-shot learning methods, which aim is to generalize the pretrained model to new unseen categories of data. In few-shot learning, the model compares the feature representations of a few-shot example with the limited labeled examples and makes predictions based on their similarity scores [26]. However, it is important to note that while few-shot learning focuses on classifying new samples, the goal in SSSL is to learn a distance function that serves as a metric to quantify the similarity between data instances. Additionally, incorporating mining hard examples can significantly enhance the effectiveness of metric learning methods [27]. Hard example mining involves identifying and selecting challenging or informative samples from the training set. By prioritizing the inclusion of these difficult samples during the training process, the model can focus on learning from the most informative instances, thereby improving its generalization. Deep metric learning can be applied using different setups, such as the Siamese scheme where similar and dissimilar pairs are created, or the Triplet scheme by designing triplets of training data instances, instead of pairs [28], [29]. Various ranking losses [30], including contrastive loss, triplet loss, margin loss, and hinge loss, offer different formulations to choose from based on the specific requirements and objectives of the problem being studied.

1.2 Subject of the PhD Thesis

This PhD thesis aims to explore deep learning strategies specifically designed to address the small sample size learning (SSSL) problem. The focus lies in developing efficient training methods for convolutional neural networks (CNNs) when there are limited data available for the target problems of interest. The research investigates the application of these strategies in domains with inherent constraints on the training data availability, such as biometric and biomedical problems.

Biometrics play a crucial role in balancing convenience, security, and user experience across everyday activities [31]. Although many modern biometric solutions are continuously being developed with a variety of input signals, handwritten signatures remain widely accepted and legally binding in many sectors such as banking, legal, and government. Signature is considered a behavioral biometric trait due to its association with an individual's learned behavioral patterns. The most challenging task is signature verification problem that authenticates the identity of a person on the basis of the claimed identity, by accepting the writer’s genuine signatures and rejecting the forgery ones [32]. While manual signature comparison seems like an ineffective way to handle the masses of documents that need to be checked in a small amount of time, automatic handwritten signature verification systems are pivotal to reduce fraud. This system automatically detects authenticity, meaning that the questioned signature owns to the claimed writer and thus it is genuine or whether the signature has been provided from anyone else and thus it is forgery. The problem of offline signature verification (OffSV), which analyses signature images after the signing process, presents an ideal scenario for investigating the proposed approaches in addressing the challenge of limited data. Additionally, during the course of this PhD research, the retraction of the largest offline signature dataset, which was the only publicly available and large enough dataset for training deep architectures, has emerged as a significant practical problem in this field. As a result, a challenging need has arisen to address this issue. Therefore, the proposed research direction seizes this opportunity and focuses on the OffSV problem. In this dissertation, different approaches are designed to tackle the challenges of SSSL in this field, and the effectiveness of the proposed methods is evaluated using several widely used offline signature verification datasets.

Biomedical research involves understanding of human health and disease, including the development of diagnostic tools, therapies, and improving overall healthcare outcomes [33], [34]. In the context of antinuclear autoantibodies (ANA) detection, the use of indirect immunofluorescence (IIF) on human epithelial type-2 (HEp-2) cells is a standard protocol due to its high sensitivity and ability to capture a wide range of antigens [35], [36]. While numerous nuclear and cytoplasmic patterns can be observed in HEp-2 samples, typically only a few classes are considered clinically relevant. The accurate classification of fluorescence patterns is crucial as specific diseases are associated with distinct staining patterns. However, manual pattern identification using a microscope is time-consuming, labor-intensive, and subjects to the physicians' experience while processing large amount of data might introduces significant oversight errors. Automatic single cell HEp-2 fluorescence images classification could enhance Computer Aided Diagnosis (CAD) systems, providing complementary information by filtering the amount of data the expert has to inspect. The unique characteristics of HEp-2 cell images, combined with the limited availability of data, provide a good opportunity to assess the efficacy of a SSSL-oriented CNN approach in the context of IIF cell classification. Two widely recognized benchmark datasets are utilized to evaluate the proposed method.

This dissertation follows a stepwise structure, studying the SSSL from multiple perspectives. It starts by exploring the utilization of hand-crafted representations and shallow learning approaches, gradually advancing towards a hybrid approach that combines hand-crafted image representations with CNNs. Next, the dissertation culminates in the study of deep learning methods that tackle the SSSL problem through various solutions, proposing novel techniques such as external data utilization, domain adaptation, and knowledge distillation, specifically tailored to the OffSV problem. Overall, this dissertation provides a comprehensive investigation of various approaches and strategies for addressing the SSSL problem. By gradually transitioning from shallow and/or hand-crafted representations to hybrid approaches and ultimately to deep learning methods, it offers a systematic exploration of techniques to enhance the generalization capability of deep learning models in scenarios with limited training data. The structure of the dissertation, in detail for each chapter, is as follows:

Chapter 2 introduces shallow representation methods that specifically focus on encoding signature image local neighborhoods. Traditional computer vision methods serve as a baseline solution to tackle the limited sample problem in the OffSV task. These methods involve the utilization of hand-crafted preprocessing and feature extraction stages, combined with trained classifiers, effectively addressing the Small Sample Size Learning (SSSL) problem. While these methods may not fully leverage the potential of automated feature learning that deep learning offers, where the features are learned directly from the data, they uncover discriminative characteristics of signature images. The features used in these methods are manually designed, drawing upon the knowledge and expertise of the human engineers in the field. Different encoding mechanisms are developed in this chapter, involving hand-crafted features using image visibility graph motifs (IVG) [37], and mapping the signature images as Symmetric Positive Definite (SPD) points within the SPD manifold [38]. Also, shallow learning is employed through a patch-based sparse representation method [39], [40], where reconstructive overcomplete dictionaries are learned by utilizing patches extracted from a set of signature images. The sparse coding procedure then maps pixel values to sparse features, resulting in an effective representation to describe the information of signature images. Finally, an introduction to deeper topologies is explored by stacking multiple sparse representation layers and training them in a layer-wise manner [41]. The analysis of these approaches not only enables us to comprehend the effectiveness of hand-crafted features but also provides valuable insights into the inherent characteristics of signature signals. This understanding can subsequently guide the design of deep learning schemes aimed at enhancing the model's capability to extract meaningful information from limited data.

Chapter 3 presents a hybrid scheme that involves an initial step of calculating hand-crafted descriptors, followed by the utilization of a CNN. The process of computing hand-crafted features transforms the image into a more informative representation that captures specific image characteristics. In this chapter, dense SIFT (Scale-Invariant Feature Transform) is implemented, where descriptors are calculated for every pixel of a grayscale image, running SIFT algorithm [42] on a dense gird of locations at a fixed scale. This approach extracts information from the input data that is richer than the raw pixel values and represents them in a suitable form for subsequent analysis. The resulting image representation, known as SIFT-Image [43], preserves the spatial structure of the original image and serves as input to the CNN model. The CNN operates regularly but with the new input representation, harnessing the learning capabilities of deep neural networks. By operating in the image space, this approach aims to address the SSSL problem enhancing the informativeness of the inputs. However, due to the degenerate nature of signature images composed of line segments and curved strokes in a uniform canvas, dense SIFT descriptors may not efficiently capture the underlying information in signature signals. As a result, it is challenging to augment signature images into a more informative representation using SIFT-Image. Hence, the proposed method's effectiveness in tackling the SSSL problem is evaluated on the biomedical Hep-2 IIF cell classification problem. Nevertheless, in order to explore the characteristics of handwriting within a straightforward classification problem, a study is conducted using the widely recognized MNIST handwritten digit dataset, which is considered the standard benchmark dataset in the field of machine learning. Although the design of this hybrid scheme is motivated by the need to address the SSSL problem, the local rotation invariance provided from the dense SIFT descriptors could be beneficial in many other problems. Therefore, the chapter delves into additional applications, such as cloud type classification using ground-based all-sky images and the lip-reading problem using video data of mouth region, where local rotation property plays a crucial role in achieving efficient results. Ultimately, the hybrid approach presented in this chapter combines the strengths of shallow hand-crafted representations with deep neural networks, offering more informative data input for training CNNs with less data. By integrating these two techniques, the goal is to overcome the limitations imposed by small sample sizes while incorporating the beneficial properties derived from the hand-crafted methods into the overall hybrid system.

Chapter 4 proposes a method that leverages external data and domain adaptation techniques to address the SSSL problem in OffSV. Training deep architectures inherently requires a substantial amount of data, however the utilization of auxiliary data from a related task could serve as an information carrier to substitute the target data. In many cases, this approach requires additional domain adaptation practices to address the distribution mismatch between the auxiliary data and the target data. This chapter focuses on addressing the problem of learning informative features by employing prior knowledge from a similar task in a domain with an abundance of training data. In particular, we demonstrate that an appropriate pretraining of a CNN model in the task of handwritten text-based writer identification task can dramatically improve the efficiency of the CNN in the OffSV task, enabling to obtain state-of-the-art performance with an order of magnitude less training signature samples. In the proposed scheme, we leverage the relevance of writing and signing processes, which is also enhanced by preprocessing the raw text data to mimic the signal characteristics of signature images through a fast and efficient text manipulation that is highly suitable for large-scale data processing. After the pretraining of the CNN in writer identification task using specially processed handwritten text data, the learned features are tailored to the signature problem though a metric learning stage that utilizes contrastive loss to learn a mapping of the signatures' features to a latent space that suits the OffSV task. Therefore, the proposed SSSL solution for deep feature learning operates both in the image space including external data and in the feature space via contrastive metric mapping. This approach enables us to leverage the knowledge and patterns present in the auxiliary data to improve the learning and generalization capabilities of the deep architecture, ultimately leading to enhanced performance on the target task. At the final stage, the proposed scheme utilizes Writer-Dependent (WD) classifiers learned on a few reference samples from each writer. Our system is tested on the three challenging signature datasets, CEDAR, MCYT75, and GPDS300GRAY and the obtained accuracy in terms of Equal Error Rates (EER) is statistically equivalent to the most popular CNN model (SigNet) in the field [44], which had been trained on the largest offline signature dataset consisting of over ten thousand signature images contributed by more than five hundred writers, despite the proposed method utilizes significantly smaller training set of signature images and no skilled forgery signatures during training.

Chapter 5 proposes a feature-based knowledge distillation (FKD) method to address the SSSL problem in OffSV. FKD focuses on transferring the knowledge of intermediate activations from a teacher model to a student model [45]. Unlike simply mimicking the teacher's output probabilities, FKD aims to align feature representations between both models. The teacher strongly guides the training of the student CNN by harnessing multiple intermediate layers, providing enhanced supervision that results in improved generalization and knowledge transfer, even when auxiliary training data were utilized. Therefore, this approach extends the work presented in Chapter 4 but operating in the model space to address the SSSL problem, using the FKD method as a means of implicit domain adaptation from the teacher model. This chapter introduces a novel approach to leverage the knowledge of existing expert models for training new CNNs. The proposed Student-Teacher (S-T) configuration, combining graph-based similarity function for local activations with global similarity measures to supervise student's training, using only handwritten text data. The feature maps' similarity in shallower layers is calculated using the geometrical criterion based on a manifold-to-manifold distance, while the final vectorial representations of CNN models are compared using either the cross-entropy function with temperature softmax or a similarity metric based on redundancy-reduction principle expresses through the cross-correlation matrix. The proposed FKD captures information from multiple semantic perspectives and exploits the hierarchical representational learning ability of multi-layer deep structures, combining both local and global information to improve the performance of the student model. The models trained using this technique exhibit comparable, or superior, performance to the teacher model across three signature datasets: CEDAR, MCYT75, and GPDS300GRAY. Based on the current state-of-the-art in OffSV research [46], we chose to utilize the popular pretrained SigNet feature extractor CNN as the teacher model, while the architecture of the student model follows the Residual Network (ResNet) family. The proposed work presents an efficient knowledge transfer from a successful CNN-based feature extractor to a student CNN of a different architecture without employing any signatures during the S-T training. This study demonstrates the efficacy of leveraging existing expert models to overcome data scarcity challenges in OffSV and potentially other related domains.

1.3 Publications

This PhD Thesis is based on research results that have been published during its elaboration in peer-reviewed scientific journals and peer-reviewed international conferences. The list of publications directly related to the material of this dissertation is as follows:

Publications on peer-reviewed journals:

- **Tsourounis, D.**, Theodorakopoulos, I., Zois, E. N., & Economou, G. (2023). Leveraging Expert Models for Training Deep Neural Networks in Scarce Data Domains: Application to Offline Handwritten Signature Verification. Submitted to Neurocomputing journal.
- **Tsourounis, D.**, Theodorakopoulos, I., Zois, E. N., & Economou, G. (2022). From text to signatures: Knowledge transfer for efficient deep feature learning in offline signature verification. Expert Systems with Applications, 189, 116136.
- Zois, E. N., **Tsourounis, D.**, Theodorakopoulos, I., Kesidis, A. L., & Economou, G. (2019). A comprehensive study of sparse representation techniques for offline signature verification. IEEE Transactions on Biometrics, Behavior, and Identity Science, 1(1), 68-81.

Publications on peer-reviewed international conferences:

- Andrianakos, G., **Tsourounis, D.**, Oikonomou, S., Kastaniotis, D., Economou, G., & Kazantzidis, A. (2019). Sky Image forecasting with Generative Adversarial Networks for cloud coverage prediction. In 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA) (pp. 1-7). IEEE.
- **Tsourounis, D.**, Theodorakopoulos, I., Zois, E. N., Economou, G., & Fotopoulos, S. (2018). Handwritten signature verification via deep sparse coding architecture. In 2018 IEEE 13th image, video, and multidimensional signal processing workshop (IVMSP) (pp. 1-5). IEEE.

Book chapters:

- Theodorakopoulos, I., & **Tsourounis, D.** (2023). A Geometric Perspective on Feature-Based Distillation. In Advancements in Knowledge Distillation: Towards New Horizons of Intelligent Systems (pp. 33-63). Cham: Springer International Publishing.

Contribution to other publications on peer-reviewed journals during my work on this PhD Thesis:

- Zois, E. N., Said, S., **Tsourounis, D.**, & Alexandridis, A. (2023). Subscripto multiplex: A Riemannian symmetric positive definite strategy for offline signature verification. Pattern Recognition Letters, 167, 67-74.
- **Tsourounis, D.**, Kastaniotis, D., Theoharatos, C., Kazantzidis, A., & Economou, G. (2022). SIFT-CNN: When Convolutional Neural Networks Meet Dense SIFT Descriptors for Image and Sequence Classification. Journal of Imaging, 8(10), 256.
- **Tsourounis, D.**, Kastaniotis, D., & Fotopoulos, S. (2021). Lip Reading by Alternating between Spatiotemporal and Spatial Convolutions. Journal of Imaging, 7(5), 91.

Contribution to other publications on peer-reviewed international conferences during my work on this PhD Thesis:

- Zois, E.N., Zervas, E., **Tsourounis, D.**, & Economou, G., (2020). Sequential Motif Profiles and Topological Plots for Offline Signature Verification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 13248-13258), 14-19 June 2020, Seattle, Washington, USA.
- Kastaniotis, D., **Tsourounis, D.**, & Fotopoulos, S., (2020). Lip Reading modeling with Temporal Convolutional Networks for medical support applications., Oral Presentation, In 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 17-19 October 2020, Yantai, China.
- Kastaniotis, D., **Tsourounis, D.**, Koureleas, A., Peev, B., Theoharatos, C., & Fotopoulos, S. (2019). Lip Reading in Greek words at unconstrained driving scenario. In 2019 10th

International Conference on Information, Intelligence, Systems and Applications (IISA) (pp. 1-6). IEEE.

- Katakis, S., Barotsis, N., Kastaniotis, D., Theoharatos, C., **Tsourounis, D.**, Fotopoulos, S., & Panagiotopoulos, E. (2018). Muscle Type Classification on Ultrasound Imaging Using Deep Convolutional Neural Networks. In 2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP) (pp. 1-5). IEEE.
- Kastaniotis, D., Ntinou, I., **Tsourounis, D.**, Economou, G., & Fotopoulos, S. (2018). Attention-aware generative adversarial networks (ATA-GANs). In 2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP) (pp. 1-5). IEEE.
- Zois, E. N., Papagiannopoulou, M., **Tsourounis, D.**, & Economou, G. (2018). Hierarchical dictionary learning and sparse coding for static signature verification. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW) (pp. 432-442).

Chapter 2

Shallow representation learning

2.1 Introduction

In this chapter, shallow representation models are investigated in the context of the offline signature verification problem. Handwritten signature is a common biometric trait, widely used for confirming the presence or the consent of a person. Signature verification can be categorized into offline (static) and online (dynamic) based on the acquisition conditions. The first case focuses on analyzing the visual information and shape of the signature using a digitized version of the signing document, while the other requires a digitizing device, such as electronic tablets, to collect additional information during signing, including pen inclination, pressure, and spatial coordinates. Offline Signature Verification (OffSV) is the task of verifying the signer using static signature images captured after the signing process is completed. This task finds many applications, especially in the domain of forensics but also for ensuring the security of financial and legal documents, such as bank and compliance forms, contracts, and mail ballots [47]–[49]. Depending on the signature verification design plan, there are two main approaches: writer-dependent (WD) and writer-independent (WI) [44]. WD methods build one model per user and WI approach uses one single (global) model for all users. The WI methods usually take advantage of the increased number of training samples by generating pairs between signatures (generally a similar pair includes signatures of the same writer while a dissimilar pair combines genuine and forgery signatures) and embed the feature representations to a dissimilarity space for obtaining the final decision. The WD approaches rely on the signatures of each writer to create a custom model dedicated to each writer and thus, they turn to be more restricted but at the same time they could provide more efficient results due to tailoring on each individual. There are different types of forgery signatures and the most common is to divide simulated signatures into three categories: the random forgeries which are generated without access to the original signature, the simple or unskilled forgeries where the forger has information about the shape of genuine but not allowed much practice during falsification, and the skilled forgeries where the impostor attempts to carefully imitate the original signature with no constraints [50]. Finally, the main challenges faced by an OffSV system are: (a) the high intra-class variability between signatures of the same user, related on the psychophysical state of the signer and the conditions under which the signature apposition process occurs, (b) the partial knowledge during the design of the system and the registration of a user since there is access only to genuine signatures of the user while its skilled forgeries are not practically available, and accordingly (c) the limited number of available samples for each user.

Shallow representations in computer vision refer to feature representations that are designed to capture low-level visual features such as color, texture, and shape, rather than high-level semantic features. The goal of these models is to encode the local features and characteristics of a signature using numerical values or vectors. To achieve this, the chapter presents different approaches for encoding the neighborhoods of signatures. The process of encoding a signature typically involves dividing the signature into smaller regions (patches), and then extracting features from each patch. The feature extraction process produces a compact and informative representation, which can be used for signature verification tasks. Additionally, the relative location of the features could be incorporated through spatial pyramid matching techniques that divide an image into multiple subregions at different scales, creating a hierarchy of image subregions that form a pyramid. The presented shallow representation models can be categorized based on their feature extraction approach, which can be either hand-crafted methods or those that involve low-level learned features. Furthermore, the proposed methods can be partitioned based on whether they learn a distance via a metric function or focus on learning discriminative predictors. Although the shallow representation models are easy to interpret and can be trained efficiently with relatively small amounts of data, they require specially designed image preprocessing steps as well as pooling strategies to express spatial information locally for improving performance.

This chapter presents three different approaches for encoding signature image neighborhoods. The first approach focuses on visibility graphs, where sequential visibility motifs within signature image patches are analyzed. These image visibility graph motifs (IVG) capture patterns of connected nodes in the visibility graph, representing substructures within the image. The second approach explores the representation of symmetric positive definite matrices using the covariance descriptor of image feature maps. This involves calculating region covariance matrices for image patches and averaging them to obtain the covariance descriptor, mapping the image to the symmetric positive definite (SPD) manifold. In addition to utilizing the covariance descriptor as a discriminative feature for signature images, a metric learning process is employed within the manifold to rearrange the SPD points improving the verification purpose. The third approach utilizes the sparse representation (SR) into the signature verification problem to extract informative features by computing sparse coefficients that minimize reconstruction error for image patches using a learned dictionary. The dictionary learning process is performed for each writer using the signature patches from a set of reference signatures. Once the dictionary is learned, it can be used to represent and encode any query signature by computing its sparse coefficients. Ultimately, these three approaches offer distinct ways to represent and encode local features of images.

To conclude, this chapter describes shallow representation models that evaluated in the offline signature verification task. It highlights their focus on low-level visual features, the encoding of signature image neighborhoods, and the incorporation of spatial information through spatial pyramid matching. Understanding and utilizing these models can contribute to improved performance in signature verification tasks, while also offering interpretability and efficiency with smaller amounts of data.

2.2 Elementary Processes for offline signature images

2.2.1 Preprocessing

The preprocessing steps play a crucial role in transforming the signature image into a simplified and standardized representation, which is essential for accurate signature verification. This preprocessing consists of two main operations: binarization and thinning. Grayscale images are first binarized using Otsu's method [51]. Subsequently, morphological thinning operations are applied to the binary image to obtain a gradual skeletonization of the signature. The outcome of the thinning operation is crucial for the verification performance as it affects the shape of the signature image. Experimental observations indicate that the optimal thinning level (OTL) varies for each writer, and hence, it is not common for all databases. The OTL for each signature and writer is defined as the number of thinning operations that result in the steepest decrease of the density function. After enrolling a set of genuine reference signatures for a writer, the median value of the associated OTL values (MOTL) is selected as the thinning level applied to all the signatures related to this writer. This ensures that the thinning level used for each signature is consistent and appropriate for the writer's signatures characteristics. For any input signature claiming an identity, the number of thinning operations is determined by the MOTL value of the claimed signing person. Finally, to prepare the image for further processing, it is first inverted to have a black background and grayscale foreground. This is achieved by subtracting each pixel from the maximum brightness value (white), after setting the background pixels to white (255) and leaving the foreground pixels in grayscale.

2.2.2 Patch extraction

The signature patches are extracted from the original grayscale signature image, indexed by the signature’s skeleton pixels after applying the thinning operation MOTL times. Specifically, the patches' centers are sampled densely at every pixel of a signature's skeleton. As a consequence, the number of image patches equals the number of pixels of the signature's skeleton. Furthermore, the patches are centered, i.e., have their average intensity been subtracted in order to have a zero-mean value. The centering of each patch produces data invariant to the mean intensity and the learned structures, like edges, are anticipated to have zero-mean as well. In all the conducted experiments the patch size is set to five and thus, every patch has 25 pixels (resolution of 5×5). The main rationale behind this selection is to keep the complexity of the local manifold of patches reasonably low. With this aim, it is valuable to consider the parameters which affect the dimensionality and shape of the underlying local manifolds. In [52] Peyré shows that the local manifold of patches from cartoon images (images that contain sharp variations along regular curves) can be parameterized by two variables, leading to a manifold topologically equivalent to the surface of a cylinder in 3D space. This parameterization holds as long as the signal within each patch can be approximated by two regions (black and white) separated by a linear segment. If the patch size becomes larger and the edges within the patches appear curved, extra degrees of freedom have to be included to the signal's model thus leading to a more complex manifold. Similarly to cartoon images, the nature of the signal within signature patches is such that can be modeled by a handful of parameters if the patch size is small-enough, indicating a low-dimensional underlying manifold structure. On the other hand, the complexity can be dramatically increased if the patch size becomes large enough to contain curves and parts from neighboring line segments. The patch size equals 5, since it is a good tradeoff between the underlying signal's complexity since for smaller patches the local manifold obviously becomes degenerate - and the overall computational complexity.

2.2.3 Spatial Pyramid Pooling

Spatial Pyramid Pooling (SPP) is a technique used to aggregate local features over regions of interest to create a final feature vector. To implement this technique, the signature images are first segmented into a grid of equimass subregions using a spatial pyramid. For each segment, local feature vectors are extracted and subjected to a pooling operation. The pooled feature vectors from the entire image and its subregions are then concatenated into a single final feature vector for the entire image. The spatial pyramid consists of either a 2 × 2, 3 × 3, or 4 × 4 equimass subregion division, with a pooling function applied at each level. The pooled vector according to different pooling functions is defined as follows:

- *Average (Avg) pooling (F_1):* $f_{\mathrm{I}}^{F1} = \{ f_{\mathrm{I}}^{F1}[j] \} = \{ \frac{1}{M} \sum_{i=1}^{M} x^i[\mathrm{j}] \}, j = 1:K$ *eq. 2.1*
- *Maximum (Max) pooling (F_2):* $f_{\mathrm{I}}^{F2} = \{ f_{\mathrm{I}}^{F2}[j] \} = \max(|x^i[j]|),\ i = 1:M,\ j = 1:K$ *eq. 2.2*

- *Standard Deviation (Std) pooling (F_3):* $f_{\mathrm{I}}^{F3} = \{ f_{\mathrm{I}}^{F4}[j] \} = \{ \sqrt{\frac{\sum_{i=1}^{M} (\mathrm{x}^{i}[\mathrm{j}] - f_{\mathrm{I}}^{F1}[\mathrm{j}])^{2}}{M-1}} \}, j = 1:K$ *eq. 2.3*

- *Normalized sum (norm) pooling (F_4):* $f_{\mathrm{I}}^{F4} = \{ f_{\mathrm{I}}^{F4}[j] \} = \frac{\sum_{i=1}^{M} x^{i}[\mathrm{j}]}{\sum_{j=1:}^{K} \sum_{i=1}^{M} x^{i}[\mathrm{j}]}, j = 1:K$ *eq. 2.4*

- *L-2 normalized sum (L-2 norm) pooling (F_5):* $f_{\mathrm{I}}^{F5} = \{ f_{\mathrm{I}}^{F5}[j] \} = \frac{\sum_{i=1}^{M} x^{i}[\mathrm{j}]}{\sqrt{\sum_{j=1:}^{K} (\sum_{i=1}^{M} x^{i}[\mathrm{j}])^{2}}}, j = 1:K$ *eq. 2.5*

where each extracted feature x^{i} for any patch i has K elements and each region of interest includes M feature vectors that are pooled together. The average pooling (F_1) function computes the mean of the feature vectors extracted from the regions of interest. In contrast, the max pooling (F_2) operation selects the most salient feature value from each region of interest. Standard deviation (F_3) is an alternative pooling function that captures second-order statistics of the vectors' elements distribution, which could potentially improve the discrimination capabilities of the final feature vector. The normalized sum pooling (F_4) function produces feature vectors that are invariant to changes in intensity. Lastly, the L-2 normalized final feature vectors produced by the (F_5) function are projected onto the unit ball, which is important for linear classification kernels.

2.3 Designing systems for offline signature verification problem

2.3.1 Image Visibility Graph motifs (IVG)

2.3.1.i Method

Visibility Graphs (VG) and Horizontal Visibility Graphs (HVG) are methods of converting ordered sequences into a graph structure, enabling the application of graph theory in analyzing the data [53], [54]. The VG connects points in the sequences that have a clear line-of-sight (or visibility) to one another and forming a network of nodes and edges, while the HVG is a modified version that considers only the horizontal visibility between points, resulting in a simplified graph structure. Image natural Visibility Graphs (IVG) and Image Horizontal Visibility Graphs (IHVG) are extensions of this concept that map scalar fields and images into graphs, where each pixel represents a node and edges are added between nodes that are mutually visible. Local features can be extracted by detecting the local properties through visibility patches, which are small subgraphs in the IVG/IHVG. The concept of visibility patch is the natural extension to images of the concept of sequential visibility graph motifs [55] by extending the visibility criteria along one-dimensional sections of the two-dimensional signal (image patches) via scanning in horizontal, vertical, and diagonal directions. Sequential natural or horizontal visibility graph motifs profiles are defined as smaller substructures of n consecutive nodes that appear with characteristic frequencies. The visibility motifs used in this work are of low order, and specifically of size four, which can be used to create a six-dimensional feature vector that is computationally efficient and highly informative [55]. To better illustrate the encoding process for each signature patch, Figure 2-1 shows how the number of six IVG and six IHVG motifs' appearances are counted in different rasterized formats, including row, column, 1st (main) diagonal, 2nd (secondary) diagonal, and column-wise patch transformed-to-vector formats. The resulted visibility code is of size 60, obtained by concatenating the five individual formats. This encoding process produces handcrafted features and allows local patches to be represented by sequential visibility graph motifs while expresses local information by counting the repetitions of motifs in the image. Finally, the spatial pyramid pooling is applied to the local features (visibility codes) corresponds to image segments in order to provide the final feature vector for any image.

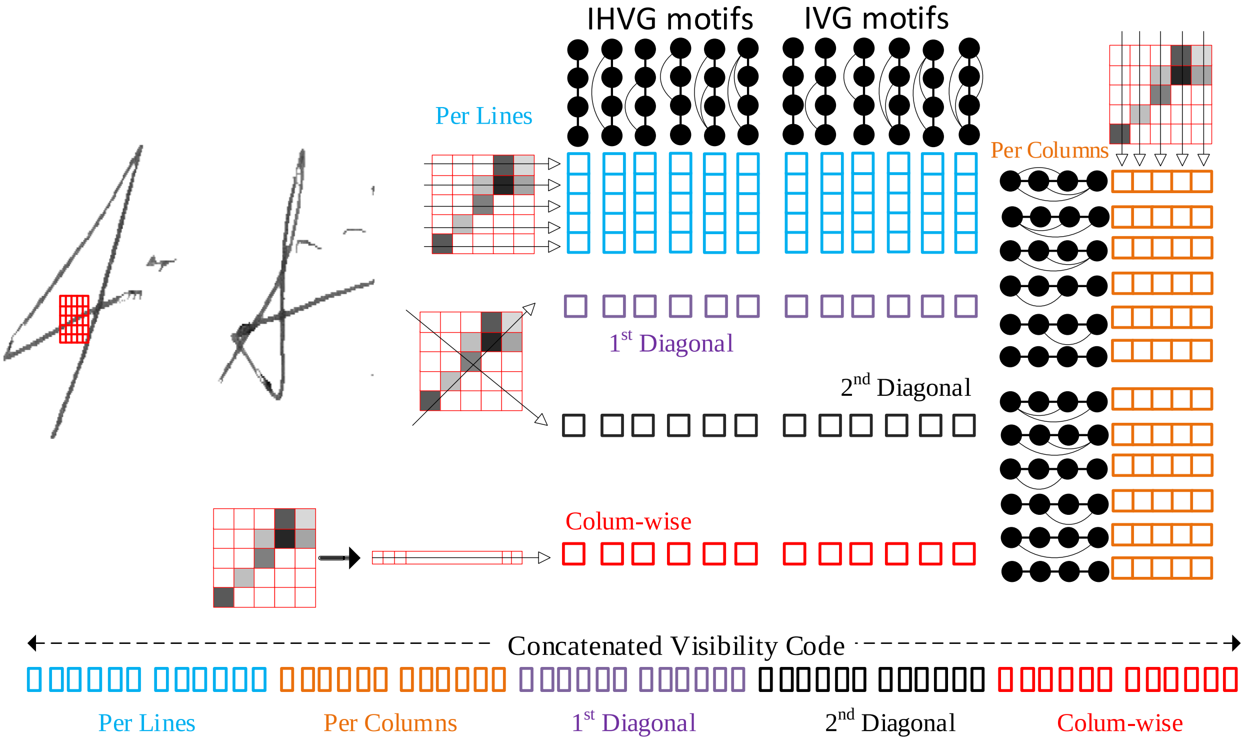

Figure 2-1: An example of calculating a local visibility code with 12 sequential motifs of size four. The local patches are extracted on the signature trace and each patch includes 25 pixels (5 × 5). The final visibility code (Concatenated Visibility Code) is the concatenation of the local IVG/IHVG motif coding procedures for each line, column, main and secondary diagonals, as well as a column-wise patch vector.

2.3.1.ii Datasets

Two popular offline signature datasets were used in order to demonstrate the effectiveness of the proposed system. The first one is CEDAR dataset with 55 writers and a total of forty-eight signature specimens (24 genuine and 24 simulated) while the skilled forgery signatures are composed from a mixture of random, simple and skilled forgeries [56]. The second signature dataset is MCYT75 with a total of 15 genuine and 15 simulated signature samples from 75 enrolled writers [57]. For this section, the thinning levels for the CEDAR and MCYT75 datasets have been set to one and two correspondingly. In addition, for the sake of simplicity results are reported when the number of equimass segments has been set to four (2×2) and sixteen (4×4) for the CEDAR and MCYT75 datasets respectively.

2.3.1.iii Classifiers

Writer dependent (WD) classifiers are trained for the offline Signature Verification (OffSV) problem. The number of genuine reference samples for each writer ($N_{REF} = 10$) has been set to ten for creating the positive class ω^+. In a similar way, a population of $N_{RF} = 30$ random forgeries (selected as random genuine samples from other signatories) create the corresponding negative class ω^-. For the local sequential motif features and any associative pooled Sequential Visibility feature Vector (SVV), the reference $SVV_{REF} \in \mathbb{R}^{10\times[60\times(\text{\#segments}+1)]}$ has been created to account for the genuine class while in a similar way, the random forgery visibility vector $SVV_{RF} \in \mathbb{R}^{30\times[60\times(\text{\#segments}+1)]}$ represents the negative class. These features are used as inputs to a binary, radial basis Support Vector Machine (SVM) classifier. A holdout cross-validation procedure returns the optimum operational parameters for the SVM margin and scale with respect to the maximum value of the Area under Curve (AUC). Moreover, the cross-validation procedure provides for each writer the scores conditioned on the positive only ω^+ class samples $CVS^{\oplus}$. The testing stage makes use of questioned (designated as: *Q*) samples that originate from: the remaining genuine signatures (14 for CEDAR, 5 for MCYT), the skilled forgeries (*S*: 24 for CEDAR and 15 for MCYT) and a number of 44 or 64 random forgeries (*R*) by taking one random sample from the remaining writers which does not participate at the training phase. Results are reported by means of the receiver operating characteristic (ROC) probabilities: the $p_{FAR(\mathrm{S})}$ and p_{FRR} error rates are computed as a function of a sliding threshold, whose extremes lie between the minimum and maximum values of the $CVS^{\oplus}$ cross validation procedure. Two different verification approaches are reported. In the first, a hard threshold is utilized to separate the genuine sample from skilled forgeries. This selection relies only on the ω^+ genuine reference samples as they are the only ones available for training the classifier. In a typical scenario, this hard threshold is set to *50%* of the average value of ω^+ scores. Additionally, the equal error rate per user threshold: *EER(S)*$_{\text{user-threshold}}$ is calculated as the point in which $p_{FAR(S)} = p_{FRR}$. The experiments were repeated ten times and their average values are reported. In addition, at this specific EER(S) threshold, the random forgery-(R) $p_{FAR(\mathrm{R})}$ error rate is calculated by using the genuine samples of the testing sets from the other writers.

2.3.1.iv Results

Table 2-1 displays the verification results achieved through various experimental protocols. For the CEDAR dataset, a Spatial Pyramid Pooling with 2 × 2 equimass segments, along with the whole image, was employed. As for the MCYT75 dataset, a Spatial Pyramid Pooling with 4 × 4 equimass segments, along with the whole image, was utilized. Also, this table includes the verification error rates when the vectors are partitioned in halves into their horizontal/natural parts. Table 2-2 presents a summary of results for the CEDAR and MCYT75 signature datasets, with other approaches found on the literature. It must be kept in mind that attaining a fair comparison between these results can be a very difficult task, because there are a number of factors that affect it during the classifier construction and evaluation. The reported results are either the Average or the Equal Error Rates (AER/EER) for the skilled forgeries (S) case. Ultimately, it could be argued that the proposed method achieves a low error of verification which is considered to be at least comparable to the ones derived from state-of-the-art methods.

Table 2-1: Verification Error Rates (%) for the CEDAR and MCYT75 datasets using different pooling strategies at the Sequential Visibility Vectors (SVV) that calculated via Horizontal, Natural and both image graph visibility motifs.

Pooling strategy	**CEDAR** **(*N_{REF} = 10*)**				**MCYT75** **(*N_{REF} = 10*)**			
	Hard Decision		**EER(S)**	**$P_{FAR(R)}$**	**Hard Decision**		**EER(S)**	**$P_{FAR(R)}$**
	$P_{FAR(S)}$	**P_{FRR}**	**user_thresh**	**@EER(S)**	**$P_{FAR(S)}$**	**P_{FRR}**	**user_thresh**	**@EER(S)**
Avg (SVV^H)	1.23	0.97	0.30	0.00	4.37	5.32	1.59	0.02
Max (SVV^H)	25.8	23.6	22.2	2.93	11.5	23.2	16.2	2.31
Std (SVV^H)	4.69	4.18	2.72	0.04	6.20	7.18	2.40	0.02
Avg (SVV^N)	2.89	2.14	1.01	0.02	6.00	6.95	2.03	0.02
Max (SVV^N)	21.1	19.8	19.4	2.15	16.7	13.9	13.8	2.19
Std (SVV^N)	4.53	4.02	2.01	0.03	8.61	8.09	2.81	0.03
Avg (SVV)	1.28	1.06	0.51	0.00	4.37	5.21	1.54	0.01
Max (SVV)	19.4	18.2	17.0	1.83	15.7	12.6	12.55	2.54
Std (SVV)	4.55	4.13	1.99	0.04	6.22	7.29	2.42	0.03
Avg & Std (SVV)	1.25	0.99	0.41	0.00	4.38	5.32	1.61	0.02

Table 2-2: Comparative summary of error rates (EER(S)%) including SOTA methods for OffSV.

CEDAR			**MCYT75**		
Method	$\mathcal{N}_{REF}$	**AER / EER**	**Method**	$\mathcal{N}_{REF}$	**AER / EER**
Gradient & concavity [58]	16	7.90	H.O.T. [35]	10	18.15
Zernike moments [58]	16	16.4	Global and Local Slant [59]	10	9.28
Partially Ordered Sets [60]	5	4.12	Partially Ordered Sets [60]	5	6.02
Gradient LBP + LRF [61]	16	3.52	L.B.P. [62]	10	7.08
Chain code [63]	12	7.84	Discrete Radon Transform [64]	10	9.87
B.O.W. with KAZE [65]	16	1.60	B.O.W. with KAZE [65]	10	6.4
V.L.A.D. with KAZE [66]		1.00			
Gradient Direction [67]	14	6.01	Contours [68]	10	6.44
Archetypes [69]	5	2.07	Archetypes [69]	5	3.97
Proposed: Average (SVV) [70]	10	0.51	**Proposed**: Average (SVV) [70]	10	1.54

2.3.2 Symmetric Positive Definite manifold (SPD)

2.3.2.i Method

Let $F \in \mathbb{R}^{w \times h \times n}$ be a feature map constructing from a stack of *n*-image planes that generated from one image I when a number of *n*-filters is applied: $F(x, y, i) = \Phi_i(I, x, y), i = 1{:}n$. Given a rectangular image region $\mathcal{R} \subset F$, let $\boldsymbol{f} = [\boldsymbol{f}_i]_{i=1,2,\ldots S} \in \mathbb{R}^{n \times S}$ be a local feature map of S total pixels that reside in $\mathcal{R}$. Then, the region $\mathcal{R}$ is modelled by its region covariance matrix $\boldsymbol{C}_{\mathcal{R}} \in \mathbb{R}^{n \times n}$ of the $\boldsymbol{f}_i \in \mathbb{R}^n$ points which is defined as follows:

$$\boldsymbol{C}_{\mathcal{R}} = \frac{1}{S-1} \sum_{i=1}^{S} (\boldsymbol{f}_i - \boldsymbol{\mu})(\boldsymbol{f}_i - \boldsymbol{\mu})^T \qquad \textit{eq. 2.6}$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$ represents the column mean vector of the $\boldsymbol{f}_i$ points and T denotes the transpose operator. Thus, the covariance matrix $\boldsymbol{C}_{\mathcal{R}}$ can be considered as a Symmetric Positive Definite (SPD) matrix. In cases where it does not meet the criteria of being SPD, it can be transformed into an SPD matrix by adding a small regularization term $\lambda \cdot \mathbf{I}_{nxn}$. All SPD matrices lie in the SPD manifold. The mapping $\Phi_i(I, x, y)$ of a signature image $I_{sign}(x, y)$ is defined as follows:

$$\left[I, I_x, I_y, I_{xx}, I_{xy}, I_{yy}, \sqrt{{I_x}^2 + {I_y}^2}, tan^{-1}(I_y/I_x), x_{n,}y_{n,} \right] \qquad \textit{eq. 2.7}$$

in which, I is the grayscale image after the preprocessing, $I_x, I_y, I_{xx}, I_{xy}, I_{yy}$ are image derivatives of $I(x, y)$, x_n, y_n are the signature pixel coordinates, normalized by their maximum number of rows and columns of the image bounding box and $\tan^{-1}(I_y/I_x)$ is the gradient direction, normalized into radians with range varying from [-π, π). The signature covariance matrix $\boldsymbol{C}_{SCM}$ of I is calculated using only the pixels that belong to the signature trace after the preprocessing. Therefore, any signature image results in a $\boldsymbol{C}_{SCM} \in \mathbb{R}^{n \times n}$ point of the corresponding Symmetric

Positive Definite (SPD) manifold. Using the equimass image division, different covariances matrices could be calculated according to relevant sub-regions.

Both writer-dependent and writer-independent verification approaches are designed. In the WD approach, the signature covariance matrix $\boldsymbol{C}_{SCM}$ is mapped on the tangent plane of a common pole $I_{n\times n}$ that creates a tangent vector with respect to a common tangential origin $\mathbf{0}_{n\times n} = \log_{\mathbf{I}_{n\times n}}(\mathbf{I}_{n\times n})$. The final feature vector representation for each signature arises by evaluating the orthonormal coordinates of the tangent vector in the common pole tangent space by a) applying the vector operator $\mathbf{v} = vec_{\mathbf{I}_{n\times n}}(\mathbf{y}) = (\mathbf{I}_{n\times n}^{-1/2}\mathbf{y}\mathbf{I}_{n\times n}^{-1/2})$ and b) selecting its *n(n+1)/2* components according to $[\mathbf{v}_{1,1}, \sqrt{2}\mathbf{v}_{1,2},..., \mathbf{v}_{\mathbf{2,2}}, \sqrt{2}\mathbf{v}_{2,3},.., \mathbf{v}_{n,n}]$. Next, discriminative predictors are learned using WD classifiers for the OffSV problem. In the WI approach, a metric similarity function is utilized to learn a mapping W from the original SPD manifold P_n to another SPD manifold $P_{m\cdot p}$ using the model $\Delta(\Theta)$, with parameters $\Theta = \{W, A, \mathrm{M}\}$, which explores diverse visual information that is stored in the $m-$individual block diagonal matrices $\in \mathbb{R}^{p\times p}$. The objective function is constructed based on contrastive loss, which incorporates similar pairs consisting of genuine signatures from the same writer, as well as dissimilar pairs comprising genuine signatures with random forgeries (i.e., genuine signatures of other writers in the dataset). Since the learnable parameters W, M lie in the Grassman and SPD manifolds respectively, the optimization procedure of the objective loss function is a non-jointly convex function of its learning parameters. However, it optimized with the stochastic gradient descent (SGD) given that $A \in \mathbb{R}^{m\times 2}$ and its update stage relies on the Euclidean gradient $\frac{\partial\mathcal{L}}{\partial A}$ while the Riemannian constraints of W, M impose the use of Riemannian gradients $\frac{\partial\mathcal{L}}{\partial \boldsymbol{W}^R}$ and $\frac{\partial\mathcal{L}}{\partial \boldsymbol{M}^R}$ [38].

2.3.2.ii Datasets

Four popular offline handwritten signature datasets of western and Indo-Aryan origin were employed in order to evaluate the proposed methods. In the Western style, the CEDAR [56] and MCYT75 [57] datasets were used, which are described in the previous section. For the Asian style, the BENGALI and HINDI, two subsets of the Indo-Aryan BHSig260 database [71], were utilized. The BHSig260 database comprises signatures in two regional languages and thus, it includes 100 signers for Bengali and 160 signers for Hindi. Each signer contributed 24 genuine and 30 forged images in both the Bengali and Hindi datasets.

2.3.2.iii Classifiers

In the WD approach, a binary, radial basis support vector machine (SVM) classifier is trained for each writer using N_{REF} genuine signatures of the writer and N_{RF} random forgeries selected from other writers in the dataset, as described in the previous section. The test set consists of the remaining genuine signatures of the writer, along with either skilled forgeries from the same writer or random forgeries from all other writers in the dataset. The results are reported by means of the average value (10 repetitions) of two corresponding Equal Error Rates (EER) with two user-dependent sliding thresholds. The first EER(S) measures the probability of rejecting genuine samples p_{FRR} against the probability of accepting skilled forgery samples $p_{FAR(S)}$ and the second EER(RF) measures the probability of rejecting genuine samples p_{FRR} against the probability of accepting random forgery samples $p_{FAR(RF)}$. To address the limited training data problem, feature points augmentation is performed using fixed Riemannian Gaussian Distributions (RGD) on the space of SPD matrices, with their maximum likelihood estimators evaluated from the original samples in the training set, which includes $N_{REF} = 3$ genuine signatures.

In the WI approach, a threshold is defined within the SPD manifold for each writer to differentiate between genuine and forgery signatures. This user-specific threshold is calculated by comparing each questioned signature with ten genuine reference samples from the same writer. Instead of using a single global $\boldsymbol{C}_{SCM}^{1\times1}$ covariance matrix, a set of four $\boldsymbol{C}_{SCM}^{\{2\times2\}}$ and nine $\boldsymbol{C}_{SCM}^{\{3\times3\}}$ covariance matrices are utilized. Thus, each image is represented by a set of fourteen covariance matrices according to fourteen regions. These fourteen local scores are sorted and then a local score is calculated by averaging. The final signature verification score is the minimum value of all scores between the questioned signature and all reference samples, which determines whether the signature belongs to the genuine or forgery side. Additionally, two experimental protocols are used to evaluate the performance. The first $\mathcal{F}_{intra}$ protocol follows a standard 5×2 blind-fold approach for each individual dataset. In this protocol, the writers in each dataset are divided into two equal sets in five different ways. One set is used for training the model $\Delta(\Theta)$, while the other set is used as the test set. The evaluation results are presented as the average across the five repetitions, providing an overall assessment of the model's performance. The second $\mathcal{F}_{inter}$ protocol resembles a transfer learning approach, which involves cross-dataset evaluation, meaning that the model $\Delta(\Theta)$ trained on one dataset is tested on another dataset.

2.3.2.iv Results

The results obtained using the WD approach are presented in Table 2-3. To generate duplicate samples for the genuine training class, the maximum likelihood estimate (MLE) of the parameters of the Riemannian Gaussian Distribution (RGD) is calculated using three reference samples. An SPD RGD model is then constructed using these parameters to draw SPD duplicates. The duplicates are created by sampling directly from three individual RGD distributions $G\left(\boldsymbol{C}_i, \hat{\sigma}(\omega_{3G}^{L+})\right)$ where $\boldsymbol{C}_i$ is one reference sample, or from three RDGs with fixed standard deviations σ_{fixed} of 0.1 or 0.01. This process generates five duplicates for each genuine reference signature. In these experiments, the I_{xy} image derivative is not included in the mapping $\Phi_i(I, x, y)$ to simplify the calculations. Although a comprehensive comparison with other methods from the literature is not possible due to differences in design and implementation, the proposed system achieved a low verification error that is comparable to related works, as demonstrated in Table 2-3. However, a limitation of this approach is that it is sensitive to setting hyperparameters, such as σ, which can significantly impact the results if not properly configured.

Table 2-3: Comparative summary of error rates (EER(S)% and EER(RF)%) including generative SOTA methods for WD OffSV.

OffSV System		**Training Set**		**CEDAR**		**BENGALI**		**HINDI**	
Method	**Augmentation method**	**(+) class (*N_{REF}*)**	**(-) class**	**EER(S)**	**EER(RF)**	**EER(S)**	**EER(RF)**	**EER(S)**	**EER(RF)**
Motifs [70]	-	10G	30RF	0.51	-	0.32	-	1.02	-
Deformations [72]	-	5	N/A	3.89	-	8.92	-	9.84	-
Synthetic signatures [73] [74]	Duplicator	3G +66D	648RF +14kD	3.04	-	6.06 (5G)	-	-	-
Feature augment [74]	Gaussian Noise			0.82	-	-	-	-	-
Generative [75]	GAN	5G	N/A	4.50	-	-	-	-	-
Proposed: SPD [76]	-	10G	50RF	0.49	0.03	0.27	0.09	1.00	0.30
	-	3G	30RF	1.18	0.14	1.52	0.41	2.50	0.69
	RGD: G(C_i,0.1)	3G +15D	753RF	1.73	0.18	5.22	1.80	2.13	0.63
	RGD: G(C_i, $\hat{\sigma}(\omega_{3G}^{L+})$)			0.74	0.15	1.41	0.53	1.82	0.57
	RGD: G(C_i,0.01)			0.55	0.11	0.92	0.25	1.62	0.45

In the WI approach, the model $\Delta(\Theta)$ has two hyperparameters, $m \; and \; p$. The parameter m determines the number of block-diagonal SPD matrices $\boldsymbol{X}^{k=1:m,}$ and, consequently, the number of sub-distances $\{D_{\boldsymbol{A}}^{k}(\cdot,\cdot)\}_{k=1}^{m}$*. The parameter* p *determines the size of each block diagonal SPD matrix* $\boldsymbol{X}^{k} \in \mathbb{R}^{p \times p}$ used in the evaluation of any $D_{\boldsymbol{A}}^{k}$. After experimental investigation, it was found that setting $m \cdot p = 10$ leads to optimal performance due to the relatively small initial dimensionality of the P_{10} manifold. Therefore, m is set to 1 and p is set to 10 and the projection matrix W reorders the samples within the same dimensional space. The number of similar and dissimilar pairs are equal and is determined by the maximum number of created similar pairs. The number of participating segments depends on the dataset being evaluated. For the CEDAR signature dataset, there are four participating segments using the Spatial Pyramid equimass division into fourteen segments (1, 2×2, and 3×3 pyramid levels), while for the MCYT75, HINDI, and BENGALI datasets, there are seven participating segments from the total of fourteen segments on all reported results. Table 2-4 presents the EER(S) values with user-defined thresholds for $\mathcal{F}_{intra}$ (Table's diagonal) and $\mathcal{F}_{inter}$ (Table's non-diagonal) protocols. The results presented in this table demonstrate the robust performance of our method. Table 2-5 summarizes the state-of-the-art results in term of EER(S) for WI-SV systems, showcasing the comparable performance of the proposed method with other works.

Table 2-4: EER(S) values for intra-dataset (diagonal values) and inter-dataset (non-diagonal values) experimental protocols.

Train / Test	CEDAR	MCYT75	BENGALI	HINDI
CEDAR	0.37	0.95	0.24	0.75
MCYT75	0.36	0.96	0.27	0.75
BENGALI	0.35	0.96	0.26	0.75
HINDI	0.35	0.97	0.27	0.77

Table 2-5: Comparative summary of error rates (EER(S)%) including SOTA methods for WI OffSV.

Method	Protocol (training writers / test writers)	$\mathcal{N}_{REF}$	CEDAR	MCYT75	BENGALI	HINDI
Triplet Nets Graph edit dist.[77]	16/8 signatures per writer	10	5.91	3.91	-	-
MSDN [78]	50/55	10	1.75	-	-	-
Point2Set DML [79]	$\mathcal{F}_{inter}$	5	5.22	4.86	-	-
DCCM [80]	10/45	5	2.10	-	-	-
Partially oriented sets [60]	12/43 for CEDAR 50/25 for MCYT75	5	2.90	3.50	-	-
CNN & BiLSTM [81]	$\mathcal{F}_{intra}$	N/A	0.00	-	1.76	2.23
SigNet-F Dichotomy Transformation [82]	$\mathcal{F}_{intra}$	12	5.86	2.99	-	-
	$\mathcal{F}_{inter}$	12	4.21	4.22	-	-
SigNet-Contrastive (trained on GPDS [83])	12/10 signatures per writer for CEDAR 10/5 signatures per writer for MCYT75		3.34 (12G)	3.52 (10G)	-	-
SigNet-Contrastive (trained on GPDS-S [84])			4.59 (12G)	3.95 (10G)	-	-
V.L.A.D. [85]	50/55 for CEDAR 50/100 for BENGALI 100/160 for HINDI (SF used for Training)	N/A	0.00	-	9.62	20.2
AVN [86]		1	3.77	-	6.14	5.65
Deep HSV [87]		1	0.00	-	11.9	13.3
IDN [88]		1	3.62	-	4.68	6.96
Proposed: SPD-WI	$\mathcal{F}_{intra}$	10	0.37	0.96	0.26	0.77

2.3.3 Sparse Representation (SR)

2.3.3.i Method

Sparse Representation (SR) is a method used in signal processing to represent data as a sparse linear combination of a set of basis vectors. It assumes that the data can be expressed using only a few elements (sparsity) from a given dictionary of basis vectors (atoms). The Sparse Representation problem combines both the dictionary learning step, which learns the dictionary $\boldsymbol{D}$ from the training data, and the sparse coding step, which finds the sparse coefficients $\boldsymbol{A}$ to represent the data using the dictionary. Many methods exploit the assumption that natural images can be represented by a sparse combination of basis vectors in a redundant dictionary [39], [40], [89], [90]. While these methods often succeed in enhancing discriminative power, the optimization process for dictionary learning is primarily generative and does not explicitly consider the discrimination task.

For a given set of *M* training signals $\mathbf{X} = [\mathbf{x}^1, \mathbf{x}^{2,} \ldots \mathbf{x}^M] \in \mathrm{R}^{\mathrm{n \times M}}$, the problem of dictionary learning can be formulated using different variants of minimization problem:

such as when the sparsity-inducing regularization function $\boldsymbol{\psi}(\cdot)$ is used as a penalty:

$$\min_{\boldsymbol{D} \in \boldsymbol{C}, \boldsymbol{A} \in \mathbb{R}^{p \times n}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{a}_i\|_2^2 \ + \ \lambda \boldsymbol{\psi}(\boldsymbol{a}_i) \qquad \textit{eq. 2.8}$$

and when the sparsity-inducing regularization function $\boldsymbol{\psi}(\cdot)$ is used as a constraint:

$$\min_{\boldsymbol{D} \in \boldsymbol{C}, \boldsymbol{A} \in \mathbb{R}^{p \times n}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{a}_i\|_2^2 \quad \boldsymbol{s.t.} \quad \boldsymbol{\psi}(\boldsymbol{a}_i) \leq \boldsymbol{\mu} \quad \textit{eq. 2.9}$$

or equivalent

$$\min_{\boldsymbol{D} \in \boldsymbol{C}, \boldsymbol{A} \in \mathbb{R}^{p \times n}} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{a}_i) \quad \boldsymbol{s.t.} \quad \|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{a}_i\|_2^2 \leq \boldsymbol{\varepsilon} \quad \textit{eq. 2.10}$$

where the regularization function $\boldsymbol{\psi}(\cdot)$ is the *ℓ-p* norm for 1≤*p*≤∞, while most popular forms are either $\|\boldsymbol{a}\|_0 \ \boldsymbol{or} \ \|\boldsymbol{a}\|_1$ using *ℓ-0* norm and *ℓ-1* norm respectively, that could encourage sparse solutions [39], [40]. Also, the columns of dictionary matrix $\boldsymbol{D}$ are constrained to have less than unit *ℓ-2* norm to avoid trivial solutions and mitigate the issue of large atoms' values leading to arbitrarily small sparce coefficients [39], [40]. Given that the above minimization problem is non-convex, many empirical optimization methods have proven effective in practical applications. The most classical approach for dictionary learning is the alternate minimization scheme, which alternates between two minimization steps: updating the dictionary with fixed sparse codes and updating the sparse codes with a fixed dictionary. Although not as fast as well-tuned stochastic gradient descent algorithms, the alternate minimization scheme is parameter-free and has shown reliability in computer vision tasks [39], [40].

This work employs a writer dependent (WD) signature verification approach to discover the underlying structure in image patches using Sparse Representation (SR). Specifically, patches are extracted from the reference signatures of each writer and utilized as the training signals in the dictionary learning step. Subsequently, the dictionary of the claimed writer is used to encode every patch from the query signature, resulting in the calculation of sparse codes for the image. Different optimization approaches for solving the SR problem have been investigated, utilizing either batch learning with ℓ-0 norm regularization [48] or online learning with ℓ-1 norm regularization [49]. Depending on the case, both the K-SVD/OMP and the SPAMS/LARS-Lasso algorithms are evaluated in the OffSV task [91]. Therefore, greedy algorithms (like the Orthogonal Matching Pursuit – OMP) deal with NP-hard problems arise from the ℓ-0 norm constrain in sparse coding (eq. 2.9 and eq. 2.10) and homotopy algorithms (like the Least Angle Regression – LARS) solve the ℓ-1 regularized problems with a penalty parameter (eq. 2.8). The resulting sparse codes from any signature image are pooled together using spatial pyramid pooling (SPP) techniques. The SPP techniques are further enhanced by incorporating an additional set of patches that correspond to points of interest as detected by the Binary Robust Invariant Scalable Keypoints (BRISK) detector [92]. The keypoints serve as indicators for the patches, and thereafter the corresponding sparse codes are pooled together in order to obtain an additional feature vector. This vector is concatenated with the spatial pyramid vector resulting to a final feature vector. This approach enables the capture of additional local information via selective attention to specific signature's points of interest.

The typical SR has been extended in two directions, one that uses a structured sparse regularization to learn a dictionary embedded in a tree and one that uses a deeper multi-layer architecture to learn several levels of SR in different abstraction levels. First, Hierarchical Sparse Coding (HSC) introduces the idea of embedding dictionary atoms in a rooted and directed tree structure [93]. Since the *ℓ-1* norm cannot model interactions between atoms, it is replaced by a more complex sparsity-inducing penalty that considers a particular tree structure of dictionary elements [93]. The tree structures are pre-defined, and the dictionary elements naturally organize themselves in the tree during the learning process. The hierarchical penalty combines the Group Lasso with the *ℓ-1* norm, promoting sparsity patterns rooted in subtrees. Consequently, a dictionary element can only be utilized in patch decomposition if its parent in the tree is also used, encouraging structured sparsity patterns. Second, Deep Sparse Coding (DSC) utilizes multiple layers and incorporates a sparse-to-dense module between these layers [41]. Unlike simply stacking SR layers, DSC accounts for the spatial information of image patches and considers that small variations in the original space might lead to significant differences in the corresponding sparse codes. To address these issues, an average pooling function and a dimensionality reduction (DR) method are applied between SR layers. Local spatial pooling ensures that higher-level features are learned from collections of nearby lower-level features, while dimensionality reduction prevents an unnecessary increase in feature dimension without gaining new information. In this way, the pooled sparse codes from the previous layer serve as high-dimensional vectors, while the dimensionality-reduced latent vectors become inputs for subsequent SR layer. Both HSC, with its pre-defined hierarchical tree structure, and DSC, with its layer-wise organization, enable hierarchical learning. These approaches facilitate the capture of increasingly abstract and complex representations of the data. These concepts are effectively leveraged in the OffSV task to enhance the discriminative power of the learned features [94], [95].

2.3.3.ii Datasets

Four popular offline handwritten signature datasets of western and Persian origin were employed in order to evaluate the proposed methods. In the Western style, the CEDAR [56] and MCYT75 [57] datasets were used and described in the previous section while another one dataset that included in the experiments is GPDS300GRAY [62]. This third signature dataset contains 24 genuine signatures and 30 simulated forgeries of 300 individuals stored in an 8-bit, grey level format. An interesting characteristic of this dataset is that the acquisition of signature specimens is carried out using two different bounding boxes of size 1.8 × 5.0 cm and 2.5 × 4.5 cm respectively (height × width). As a result, the images of this dataset have two different aspect ratios. The fourth signature dataset used is the Persian UTSIG, created by Soleimani et al. [96]. It contains specimens from 115 writers where each one has 27 genuine signatures, 3 opposite-hand signatures, and 42 skilled forgeries made by 6 forgers. Notably, the acquisition of signature specimens in the UTSIG dataset was performed using six different bounding box sizes, simulating real-world conditions and public services application forms.

2.3.3.iii Classifiers

The writer dependent (WD) classifiers stage is implemented similarly to the experiments with Image Visibility Graph motifs (IVG) described previously. In this stage, the training set of each writer's classifier comprises both genuine signatures of the writer and randomly selected forgeries from other writers in the dataset. The number of training random forgeries is twice the number of training genuine signatures, i.e., $N_{RF} = 2 \cdot N_{REF}$. The test set, on the other hand, consists of the remaining genuine signatures of the writer along with skilled forgeries corresponding to the same writer. This approach allows for evaluating the performance of the WD classifiers with a user-defined threshold.

2.3.3.iv Results

The dictionary size (number of atoms) is set to 60, with a sparsity level of $\mu = 3$ for the K-SVD/OMP algorithms and a penalty parameter of $\lambda = 0.15$ for the SPAMS/LARS-Lasso algorithms, determined through a grid search. In Table 2-6, the impact of different image division scenarios on verification performance is investigated specifically for the CEDAR dataset, using ℓ-0 sparsity regularization for the SR problem. Six division scenarios are examined, including: a) using the entire image (first level of SPP), b) employing only the second level of SPP with four equimass regions (2 × 2), c) using the full SPP, d) considering only the BRISK keypoints, e) a combination of all previous scenarios, and f) a combination of SPP with nine equimass regions (3 x 3) and BRISK keypoints. It is observed that the combination of all scenarios together outperforms the others, irrespective of the pooling function. However, increasing the number of equimass regions in the second level of the SPP technique is not always the optimal solution. Therefore, subsequent tables present results for both 2 × 2 and 3 × 3 equimass regions in the SPP, and the SR problem is optimized using either ℓ-0 sparsity regularization with K-SVD/OMP algorithms or ℓ-1 sparsity regularization with SPAMS/LARS-Lasso algorithms.

Observing the results in

Table 2-7 and Table 2-8, several interesting conclusions can be drawn. Firstly, the standard deviation pooling function (F3) consistently outperforms other pooling functions in most cases. It is notable that although max pooling performs well in other computer vision applications, it is not as effective for signature images. This could be attributed to the unique nature of signature images, which exhibit a degenerate structure. Due to the limited set of structural elements shared by all signature images, the sparse coefficients of different signatures may not differ significantly in terms of first-order statistics, as observed in more complex image structures. Therefore, higher-order statistics are likely to provide better discrimination between the distributions of sparse coefficients, resulting in improved verification performance. Secondly, the verification error rates for the CEDAR and MCYT datasets do not show substantial variation when using either ℓ-0 or ℓ-1 form for the SR problem. Theoretically, for an appropriate dictionary, the solution obtained using the ℓ-1 norm is equivalent to an ℓ-0 norm solution with full probability [97]–[99]. Nevertheless, his equivalence does not hold for the joint regularized and constrained dictionary learning problem, as different formulations of the problem may lead to different solutions, even when using the same ℓ-1 norm [40]. However, due to the similar performance achieved using both approaches, only the results obtained from the ℓ-0 SR problem are presented in the Table 2-9 for the GPDS300GRAY and UTSIG datasets.

Thirdly, the experimental results indicate that the optimal performance on the CEDAR dataset is achieved with a spatial pyramid consisting of 2 × 2 equimass regions, while for the MCYT75 dataset, the best results are obtained with a spatial pyramid of 3 × 3 equimass regions. This observation can be attributed to the difference in scanning resolutions, with CEDAR signatures having a resolution of 300dpi and MCYT75 signatures having a resolution of 600dpi, resulting in higher pixel density in MCYT75 segments. It is worth noting that even with a spatial pyramid of 2 × 2 on the CEDAR dataset, there are some patches within certain segments that do not contribute significantly to discrimination, which is not observed in the case of the MCYT75 dataset. Additionally, increasing the spatial pyramid size from two to three has a positive impact on performance when using the UTSIG dataset, but it has a negative impact on the GPDS300GRAY dataset. Since the GPDS300GRAY dataset utilized only two acquisition bounding boxes and UTSIG utilized six, designing more regions seems to be beneficial when many bounding boxes are employed, despite the spatial pyramid equimass division does not consider the different aspect ratios of the acquisition bounding boxes.

Table 2-6: EER(S) values for different image division profiles using CEDAR dataset.

CEDAR $\mathcal{N}_{REF}$ **= 5**	***ℓ*-0 sparsity regularization:** **K-SVD/OMP**					
Pooling	**SPP:** **1**	**SPP:** **2 × 2**	**SPP:** **1 & 2 × 2**	**keypoints only**	**SPP: 1 & 2 × 2 and keypoints**	**SPP: 1 & 3 × 3 and keypoints**
	Feature dim = 60	**Feature dim = 240**	**Feature dim = 300**	**Feature dim = 60**	**Feature dim = 360**	**Feature dim = 660**
Avg	8.87	4.10	2.78	8.35	2.67	2.80
Max	10.3	5.73	6.30	10.3	4.17	3.21
Std	4.36	2.98	1.95	4.30	**1.44**	1.80
Norm	9.37	9.92	8.02	9.22	7.51	13.4
L-2	8.20	4.38	3.04	8.47	3.08	2.48

Table 2-7: Performance evaluation using Sparse Representation methods for CEDAR dataset.

CEDAR $\mathcal{N}_{REF}$ **= 5**	***ℓ-0*** **sparsity regularization:** **K-SVD/OMP**				***ℓ-1*** **sparsity regularization:** **SPAMS/LARS-Lasso**			
Pooling	**Hard Decision**		**EER(S)**	**$P_{FAR(R)}$**	**Hard Decision**		**EER(S)**	**$P_{FAR(R)}$**
	$P_{FAR(S)}$	**P_{FRR}**	**user_thresh**	**@EER(S)**	**$P_{FAR(S)}$**	**P_{FRR}**	**user_thresh**	**@EER(S)**
	SPP 1 & 2 × 2 and keypoints: final feature dimensionality is 360							
Avg	6.83	7.32	2.67	0.43	6.99	7.60	2.65	0.37
Max	9.45	9.83	4.17	2.93	9.42	9.71	3.45	1.55
Std	4.76	4.91	**1.44**	0.17	4.61	4.64	**1.62**	0.12
Norm	8.12	9.94	7.51	1.95	8.19	8.73	3.52	0.56
L-2	5.31	5.27	3.08	0.35	7.05	7.81	3.38	0.41
	SPP 1 & 3 × 3 and keypoints: final feature dimensionality is 660							
Avg	6.13	7.18	2.80	0.27	6.59	6.61	3.08	0.23
Max	8.83	8.65	3.21	0.76	8.52	8.01	2.80	0.48
Std	4.78	5.63	**1.80**	0.12	4.88	5.81	**2.01**	0.08
Norm	12.7	20.2	13.4	4.82	9.17	8.71	3.97	0.41
L-2	5.05	5.11	2.48	0.24	5.49	6.48	3.07	0.22

Table 2-8: Performance evaluation using Sparse Representation methods for MCYT75 dataset.

MCYT75 $\mathcal{N}_{REF} = 5$	***ℓ-0*** **sparsity regularization: K-SVD/OMP**				***ℓ-1*** **sparsity regularization: SPAMS/LARS-Lasso**			
Pooling	**Hard Decision**		**EER(S)**	**$P_{FAR(R)}$**	**Hard Decision**		**EER(S)**	**$P_{FAR(R)}$**
	$P_{FAR(S)}$	**P_{FRR}**	**user_thresh**	**@EER(S)**	**$P_{FAR(S)}$**	**P_{FRR}**	**user_thresh**	**@EER(S)**
	SPP 1 & 2 × 2 and keypoints: final feature dimensionality is 360							
Avg	10.4	7.32	3.80	0.22	10.46	8.57	3.91	0.24
Max	14.7	14.6	10.9	3.16	15.8	14.3	11.20	3.09
Std	8.09	6.75	**3.18**	0.16	8.35	7.58	**3.71**	0.22
Norm	12.99	16.17	10.8	3.08	12.2	13.43	3.93	0.30
L-2	7.64	7.76	3.23	0.14	7.87	7.59	3.66	0.20
	SPP 1 & 3 × 3 and keypoints: final feature dimensionality is 660							
Avg	9.18	7.12	3.19	0.14	7.82	7.68	3.58	0.14
Max	15.6	11.5	8.65	1.12	13.9	12.3	8.60	1.08
Std	8.07	6.21	**2.82**	0.07	7.59	7.29	**3.40**	0.10
Norm	11.0	23.7	16.7	6.17	9.82	9.68	3.53	0.15
L-2	7.77	7.92	3.46	0.13	7.07	7.69	3.05	0.12

Table 2-9: Performance evaluation using Sparse Representation for GPDS300GRAY and UTSIG datasets.

$\mathcal{N}_{REF} = 5$	**GPDS300GRAY**				**UTSIG**			
	ℓ-0 **sparsity regularization: K-SVD/OMP**							
Pooling	**Hard Decision**		**EER(S)**	**$P_{FAR(R)}$**	**Hard Decision**		**EER(S)**	**$P_{FAR(R)}$**
	$P_{FAR(S)}$	**P_{FRR}**	**user_thresh**	**@EER(S)**	**$P_{FAR(S)}$**	**P_{FRR}**	**user_thresh**	**@EER(S)**
	SPP 1 & 2 × 2 and keypoints: final feature dimensionality is 360							
Avg	6.91	6.22	2.47	0.59	18.3	16.9	11.7	1.91
Max	8.27	8.89	3.98	3.70	25.3	21.6	17.0	7.02
Std	6.73	6.13	**1.50**	0.33	16.1	13.4	**9.94**	1.27
Norm	16.1	19.3	15.2	6.23	24.5	27.6	21.1	6.39
L-2	7.79	7.96	3.30	0.55	20.7	24.6	12.2	1.48
	SPP 1 & 3 × 3 and keypoints: final feature dimensionality is 660							
Avg	6.46	6.47	3.14	0.44	16.9	12.4	9.82	1.94
Max	11.79	9.37	4.43	1.82	23.6	15.7	13.35	3.33
Std	5.01	5.92	**1.97**	0.19	15.8	12.5	**8.56**	1.25
Norm	18.6	29.4	20.4	10.0	29.8	29.5	25.15	12.5
L-2	7.58	6.83	3.53	0.30	17.3	15.2	10.48	0.92

Table 2-10 presents additional experiments involving Sparse Representation (SR) with different preprocessing and pooling options, as well as results from Hierarchical Sparse Coding (HSC) and Deep Sparse Coding (DSC) methods. It also includes some popular Deep Learning (DL) methods for OffSV. The performance of SR methods is found to be sensitive to user-defined hyperparameters in the preprocessing and Spatial Pyramid Pooling (SPP) stages, such as the thinning level and the choice of pooling function or number of pyramid segments. Both HSC and

DSC approaches, despite their complex implementations, did not achieve superior performance compared to the baseline SR method. This may be attributed to the limitations of layer-by-layer optimization and the need for setting hyperparameters specifically for the OffSV task, such as branch size in HSC or the dimensionality reduction method in DSC. However, DSC serves as an introduction to deep learning architectures by utilizing similar optimization solvers, such as Stochastic Gradient Descent, when employing Dimensionality Reduction by Learning an Invariant Mapping (DrLIM) [28]. Additionally, the layer-by-layer optimization using pairs in DSC requires fewer initial training images compared to CNN-based writer identification tasks in other DL works presented in the table. In conclusion, comparing DL and SR methods directly in the context of OffSV is challenging due to the differences in system design. However, it is worth noting that SR methods, like the one employed here, can be effective in the OffSV problem but require careful tuning of numerous hyperparameters (from preprocessing to feature extraction and classifiers stages) for optimal performance.

Table 2-10: Comparative summary of error rates (EER(S)%) including Deep Learning methods and extension of Sparse Representation for OffSV.

Method		N_{REF}	**CEDAR**	**MCYT75**	**GPDS300 GRAY**	**UTSIG**
Name	**Description**					
HOCCNN [100]	Combine low-level and high-level features into a hierarchical CNN	12	4.94	5.46 (10G)	-	12.88 (12G)
Deformations CNN [72]	Location information as a feature for capturing micro deformations	12	2.76	-	-	6.14
MLSE [101]	Multi-Loss Snapshot Ensemble objective function	10	-	2.93	-	6.17
SigNet [102]	Training with genuine signatures	12	4.76	2.87 (10G)	3.15	-
		5	5.87 (4G)	3.58	3.92	-
SigNet-F [102]	Training with genuine and skilled forgery signatures	12	4.63	3.00 (10G)	1.69	-
		5	5.92 (4G)	3.70	2.42	-
SigNet-SPP [103]	Training with signatures of different sizes	10	3.60	3.64	3.15	-
SigNet-SPP (fine-tuned) [103]		10	2.33	3.40	-	-
Proposed: SR [91]	SPP: 1 & 2 × 2 and keypoints (std)	10	0.90	1.63	0.76 (12G)	7.36 (12G)
	SPP: 1 & 2 × 2 and keypoints (std)	5	1.44	3.18	1.50	9.94
	SPP: 1 & 3 × 3 and keypoints (std)	10	1.14	1.97	0.92 (12G)	6.22 (12G)
	SPP: 1 & 3 × 3 and keypoints (std)	5	1.80	2.82	1.97	8.56

Method		N_{REF}	CEDAR	MCYT75	GPDS300 GRAY	UTSIG
Name	Description					
	Preprocessing: Thinning level per signature and SPP: 1 & 2 × 2 and keypoints (std)	5	2.89	4.98	5.71	10.3
	Preprocessing: Thinning level per signature & SPP: 1 & 2 × 2 and keypoints (avg)	5	3.45	4.89	6.23	12.5
Proposed: SR [104]	SPP: 1 & 2 × 2 (avg)	5	2.78	3.67	2.70	-
Proposed: HSC [94]	Depth: 3, Branch: [20 2] SPP: 1 & 2 × 2	5	2.30	4.01	-	-
	Depth: 3, Branch: [20 2] SPP: 1 & 3 × 3		2.42	3.52	-	-
	Depth: 4, Branch: [6 3 2] SPP: 1 & 2 × 2		2.56	4.29	-	-
	Depth: 4, Branch: [6 3 2] SPP: 1 & 3 × 3		2.72	3.70	-	-
Proposed: DSC [95]	Layers: 2, DR: PCA SPP: 1 & 2 × 2	5	3.98	-	-	-
	Layers: 2, DR: DrLIM SPP: 1 & 2 × 2		3.05	-	-	-
	Layers: 2, DR: Random projection, SPP: 1 & 2 × 2		2.85	-	-	-
	Layers: 2, DR: Random Ortho proj. SPP: 1 & 2 × 2		2.82	-	-	-
	Layers: 3, DR: PCA SPP: 1 & 2 × 2		5.12	-	-	-
	Layers: 3, DR: DrLIM SPP: 1 & 2 × 2		4.93	-	-	-
	Layers: 3, DR: Random projection SPP: 1 & 2 × 2		3.30	-	-	-
	Layers: 3, DR: Random Ortho proj. SPP: 1 & 2 × 2		2.87	-	-	-

2.4 Conclusions

The shallow representation models discussed in this chapter provide a transparent and interpretable way to extract features from signature images. The image encoding approaches using image visibility graph motifs, SPD manifold mapping, and sparse representation offer insights into the underlying characteristics of the signatures. Although these models demonstrate state-of-the-art performance in capturing relevant features from signature images, they require meticulous tuning of numerous hyperparameters to achieve optimal results. However, the major advantage of these models lies in their efficiency even with a limited number of training samples. This is due to their shallow architectures, which provide low-level features and require training procedures only for the final classifiers.

On the other hand, deep learning methods have demonstrated exceptional performance in various computer vision tasks, including signature verification. By leveraging convolutional neural network architectures, deep learning models can automatically learn hierarchical representations from data, eliminating the need for handcrafted feature engineering and explicit hyperparameter tuning. The end-to-end training process allows CNN models to capture intricate patterns and nuances in images, potentially leading to higher accuracy and robustness. However, the success of deep learning models heavily relies on having a large and diverse dataset for training, which can be challenging to acquire in certain domains.

Considering these aspects, a promising direction for research lies in exploring hybrid approaches that combine the strengths of both shallow representation methods and deep learning models. This approach is presented in the next chapter, where hand-crafted image descriptors are combined with CNNs.

Chapter 3

Hybrid combination of shallow representations and deep learning

3.1 Introduction

Hand-crafted features have been widely employed in computer vision problems, mainly for the task of image classification [42], [105], [106]. These features are derived from non-learning processes by directly applying various operators on image pixels and can provide several properties, like rotation and scale invariance [42], [107], due to their ability to efficiently encode local gradient information. However, hand-crafted features suffer from three primary limitations. First, hand-crafted features extract a low-level representation of the data, lacking the ability to provide a prominent abstract representation necessary for recognition tasks [108]–[110]. Secondly, the local descriptors like SIFT (Scale-Invariant Feature Transform) do not offer a fixed-length vector representation of the input image, necessitating additional logic for the local descriptor encoding [109], [111], [112]. Thirdly, hand-crafted features have limited capacity as they are predetermined mappings from the data to the feature space, which remains fixed regardless of the requirements of specific recognition problems.

In the past decade, hand-crafted based methods have been replaced by Deep Convolutional Neural Networks (CNNs), which employ an end-to-end learning scheme, typically in a supervised manner [113]. CNNs operate by associating each input image with a ground-truth label specific to the computer vision task at hand. The predictive model of the CNN produces a score, which is compared to the corresponding label, and the model weights are adjusted until the output achieves an acceptable level of accuracy. Consequently, CNNs build a hierarchical organized feature representation of the input data through a learning process that minimizes a given (differentiable) cost function. Thus, the CNNs learn both the feature representation and the feature encoding directly from images. The result is a learnable model that can provide high-level feature representations of input data once trained on a particular dataset and task. However, CNNs have certain limitations. They demand extensive amounts of data for training and are highly dependent on the quality of the data, including the availability of corresponding labels. After all, training deep architecture presents challenges, such as the requirement for large annotated datasets and difficulties in ensuring scale, rotation, or geometric invariance properties [114].

This chapter explores the combination of local descriptor representations with deep learning architectures. Our objective is to assess the ability of local descriptors to provide higher-level information to CNNs, thereby improving their performance in handling rotations, complex textures, and patterns. Initially, we calculate SIFT descriptors on a dense grid of image locations, considering the neighborhoods of all pixels in the image. The center pixel of each image neighborhood is mapped to a histogram, creating a new image representation called the SIFT-Image [43]. The spatial resolution of the SIFT-Image can match that of the input image (unless subsampling is applied with a stride greater than one), and the depth of the SIFT-Image corresponds to the dimensionality of the SIFT descriptor. The new image representation is used as input to CNN, resulting in a framework called SIFT-CNN. Thus, the proposed SIFT-CNN consists of two stages: the unsupervised calculation of the dense SIFT descriptors to generate the local descriptor representation [115], followed by utilizing the produced SIFT-Images as inputs for the supervised training of a CNN model in a classification task. Therefore, in the proposed SIFT-CNN framework, we undergo a transformation of the input space representation to incorporate desirable properties.

The SIFT-CNN incorporates local scale and local rotation invariance property and hence robustness against affine distortion, viewpoint changes, illumination variations, and noise. The SIFT descriptors are used here as a mapping of the input pixels into a robust representation equipped with the SIFT properties, and thus, the local rotation invariance is integrated implicitly to the framework This is achieved by conducting SIFT-CNN training using SIFT-Images instead of directly operating on image pixels. Also, the SIFT-CNN takes advantage of both domains, the hand-crafted SIFT descriptors and the learned features from CNNs. Through evaluation on different problems, we demonstrate that this novel combination yields improved efficiency. Finally, the SIFT-CNN emphasizes the representation of the input images rather than on specific CNN architectures or loss functions and offers an alternative approach to enhancing performance and addressing the limited sample size if it is presents. Initially, an investigation of SIFT-CNN framework is developed on the widely-used MNIST ("Modified National Institute of Standards and Technology") dataset, consisting of handwritten digit images [116], [117]. The MNIST database serves as an ideal testbed for benchmarking classification algorithms. Here, the experiments with varying number of training examples showed that the SIFT-CNN can be efficiently trained with a limited number of samples. Next, our study explores mostly the local rotation invariant property by examining scenarios where local areas within an image undergo rotation without affecting the overall classification category, eliminating the need to rotate the entire image. Examples of such problems include indirect immunofluorescence (IIF) cell images, ground-based all-sky cloud images, and human Lip-Reading image sequences, where cell, clouds, or part of the mouth area can be rotated within the image, but the final image class decision should be preserved, as one can observe from some example data in Figure 3-1. In the case of the biomedical problem involving Human Epithelium Type-2 (HEp-2) cell images [118], the proposed SIFT-CNN framework outperforms the network trained directly on image pixels. Also, the experiments conducted on both the cloud type classification using a large dataset of all-sky images (GRSCD) [119], [120] and the sequence modelling task of word-level Lip-reading recognition (LRW) [121] showed that SIFT-CNN achieved state-of-the-art performance. The experimental results across these various tasks consistently indicate that the proposed SIFT-CNN approach provides significant improvements compared to CNNs trained directly on pixel images. Hence, the SIFT-CNN demonstrates its effectiveness in enhancing performance across a range of computer vision problems, making it a promising and efficient approach.

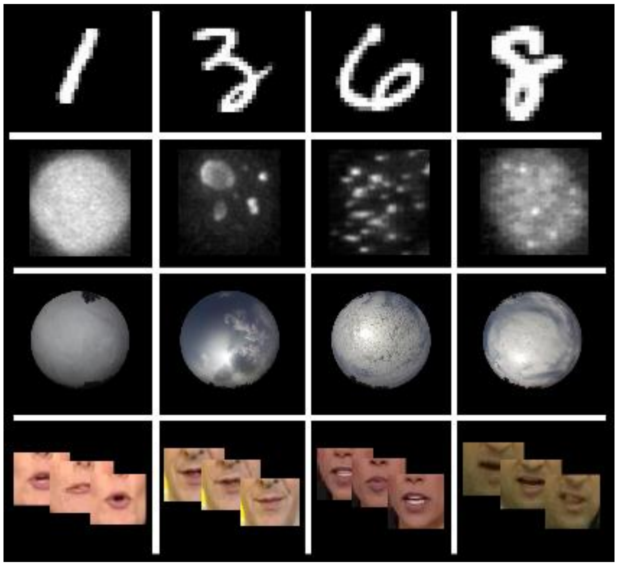

Figure 3-1: Some representative data examples of the related problems. The first row shows handwriting digits from the most popular machine learning dataset of MNIST. The second row presents IIF cell images for the HEp-2 cell classification task. The third row includes the all-sky images of the whole sky dome, where different types of clouds are recognized. The fourth row corresponds to the lip-reading classification task, where each sample is an image sequence (29 frames) representing a spoken word. Obviously, local rotation invariance is a sought-after property in these tasks.

The rest of this chapter is organized as follows: A brief overview of the existing combination of hand-crafted SIFT features with the deep learning topologies are given in Section 2. The proposed method is detailed in Section 3. The experimental procedure on four different classification tasks, incorporating handwritten digit images (MNIST), Human Epithelium Type-2 (HEp-2) cells microscope images, all-sky cloud (GRSCD) images, and Lip-Reading video (LRW) along with the corresponding results are given in Section 4. Finally, the conclusions are drawn in Section 5.

3.2 Related Work

For many years, computer vision research focused on developing various calculation methods for hand-crafted features, particularly local descriptors such as SIFT, along with feature encoding mechanisms to provide a robust image representation. However, over the past decade, the dominance of CNNs has reshaped the field. Recently, there has been a growing interest in combining SIFT descriptors with CNNs, recognizing the potential synergy between the two approaches [122]. In most of the proposed works, the SIFT features are merged with the CNN features at the final stage just before the classification topology [123], [124]. Thus, two streams are utilized independently, on the one hand is implemented the calculation of the SIFT descriptors along with k-means algorithm for the bag-of-words encoding and on the other hand the CNN features are extracted utilizing a deep learning model. The outputs of the streams are fused, and the result is fed a classifier consisting of fully connected layers. Next, only the CNN-stream is updated through backpropagation on the respective stream. In this manner, many different approaches are proposed for the calculation of local descriptors either exploiting key-point SIFT [125], [126] or jointly with dense SIFT features [127]. Besides, the fusion method is varied from simple concatenation to more sophisticated attention mechanisms [123], [128], [129]. The previous dual-stream logic is modified by redoubling each stream and implement a Siamese scheme [130]. Additionally, hybrid CNN and SIFT methods are evaluated on sequence modelling tasks to capture video dynamics in opposition to optical flow [131], [132].

In scenarios with limited data availability, the use of shallow representations achieved through hand-crafted descriptors has demonstrated significant advantages. These approaches prove particularly valuable in addressing the challenges posed by data scarcity, enabling effective analysis and solutions [133], [134]. In an effort to reduce the number of learnable parameters in a CNN model, several works have proposed replacing the learnable parameters in the initial layers with user-specified functions and pre-defined handcrafted filters. One such approach involves incorporating Gabor filters into CNNs to enhance the resistance of deep learned features to the orientation and scale changes [135]. Another method combines the ScatNet, utilizing a pre-defined Morlet filter bank to extract features, with a CNN architecture to create a deep hybrid network [136]. The PCANet [137] employs Principal Component Analysis (PCA) for the filter banks, and its modifications include LDANet, which trains cascade filters using Linear Discriminant Analysis (LDA), and MLDANet, which combines PCANet and LDANet [138]. Additional variations of the PCANet approach are the PCA-based Convolutional Network (PCN) [139], a multi-stage convolutional network that can be trained layer-wise in an unsupervised manner, and KPCANet [140], which incorporates kernel PCA and is invariant to illumination while remaining stable under slight non-rigid deformation. Another notable technique involves replacing regular CNN filters with Active Rotating Filters (ARFs), leading to a significant reduction in network parameters and improved classification performance. ARFs ensure within-class rotation invariance by actively rotating during convolution, resulting in feature maps that explicitly encode location and orientation information [141]. However, notable mechanism widely used to address translation, scale, and rotation invariance in CNNs is the Spatial Transformer Networks (STNs) [142]. STNs are a generalization of differentiable attention to any spatial transformation. Unlike the previously mentioned approaches, STNs do not aim to reduce the number of network parameters. Instead, they learn the parameters of an affine transformation that is applied to the entire input image at the early stages of the CNN, thereby enhancing the geometric invariance of the model. Finally, an approach that incorporates underlying physics in the input representation is known as physics-informed neural networks (PINNs) [143], [144]. These networks integrate (noisy) data and mathematical models and utilize additional information obtained by enforcing physical laws during training in order to be trained from additional information obtained by enforcing the physical laws. By combining data-driven learning with the constraints imposed by the underlying physics, PINNs enable more accurate and robust predictions, offering an effective framework for capturing complex phenomena while incorporating prior knowledge of the physical laws. In the end, providing some kind of invariance in the first layers of CNNs seems to be very important for learning more robust representations without requiring large amounts of data or extreme data augmentation [145]. The integration of hand-crafted feature representations with CNNs provides a compelling approach that leverages the strengths of both feature engineering techniques and deep learning architectures [146], [147]. An example of this is the utilization of Local Binary Pattern (LBP) descriptors to analyze facial texture, where a CNN is trained using LBP-encoded images to enhance its global texture perception [14], [148]. The combination of hand-crafted feature representations and CNNs offers a promising approach to extract meaningful information from data and improve the overall performance of machine learning models.

In this work, we introduce a novel method for directly integrating dense SIFT descriptors into CNNs as inputs. While the concept of SIFT-Images and the fusion of SIFT and CNN features have been previously proposed, to the best of our knowledge, the benefits of using SIFT-Images as inputs to CNNs have not been thoroughly investigated. Our approach utilizes dense SIFT, and the SIFT-Image transformation maps a single-channel image to an M-channel image, where M corresponds to the dimension of the SIFT descriptor and consequently the number of the SIFT-Image channels while preserving the spatial resolution of the original image. What sets our approach apart is the utilization of these SIFT-Images as multi-channel inputs for training the CNN model in various classification problems. Our proposed SIFT-CNN framework leverages the feature extraction capabilities of the CNN model while implicitly incorporating the local rotation invariance of the SIFT descriptor within a unified system. By utilizing the consecutive integration of these two components, our framework synergistically combines the strengths of both approaches, resulting in improved performance in various computer vision tasks.

3.3 Proposed Method

3.3.1 The SIFT-CNN framework

In a typical CNN-based system, the input consists of the pixel values of an image, and the output is the classification result for that image. We will refer to this approach as Pixel-CNN throughout the rest of this chapter. In the case of dense SIFT, where SIFT descriptors are calculated on every pixel of the image, we obtain the SIFT-Image representation. The SIFT-Image is then fed into a CNN, resulting in the complete framework known as SIFT-CNN. The overview of the two frameworks is presented in Figure 3-2.

By using the SIFT-CNN approach, the network learns spatial relations directly from the histograms of gradients in neighboring pixels. This differs from learning directly from intensity pixels and allows the network to focus on the relationships between statistical properties of the pixel regions. Specifically, CNN learns the connections between histogram bins that encode the frequency of gradient directions in the vicinity of each pixel. Importantly, the spatial resolution of the input image remains unaffected, enabling CNN to learn features with high spatial detail while utilizing the entire spatial image domain. At the same time, the spatial resolution of the input image remains unaffected, enabling the CNN to learn features with high spatial detail utilizing the total spatial image domain [43]. In essence, the SIFT-CNN leverages the properties of SIFT, leading to the implicit integration of local rotation invariance within the framework.

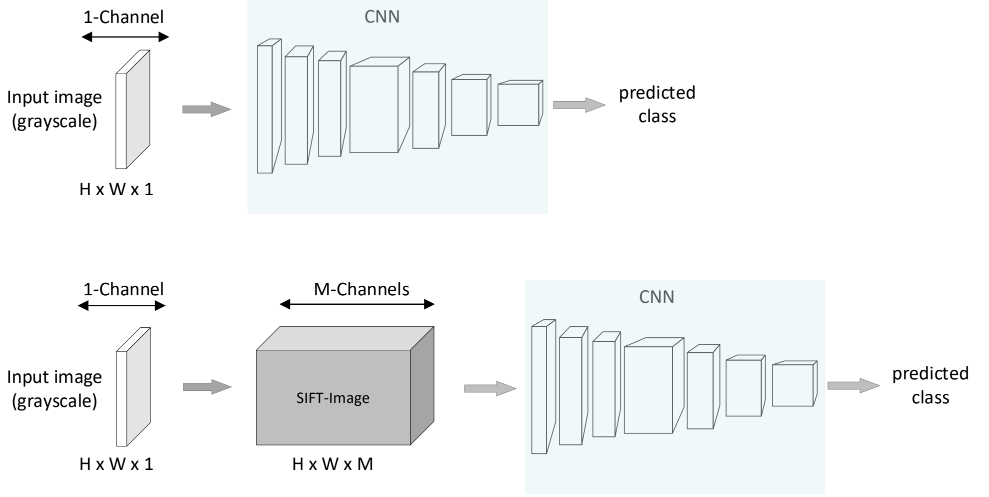

Figure 3-2: Overview of the Pixel-CNN and SIFT-CNN frameworks for image classification. Top scheme: Pixel-CNN, the regular implementation of CNN where the pixels' values of grayscale image are used directly as inputs into CNN. Bottom scheme: SIFT-CNN, the SIFT-Image representation is used as input into a CNN and thus, the SIFT-CNN is guided to learn features from local gradient information of images, something that results SIFT-CNN to incorporate implicitly the local rotation invariance property.

3.3.2 Mapping pixels to SIFT descriptors

The SIFT descriptor is computed for every pixel of a grayscale image using a method called Dense SIFT feature descriptor [149], which is roughly equivalent to running SIFT on a dense grid of locations at a fixed scale and orientation. In this work, we adopt the dominant scale approach, as suggested by previous studies [110], [150], which found that a single scale is sufficient to capture the necessary information. The dominant scale is computed by executing the SIFT detector using the training images and then, estimating the distribution's mean of all the scales. For every pixel of an image, a neighborhood of size N × N pixels is defined around it, where N is specified by the scale parameter and is set to N = 8. This local area is divided into 4 × 4 regions called cells. For each cell, an 8-bin histogram is computed and therefore, each pixel is represented as an M-dimensional feature vector, where M = 128 represents the number of bins in the SIFT histogram when all cells stacked together. As a result, each grayscale input image is transformed into a new image with M-channels, formed by the M-dimensions descriptors for every pixel. This process is presented on the following Figure 3-3. The descriptors encode statistical information related with the orientation of the gradients in the local neighborhood of pixels' area. This representation exhibits local rotation and scale invariance while also expands the receptive field-of-view of the first layer of the CNN. The increased input receptive field enables the CNN to capture higher-level features with its first layer leveraging the previous SIFT encoding. Furthermore, the properties of the SIFT descriptors are implicitly incorporated in the training process of the CNN. Training a Deep CNN with these M-channel SIFT-Images has the potential to improve generalization with fewer augmentations or training data, while implicitly transfusing a sense of local rotation invariance into the CNN.

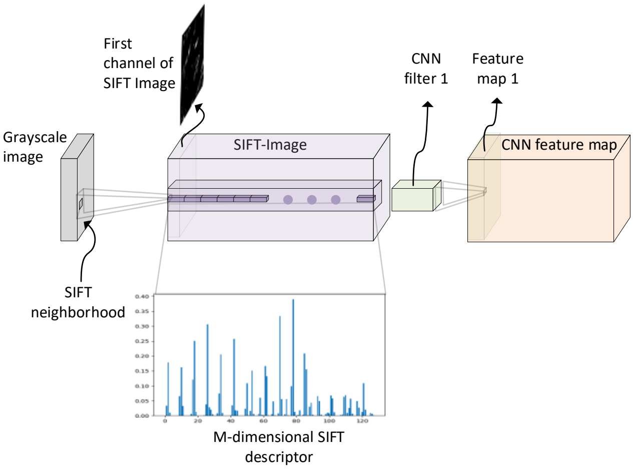

Figure 3-3: Given a grayscale image, one SIFT descriptor is computed for each pixel of the image capturing a neighborhood around every pixel. Thus, each pixel is mapped to an M=128-dimensional SIFT descriptor. For all the pixels of the grayscale image, the corresponding result is a new image that is called SIFT-Image. In the SIFT-CNN framework, every input convolutional layer of the CNN (e.g., CNN filter 1) operates directly on the SIFT-Image such as in a multiscale input image with the regular convolution process. In this way, the output of the first convolutional layer is an ordinary CNN feature map.

3.4 Experimental Results

The efficiency of the proposed SIFT-CNN is evaluated on many different challenging tasks. In all cases, the ability of SIFT-CNN to perform better than (or to be combined) regular Pixel-CNNs is presented. The ResNet-18 architecture is used as the standard CNN in the SIFT-CNN framework, since ResNets have demonstrated remarkable performance in transfer learning [151]. Optimization is performed by minimizing the loss with Stochastic Gradient Descent (SGD) for 100 epochs, starting with an initial learning rate of 0.1, which is divided by 10 every 30 epochs, unless specified otherwise. The size of the minibatch is determined by the maximum memory on GPU, with 64 images for image classification problems and 8 for the sequence classification task. However, our preliminary investigation using smaller minibatches (i.e., 8,16,32) results in performance degradation of less than 1% for each reduction. Unless otherwise stated, no specific data augmentation procedures are incorporated into the training procedures.

3.4.1 Datasets

The MNIST dataset consists of 60,000 training images and 10,000 test images, representing 10 handwritten digit categories ranging from zero to nine [116], [117]. Upon conducting experimental investigations using the 28 × 28 pixels images from MNIST, we observed the need for a larger input spatial area to efficiently train CNNs with SIFT-Image representations. This is because CNNs require the ability to handle the additional information provided by the SIFT descriptors. Therefore, for all experiments involving Pixel-CNN and generating SIFT-Images, the grayscale MNIST images with a resolution of 28 × 28 pixels are resized to 64 × 64 pixels using bicubic interpolation.

Two publicly available biomedical datasets with single-channel (grayscale) images of Human Epithelium Type-2 cells (HEp-2 cells) are used for the task of cell image classification [118]. These datasets were introduced in two contests and are known to be highly challenging. The first one is ICPR 2012 HEp-2 cell dataset, consisting of 721 training images and 734 test images across six categories [152]. The training and test sets are already provided by the contest organizers. The second dataset is ICIP 2013 HEp-2 cell contest dataset with 13,652 cell Images and six cell classes [153]. From the total of 13,652 images, 1,186 were used for training and the rest 12,466 were allocated for testing. To ensure consistency in the experiments, all grayscale cell images are resized to a resolution of 128 × 128 pixels. This resolution is applied to both the input images for Pixel-CNN and the generation of SIFT-Images.

The TJNU ground-based remote sensing cloud database (TJNU-GRSCD) contains 8000 cloud images captured by the sky camera with fisheye lens [119], [120]. The images are collected for a long period of time from 2017 to 2018 in Tianjin, China. Every ground-based sample is an RGB image of the sky dome with the resolution of 1024 × 1024 pixels and preserved in the JPEG format. The sky conditions are divided into 7 sky types: 1) cumulus, 2) altocumulus and cirrocumulus, 3) cirrus and cirrostratus, 4) clear sky, 5) stratocumulus, stratus, and altostratus, 6) cumulonimbus and nimbostratus, 7) mixed cloudiness, according to cloud genera definitions of the World Meteorological Organization (WMO) and the visual similarity of cloud in practice. The GRSCD is composed of 4,000 training samples and 4,000 test samples from 7 classes, as provided by the creators. The RGB images are converted to grayscale and resized to 280 x 280 pixels in order to allow the image augmentations of random crop into resolution of 256 × 256 and random horizontal flip during training.

The Lip-Reading problem is tackled using the LRW-500 dataset, which is a challenging large-scale dataset specifically designed for lip reading tasks [121]. This LRW (Lip Reading Words) dataset contains words extracted from short video clips captured automatically from BBC TV broadcasts. Each spoken word is represented with 29 RGB frames. The dataset contains a total of 500 different word classes, with 488,766 training samples and 25,000 samples for both validation and testing. To maintain a consistent frame length, the creators of the dataset cropped fixed windows with the target word positioned at the center. For our evaluation, each image was cropped to an 88 × 88 pixels region around the mouth area. These cropped images were transformed to grayscale and then into SIFT-Images. This mapping process converts each grayscale image sequence into a corresponding SIFT-Image sequence, which is subsequently processed by the CNN architecture.

3.4.2 Classification Results on MNIST dataset

The evaluation of the SIFT-CNN framework on the MNIST dataset enables us to conduct two distinct investigating directions: a) analyzing the impact training samples number, and b) assessing the robustness of SIFT-CNNs in terms of rotation invariance. By conducting experiments on the well-known MNIST dataset, we can examine the framework's performance under varying numbers of training samples and determine its ability to handle image rotations effectively.

3.4.2.i Study the behaviour of SIFT-CNN with limited samples

The objective is to investigate the behaviour of SIFT-CNNs concerning the number of available training samples and thus experiments are conducted by varying the number of training images. The number of training images is reduced to 30,000 , 10,000 , 5,000 , and 1,000 in addition to the full training dataset of 60,000 images. For each training set, the SIFT-CNN was compared with Pixel-CNN using the ResNet-18 model, and the results are presented in Table 3-1. In cases with less than 60,000 images, the average accuracy after 100 repetitions with randomly sampled training images is demonstrated. The results indicate that Pixel-CNN can achieve similar performance to SIFT-CNN only when it is trained with a large number of training samples. On the contrary, as the number of training samples decreases, the SIFT-CNN consistently achieves better accuracy than Pixel-CNN. From the results in Table 3-1, we can observe that the accuracy of SIFT-

CNN remains more stable, consistently above 99%, even when using less than 10% of the original available samples, namely 5,000 images.

Table 3-1.Classification Accuracy on MNIST hand-digit dataset using different training set sizes.

Number of Training Samples	**Classification Accuracy (%) on MNIST**	
	SIFT-CNN	**Pixel-CNN**
60000	99.47	99.49
30000	99.45	99.43
10000	99.17	98.78
5000	99.09	98.42
1000	97.46	95.12

3.4.2.ii Rotation invariance and comparison with Spatial Transformer Networks

In this section, our objective is to study the behaviour of SIFT-CNN framework concerning rotations. To do this, only the test set is subjected to rotations of various degrees. One of the most effective methods for achieving global rotation invariance is the of Spatial Transformer Networks (STN) [142]. Therefore, in this experiment, we study the rotation invariance properties of SIFT-CNN its compare them with Spatial Transformer Networks (STN) and Pixel-CNN. To assess the ability of both Pixel-CNN and STN-CNN to overcome the challenge of rotation invariance, data augmentations are included in their trainings too. These augmentations applied random image rotations between 0^0 and 90^0 degrees. In this experiment, ResNet-18 models are trained from scratch to avoid using weights learned with extreme augmentations from other pretrained datasets. The training set consists of 1000 training samples by randomly sampled 100 training images per class digit. The experiments are performed by running 100 random repartitions to stress the generalization capabilities of the frameworks. During these experiments, the 10,000 test images are randomly rotated at angles between 0^0 and 45^0 degrees, while the training set remains unmodified in means of rotation. The average classification accuracy values (average over 100 random repetitions) for the different frameworks, considering various rotation angles, are presented in Table 3-2. The Results demonstrate that SIFT-CNN could equip the framework with robustness against arbitrary rotations of input test images. Even for extreme rotations, the SIFT-CNN exhibits a descent performance. Interestingly, the rotation augmentations appear to have a negative impact on the models' performance, particularly in cases where the training data already have intrinsic small perturbations, such as the MNIST dataset. It is also worth-noting that the STN-CNN outperforms the Pixel-CNN in scenarios involving large rotation angles. However, Pixel-CNN demonstrates superiority in cases of small rotations where the STN may not function optimally.

Table 3-2. Evaluating different frameworks to cope with arbitrary rotations of the MNIST test set.

Test set Rotation Angles (in degrees)	**Classification Accuracy (%) on MNIST**				
	SIFT-CNN without rotation augmentations	**Pixel-CNN without rotation augmentations**	**Pixel-CNN with rotation augmentations**	**STN-CNN without rotation augmentations**	**STN-CNN with rotation augmentations**
0^0	97.46	95.12	89.66	92.43	81.16
5^0	97.20	93.66	89.75	92.30	78.60
15^0	93.75	82.06	81.22	91.46	77.67
30^0	84.15	51.72	64.39	78.92	72.29
45^0	73.75	27.60	52.45	71.26	67.90

3.4.3 Classification results on ICPR 2012 and ICIP 2013 HEp-2 cell image datasets

The Hep-2 cell image classification problem motivated us to design the SIFT-CNN method, and thus some critical parameters of the framework are optimized on the ICPR2012 dataset. Firstly, dimensionality reduction methods were applied to the 128-channel image representation to reduce the input 3D volume's depth. Half of the training images were used to learn the undercomplete dictionary for NMF (Non-negative Matrix Factorization) or the projection matrix for PCA (Principal Component Analysis), resulting in a reduced dimensionality of $M < 128$. The kernel size of filters in the first layer of the CNN was also investigated, considering that the SIFT-Image pixels already encompass a larger receptive field with an encoding neighborhood of size $N \times N$ pixels ($N = 8$). Two sizes were tested: 1×1 and 3×3 pixels. Results on Table 3-3 indicate that the 3×3 convolution kernel demonstrated notably higher accuracy, supporting our hypothesis that transforming the data into a more informative representation leads to a more efficient starting point for training the CNN model. By enhancing the information encoding in initial layers, larger convolution kernels prove beneficial for capturing more extensive neighbor information. NMF encoding showed performance improvement with an increase in the number of dimensions, while PCA-based representation was inferior, especially for larger dimensions. The evaluation showed that dimensionality reduction did not significantly contribute, leading us to focus on transfer learning in the following experiments.

Table 3-3: Parameter analysis in input representation depth and kernel size of first convolutional layer using the ICPR2012 dataset.

Parameters	**Classification Accuracy (%) on ICPR2012**			
Kernel size of filters in the first layer of CNN	**1 × 1**		**3 × 3**	
SIFT-Image input (M=128 Channels)	70.65		73.09	
Dimensionality Reduction (DR) method	**NMF**	**PCA**	**NMF**	**PCA**
Channels of Input Image after DR (M)				
4	65.76	65.48	65.76	65.75
8	66.44	66.30	65.48	66.04
16	69.70	70.52	69.29	68.75
32	69.15	68.20	70.78	70.24
64	69.43	68.47	70.92	70.92
76	66.44	66.03	70.24	68.47
96	67.25	66.30	72.14	68.85

The availability of two HEp-2 cell image datasets sharing the same classes, namely the ICPR 2012 and ICIP 2013, provides the opportunity for a comprehensive experimental analysis. On the one hand, the CNN is studied separately in each dataset and on the other hand, the transferability of the learned features across datasets is investigated. In the first case, the ResNet-18 is trained using only the training set of each dataset. Subsequently, the trained model is tested on the respective test set of each dataset. This approach is referred to as "without transfer learning" in the results presented on the Table 3-4 below. In the second case, the ResNet-18 is initially trained using the training images of one dataset. Then, the trained model serves as the initialization point for further training on the other dataset. The accuracy is reported on the test set of the final dataset. This approach is referred to as "with transfer learning" and the results are also presented in Table 3-4. To ensure a fair comparison, Pixel-CNN is tested using the same protocols as SIFT-CNN. The SIFT descriptors are computed using the patch dominant orientation option, as depicted in Figure 3-4.

Table 3-4. Classification Results on the Hep-2 cell image biomedical datasets.

Methods	**Classification Accuracy (%)**	
	ICPR 2012	**ICIP 2013**
Proposed: Pixel-CNN(ResNet-18) without transfer learning	66.3	84.47
Proposed: Pixel-CNN(ResNet-18) with transfer learning	68.5	86.12
Proposed: SIFT-CNN(ResNet-18) without transfer learning	73.0	89.18
SIFT + VHAR [110]	73.4	**-**
SIFT-SURF + BoW [154]	75.0	**-**
SPM [154]	75.0	-
Proposed: SIFT-CNN(ResNet-18) with transfer learning	75.0	89.21
SBoW [155]	78.0	-

The SIFT-CNN demonstrates a notable improvement compared to the regular Pixel-CNN both when transfer learning is performed or not. These results highlight the superior performance of SIFT-CNN, indicating that the combination of SIFT-Image with a CNN model allows the CNN to effectively leverage the dense SIFT properties and handle the complex textures present in the cell images, surpassing the utilization of pixel values alone. Considering that images captured from fluorescence microscopy often contain noise, it is evident that SIFT descriptors provide more robust representations compared to the noisy pixels. Furthermore, the performance of SIFT-CNN is statistically comparable to traditional methods that employ SIFT descriptors along with Vector of Locally Aggregated Descriptors (VHAR) or frequency-related Bag-of-Words (BoW) encoding. The effectiveness of hand-crafted features, as opposed to Pixel-CNN's learned features, is primarily attributed to the presence of noise in microscope images rather than the limited number of training samples. However, it is worth mentioning that both Pixel-CNN and SIFT-CNN exhibit the ability to transfer knowledge between tasks in all cases.

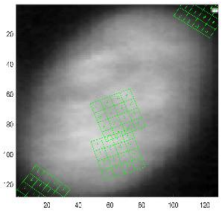

Figure 3-4: SIFT descriptors computed using patch dominant orientation on a cell image.

3.4.4 Classification results on GRSCD with all-sky images

The SIFT-CNN was evaluated on the Ground Based Remote Sensing Dataset (GRSCD), utilizing only the visual information since another version of the dataset includes additional multimodal information for every image. Table 3-5 presents a variety of methods including both traditional techniques and deep learning architectures. The traditional-based features are calculated using the SIFT descriptors together with bag-of-words (BoW), with the uniform invariant local binary patterns (LBP with the (P, R) set to (24, 3), respectively), and the completed LBP that is a joint combination of local central information, signs, and magnitudes of the local differences (CLBP with P = 24 and R = 3). Many popular CNN topologies are also presented in Table 3-5, such as the VGG-16, the AlexNet-like for CloudNet, deep convolutional activation-based features (DCAFs), and ResNet variations. The deep learning methods exhibit a significant advantage over hand-crafted methods in ground-based cloud classification. This is attributed to the unique characteristics of cloud images, which possess large intraclass and small interclass variances in terms of texture (e.g., similar clouds at different heights) and color (e.g., different daytime). CNNs prove effective in learning distinctive representations from challenging fisheye sky images, while could incorporate additional mechanisms in their end-to-end training. The utilization of these mechanisms, such as the ResNet model trained with dual guided loss (DGL), hierarchical fusion of intermediate feature maps of deep visual features, and the attention mechanism for exploiting local visual features (Attentive Network) is crucial for capturing inherent structural information and improving performance. Hence, fusion techniques enhance performance by extracting information from multiple CNN-based streams and combining the outputs, as seen in the Improved Combined Convolutional Network (ICN). Considering this, the proposed SIFT-CNN is combined with PixelRGB-CNN using a late fusion mechanism to further improve results. Finally, to optimize the decision boundary, a support vectors machine (SVM) classifier at the top of the final extracted features seems advantageous for the cloud type classification task, as evidenced by increased classification results.

The experiments using both Pixel-CNN and SIFT-CNN performed with the stochastic gradient descent (SGD) optimizer started with a learning rate 0.001 and a weight decay and momentum set to 0.0002 and 0.9 respectively. The learning rate was decreased every 30 epochs using a step function by a factor of 0.1 for a total of 100 epochs when the minibatch had 64 images. Results on Table 3-5 indicate that SIFT-CNN provides an efficient way to encode and utilize the SIFT descriptors. By comparing it with the standard approach for encoding SIFT descriptors into a histogram of occurrences (BoW), SIFT-CNN provides an improvement of about 16%. Moreover, SIFT-CNN surpasses pixel-CNN regardless of the used architecture, like the ResNet-18, ResNet-50, AlexNet-like, and VGG-16. However, on its own, SIFT-CNN falls short in achieving a high classification score. To improve the results, additional mechanisms such as a fusion scheme should be employed, as evident from the relevant works listed in the table.

Table 3-5. Summary of the state-of-the-art results on the GRSCD.

Methods	**Classification Accuracy (%) on GRSCD**
SIFT + BoW [119], [156]	66.13
LBP (P = 24, R = 3) [119], [157]	50.20
CLBP (P = 24, R = 3) [119], [158]	69.18
VGG-16 [119], [159]	77.95
CloutNet (ALexNet-like) [119], [160]	79.92
Proposed: PixelRGB-CNN (ResNet-18) [161]	82.52
DCAFs & SVM [119], [156]	82.67
ResNet-50 [46]	83.15
Proposed: SIFT-CNN (ResNet-18)	83.90
ResNet-50 + DGL [162]	85.28
ResNet-50 + hierarchical fusion & SVM [163]	85.12
ResNet-50 + Attentive Net & SVM [119]	86.25
Proposed: Late Fusion PixelRGB-CNN and SIFT-CNN (Resnet 18) [161]	87.22
Proposed: Late Fusion PixelRGB-CNN and SIFT-CNN (Resnet 18) & SVM [161]	87.55
TGCNN (with ResNet-50) [120]	89.48
Fusion of ICN, ResNet-50 and VGG-16 [164]	90.08

3.4.5 Classification results on Lip-Reading LRW dataset

Previous experiments have demonstrated the effectiveness of SIFT-CNN for single image classification. In this section, the capability of SIFT-CNN in a problem of sequence modelling is investigated. For this purpose, the Lip-Reading (LR) problem is approached using a very challenging and large-scale dataset consisted of 500 English spoken words. LR is a challenging image sequence classification task where the CNNs are asked to learn very high-level, abstract patterns of mouth motion from sequences of frames [165]. Traditionally, sequence encoding tasks have relied on recurrent neural networks (RNNs) such as GRU and LSTMs. However, in recent years, Temporal Convolutional Networks (TCNs) have garnered significant attention for various sequence learning tasks, including action recognitions [166], weather predictions [167], and also LR task [168], [169]. TCNs offer an alternative approach to sequence modeling, utilizing convolutional layers to capture temporal dependencies and extract meaningful features from sequential data. Towards this direction, a state-of-the-art implementation has been obtained by combining spatiotemporal convolutions, also known as 3D convolutions, with ResNet-18 CNNs and Multiscale Temporal Convolutional Networks (MS-TCN) [168]. In this approach, the frames of the sequence are passed through a 3D convolutional network and then processed frame-wise from a ResNet-18 to extract a feature vector from each frame. Finally, the TCNs are used to map the sequence of vectors into a fixed length vectorial representation, thereby providing sequence encoding. Our goal is to explore the power of the input image representation utilizing the SIFT-Image in combination with a deep architecture. Thus, we train the MS-TCN based Lip-Reading system proposed by [168], using the SIFT-Images as input, ensuring fair comparison with plain rule as possible. More specifically, given a gray scale image with a resolution of 88 × 88 pixels as input, we compute the SIFT-Image, resulting in a tensor of size 88 × 88 × 128 (Height × Width × Channels). Next, we apply two convolutional layers with a kernel size of 3 and a stride of 2 in order to map the channels from 128 to 64, and then from 64 to 64, respectively. Subsequently, a third convolutional layer with a stride of 1 and a kernel size of 3 is used to transform the 64 channels to 64. Padding of 1 is applied to all images. This downscales the dimension of the SIFT-Image by a factor of four resulting in an input tensor of size 22 × 22 × 64, as proposed by the work of [168]. Moreover, the essential 3D learning module in the initial stages of the LR system is used, similar to [168], but with the corresponding SIFT-Image sequence as input.

Table 3-6 presents the classification accuracy of the state-of-the-art methods on the word-level LRW dataset. The experimental results indicate an advantage of SIFT-CNN – MS-TCN over Pixel-CNN – MS-TCN [168]. To provide a comprehensive comparison, we also trained the Pixel-MS TCN of [168] from scratch, however we achieved only 79.38%, something that indicates Pixel – MS-TCN needs some particular treatment as mentioned by its authors too, like the pre-training in a few words and then gradually increasing the number of words as well as a transfer learning process by training on a different task first. The increased classification accuracy of SIFT-CNN can be attributed to the robustness in brightness, constancy, and piecewise smoothness of SIFT-flow. Additionally, the local rotation invariance properties and the higher-level information captured by SIFT-descriptors (through local gradient encoding) contribute to the superior performance of the proposed system compared to using regular pixel image as input.

Table 3-6. Summary of the state-of-the-art results on the LRW-500 dataset.

Method			**Data**		**LRW-500**
Authors (Year)	**Frontend**	**Backend**	**Input image size**	**Input and data managing policy**	**Classification Accuracy WRR (%)**
Chung et al. (2016) [121]	3D &VGG M	-	112 × 112	Mouth	61.10
Chung et al. (2017) [170]	3D & VGG M version	LSTM & Attention	120 × 120	Mouth	76.20
Petridis et al. (2018) [171]	3D & ResNet-34	Bi-GRU	96 × 96	Mouth	82.00
Stafylakis et al. (2017) [172]	3D & ResNet-34	Bi-LSTM	112 × 112	Mouth	83.00
Cheng et al. (2020) [173]	3D & ResNet-18	Bi-GRU	88 × 88	Mouth & 3D augmentations	83.20
Wang et al. (2019) [174]	2-Stream ResNet-34 & DenseNet3D-52	Bi-LSTM	88 × 88	Mouth	83.34
Courtney et al. (2019) [175]	alternating ResidualNet Bi-LSTM	alternating ResidualNet Bi-LSTM	48 × 48, 56 × 56, 64 × 64	Mouth (& pretraining)	83.40 (85.20)
Luo et al. (2020) [176]	3D & 2-Stream ResNet-18	Bi-GRU	88 × 88	Mouth and gradient policy	83.50
Weng et al. (2019) [177]	deep 3D & 2-Stream ResNet-18	Bi-LSTM	112 × 112	Mouth & optical flow	84.07
Xiao et al. (2020) [178]	3D & 2-Stream ResNet-18	Bi-GRU	88 × 88	Mouth & deformation flow	84.13
Zhao et al. (2020) [179]	3D & ResNet-18	Bi-GRU	88 × 88	Mouth and mutual information	84.41

Method			**Data**		**LRW-500**
Authors (Year)	**Frontend**	**Backend**	**Input image size**	**Input and data managing policy**	**Classification Accuracy WRR (%)**
Zhang et al. (2020) [180]	3D & ResNet-18	Bi-GRU	112 × 112	Mouth (Aligned)	85.02
Feng et al. (2020) [181]	3D & SE ResNet-18	Bi-GRU	88 × 88	Mouth (Aligned) & augmentations	85.00
Pan et al. (2022) [182]	3D & MoCo	Transformer	112 × 112	Mouth (& pretraining)	85.00
Martinez et al. (2020) [168]	3D & ResNet-18	MS-TCN	88 × 88	Mouth (Aligned)	85.30
Kim et al. (2022) [183]	3D & ResNet-18	Bi-GRU	112 × 112	Mouth (& pretraining)	85.40
Tsourounis et al. (2021) [184]	alternating ALSOS & ResNet-18 layers	MS-TCN	88 × 88	Mouth (Aligned)	85.96
Proposed	SIFT- 3D & CNN(ResNet-18)	MS-TCN	88 × 88	Mouth (Aligned)	86.46
Kim et al. (2022) [185]	3D & ResNet-18	MS-TCN + KD (ensemble)	88 × 88	Mouth (Aligned)	88.50
Koumparoulis and Potamianos (2022) [186]	3D & EfficientNetV2 & + Transformer	TCN	88 × 88	Mouth (Aligned)	88.53

3.5 Conclusions

The combination of hand-crafted descriptors with the deep learning methods is an open research domain that bridges the experience of computer vision community with hand-crafted features with the feature representation capabilities of deep learning. Our attempt to combine these two worlds resulted in the SIFT-CNN framework, which consists of a mapping that produces a new image representation based on SIFT descriptors and a learning process based on efficient CNN architecture. For every pixel of an input single-channel (grayscale) image, the SIFT descriptor is calculated generating the SIFT-Image with channel size equals to 128 (as SIFT descriptor dimension) and spatial size as the input grayscale image. Next, the SIFT-Images are fed into a CNN model under a final classification task. The SIFT-CNN framework has its own set of advantages and limitations. One limitation of the SIFT-CNN is its inability to inherently encode color information. In cases where color plays a crucial role in discriminating different classes, SIFT-CNN needs to compute SIFT descriptors per color channel and employ fusion mechanisms on the outputs, increasing computational complexity. The computation of dense SIFT introduces extra initial procedures, leading to higher processing resources and time compared to frameworks working with pixel images as input. However, the time cost during training and testing is not significantly impacted due to the implementation of SIFT computations on GPUs, and only the descriptor calculation stage is executed. At last, the utilization of a larger input volume (H×W×128 instead of H×W×3 or H×W×1) has negligible impact on process time but needs more memory requirements that evidently restricts the size of the minibatch. However, we observed that the proposed framework does not expect large minibatches to be efficient. On the other hand, SIFT-CNN offers several advantages. First, for every pixel, the surrounding pixels gradient information is encoded into a histogram and thus, information is encoded channel-wise in SIFT-Image. In this context, every pixel across channels encodes the occurrence of gradient patterns. This mapping allows the CNN to be trained directly on the values formed by the SIFT histograms using an end-to-end learning scheme. In this manner, the SIFT-CNN can be advantageous in small datasets where regular Deep learning methods are prone to overfit as the try to learn all the feature representation and the encoding while SIFT-CNN enforces these networks to be trained on statistical information that might be encoded gradually in an end-to-end manner by the CNN. Secondly, the SIFT representation inherently provides strong local rotation invariance, which can be implicitly incorporated into the SIFT-CNN framework. Finally, experimental results demonstrate that the proposed SIFT-CNN operates better than the CNNs trained directly on pixels' values in all investigated tasks. This indicates that incorporating SIFT-Images as input to a CNN could be an effective and easy alternative that enhances system's efficiency. By striking a balance between the SIFT-based features and CNN-based features, the SIFT-CNN benefits from the local rotation invariance and the data-driven learning capability. In conclusion, the SIFT-CNN framework offers a promising approach by combining the strengths of hand-crafted descriptors and deep learning.

Chapter 4

Deep learning with auxiliary external data

4.1 Introduction

The scarcity of training data poses a significant challenge in pattern recognition applications [9], [187]. Data limitations though are really inherent in the signature verification task, since a practical Handwritten Signature Verification system (HSV) should be designed and efficiently trained using just a small number of reference signature from each user and also, enabling easy model updating since the signature of a writer may change -deliberately or not- through the years. A typical structure of an HSV is presented on Figure 4-1. Thus, the solution to the small sample size problem of HSV is either the “in-vitro” training using a large signature dataset and a transfer-learning approach [102] or data augmentations via generating more samples based on the existing signatures [188]. In the case of Offline Signature Verification (OffSV), significant amounts of signature images can be found in the GPDS-960 corpus database with more than half a thousand writers used for training, having 24 genuine and 30 forgeries signatures per writer [44], [189], [190]. Unfortunately, this dataset is no longer publicly available due to the General Data Protection Regulation (EU) 2016/679[1], thus hindering the efforts of the research community to develop more complex methods that require more training data. Additionally, efficient augmentation of signatures poses a challenge, primarily due to the need for characterization of the resulting images as genuine or forgeries. The utilization of augmented signature images, whether obtained through geometrical transformations [73], [191] or generative learning models [75], [192]–[194], is problematic when assuming them to be genuine because these images are generated through a third-party process. On the other hand, considering them as forgeries could introduce bias to the forgery class with the characteristics of the augmentation method. Thus,

[1] www.gpds.ulpgc.es

the utilization of signature duplications in feature learning methods necessitates special processes to ensure their effectiveness. Otherwise, they are susceptible to deviating from the realistic intra-subject variability criterion and poor image quality [193].

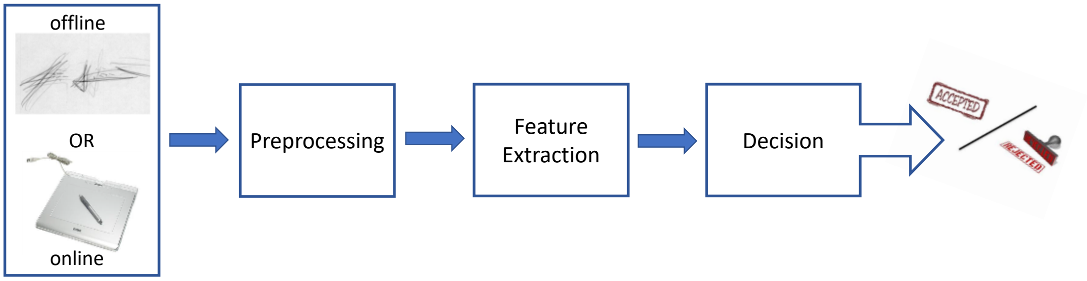

Figure 4-1: Overview of an automatic Handwritten Signature Verification system (HSV), which builds up with the Preprocessing stage for input data, the Feature Extraction stage for vectorial representation of inputs, and the Decision stage for classifying the result. A query signature (either an offline or online signature) along with the claimed identity of the user are the inputs of the HSV system and the output result is accepted if the query signature classified as genuine or rejected if the query signature regarded as forgery. Ultimately, the HSV answers the question "is the user really who he/she claims to be?".

In this work we explore an alternative path that could enable the continued incorporation of modern deep-learning techniques to OffSV systems, despite the setback caused in the OffSV field by the unavailability of the largest (to date) public dataset. In this context, we demonstrate that state-of-the-art performance can be achieved by harnessing other types of data via appropriately designed training procedures. In particular, we present an OffSV system based on a transfer learning process for training a deep Convolutional Neural Network (CNN) that is utilized as the feature extraction stage of the OffSV system. In order to enrich the feature representations learned by the CNN without the need of a vast number of signature images, we opted for transferring the larger part of the data-intensive training procedure in a domain similar to OffSV, but with an abundance of training data. To that purpose, the CNN is first trained for solving the writer identification problem using handwritten text data. The rationale behind this decision is that since both signature and text handwriting are complex high-level tasks associated with the person's motoric system and psychophysical state, it is reasonable to expect that features learned in one task can be useful to the other. We were inspired by the fact that the nature of the data is very similar for the two tasks, being comprised by scanned images of handwritten strokes. In this sense, features learned for such task should be far for informative to the OffSV compared to the usual approach for transfer learning where CNNs are pretrained to large-scale databases with natural images. Hence, in this work we attempt to operate with an auxiliary domain of handwritten text data aiming to transfer knowledge to the target domain of handwritten signature data. In more detail, the explored domains have the following characteristic:

- Auxiliary domain: A public Latin-based (western) handwriting dataset is utilized in this work, where several subjects write some predefined pieces of text in certain forms. The images of the filled forms of text are considered as the raw data of this domain. Such data though, should be processed in an appropriate way in order to generate data that are as closely related to the signature data as possible. Therefore, we designed a novel process of extracting multiple images of text from every handwritten form, taking care of preserving the personal handwriting information. CNN is trained in the writer identification problem using the generated text images, labeled with the writer's ID.
- Target domain: Three well-known western Offline signature datasets are used for evaluating the proposed OffSV system. The signature images of each dataset are utilized either in WD classifiers for estimating the performance of the system using the genuine and skilled forgery signatures or in a cross-validation way for additional training the system with the genuine signatures of one dataset and tested the system on the other datasets using the genuine and skilled forgery signatures. In all cases, a WI feature learning scheme along with WD classifiers is followed for OffSV.

After the pretraining of the CNN in an auxiliary domain, the learned features can be tailored to the OffSV task through different techniques, in an intermediate fine-tuning step. In this work we demonstrate that a metric learning stage can be used to learn an efficient mapping of the signatures' features to a latent space. A module that learns a metric or similarity measure between signatures can be trained independently of the CNN model, based on the features extracted from the model using the signature images as input. Such function can be learned using just pairs of signatures, which are considered as similar when the two signatures come from the same writer and dissimilar when the two signatures originate from different writers. We provide evidence that this process can be successfully realized using only pairs of genuine-genuine and genuine-random forgery for learning such mapping function. In the last stage of the presented OffSV system, the extracted and mapped features, are used to verify the validity of a query signature using WD kernel based SVM classifiers, each one trained individually on the reference signatures of the corresponding signer and some randomly sampled genuine signatures from other signers (used as random forgeries). As a consequence, there is no need of skilled forgery signatures for any of the training stages of the pipeline, thus eliminating the requirement for such scarce data samples that characterize many state-of-the-art OffSV systems [65], [103]. In addition, a key advantage of the proposed system is that since it exploits the handwritten text data for the learning of the CNN, it requires a significantly smaller amount of signature images for learning the final feature representation. Our system achieves state-of-the-art performance on three popular Latin Offline signature datasets, and it is competitive with systems trained on thousands of signature images using datasets which are no longer available in the public domain.

The rest of the chapter is organized as follows: Section 2 presents a brief overview of the literature related to OffSV problem, emphasizing in the deep-learning implementations. Section

3 provides an overview of the proposed approach whereas Section 4 contains a detailed description of the proposed OffSV system's pipeline. Experimental set up and results are presented in Section 5 and 6 respectively, while conclusions are drawn in Section 7.

4.2 Related Work

Feature extraction for signature images is a fundamental part of an OffSV system and various techniques have been employed for this task [195]. Although many taxonomies of such methods can be made, the most common distinction is between techniques that rely on hand-crafted features and learned features.

The hand-crafted methods aim to capture the shape of the signatures or the direction of the stokes, designing geometric, graphometry and directional features [70], [188], [196]–[204]. Also, mathematical transformations, such as Wavelets and Counterlets are utilized for feature extraction [205]–[207]. Moreover, texture descriptors and interest key-points detection techniques (e.g. SIFT, SURF, BRISK, KAZE, FREAK) are frequently used in OffSV to generate vectored representations [62], [208]–[215]. All the above methods handle with the task of producing the more compatible hand-engineered descriptors for signature images.

The learning-based approaches seem to be more efficient in OffSV task since the features are learnt directly from the images [44], [91]. The most prominent classes of algorithms from this group are the methods that rely on learning a dictionary from signature images, while the images are subsequently encoded using the learned dictionaries [91], [94], [216], [217] and methods based on deep learning [44], [77], [101], [103], [218], [219]. The first approach of harnessing deep representations for OffSV is, to the best of authors' knowledge, the utilization of Restricted Boltzmann Machine for learning an encoding/representation function [220]. Later, CNNs are used as feature extractors in the work of [221]. Generative Adversarial Networks (GAN) were utilized in [222], where the discriminator was used for extracting the signature features. Subsequently, a feature extraction CNN explicitly designed for OffSV called SigNet was proposed in [223], and latter modified effectively by [224]. In the latter approach, the SigNet is trained in the writer identification task with signature images and then, it is used as fixed feature extractor for any new test signature image. A testimony of the SigNet's efficiency are the many works in OffSV that used it, either in its original form or with various modifications [82], [101]–[103], [225]–[228]. Of course, different architectures are also investigated, such as the Capsule CNN [218], a combination of Recurrent Neural Network with Local Binary Patterns [219], LSTM models [199], and networks from the family of ResNets [77], [229], [230], however the reported results are inferior to SigNet.

A sub-class of learning-based methods are those that utilize metric-learning methods [231]. The metric learning aims to transfuse the notion of similarity between samples into the system since it is not based on the absolute positions of the embedded samples but on their relative positions to each other. The process of learning a distance between signatures is achieved either using pairs of signatures or triplets of signatures both in WI and WD systems [77], [223], [232], [233]. The triplets consist of a reference genuine signature from a writer as anchor sample with another genuine signature of the same writer as positive sample and a genuine signature of another writer or a skilled forgery signature of the same writer as negative sample. The OffSV system is trained to minimize the anchor-positive distance and maximize the anchor-negative distance and then a threshold is applied for the final verification decision [77], [232]. The pairs between two genuine signatures of the same writer and one genuine signature of one writer with one genuine signature of another writer or one skilled forgery signature of the same writer are used for training variations of Siamese-like systems and the operation of a threshold enables the OffSV decision [223], [233]. The Signature Embedding method proposed by [232], is equipped with the reduced version of VGG-16 CNN which provides a 128-dimensional feature representation for each input signature. Their scheme is designed as a WI OffSV system which is trained with signature triplets, requiring the availability of skilled forgeries. The triplet network scheme of [77] instead uses only genuine signatures for training, evaluating both the ResNet-18 and the DesnseNet-121 CNNs. Nevertheless, the performance of the generated features under the WD setting is competitive only when combined with a structural approach based on Graph Edit Distance. The WD approach of [233], named Deep Multitask Metric Learning (DMML), utilizes pairs of similar/dissimilar signatures, but the DMML was always trained on the same dataset (with the same subjects) used for testing, thus limiting the practical applicability of their technique. The Siamese architecture of [223] utilized the Contrastive loss to build a WI system but it was used an older version of SigNet with extracted features of 128 dimensions using also skilled forgeries signatures to train the CNN model. Finally, Vianna et al. [83], [83] introduced a training approach for the SigNet model, which involves two sequential tasks. Firstly, the model is trained on the writer identification task, where the goal is to bring signature samples of the same user closer together in the feature space while ensuring separation from signatures of different users. Secondly, the model is trained using the contrastive loss function, which helps fine-tune the representations of skilled forgeries by incorporating contrastive losses and enabling effective hard negative mining techniques.

To the authors’ knowledge, the only work that investigates the text-based writer identification as a domain for mining knowledge for the OffSV task is that of [229]. In that work, authors trained a ResNet-8 CNN with text data of Persian language and subsequently utilized it in OffSV, but the followed approach and study had some important disadvantages. First, it provides a limited investigation of the task since it did not consider any sophisticated preprocessing in order to improve the similarity of data from the two domains. Second, the use of a different CNN architecture does not allow a direct comparison with the state-of-the-art SigNet network, in order to highlight whether the implemented transfer learning Offers any performance benefits to the OffSV task.

In contrast to the above works, the method presented here addresses the OffSV problem by utilizing the SigNet architecture with completely different training philosophy. We exploit both properly processed text data as well as specialized mapping functions through metric learning. In particular, the handwritten text data from the auxiliary domain are processed by a specially designed algorithm in order to create an auxiliary task whose data resemble more to those of the target domain (handwritten signature images). We propose this technique as a more convenient and elaborate transfer learning methodology for efficiently training any CNN model using largely available auxiliary text data in order to address the problem of limited availability of actual signature data. Also, we design a self-contained learning module based on contrastive loss that maps the signatures' features, extracted from SigNet, into an embedded space. Unlike previous metric-learning approaches, our proposed mapping module, after being independently trained using either text data or genuine signatures (i.e., without the requirement of skilled forgeries), can be directly applied to any input feature from any signature dataset. This design offers versatility and flexibility, making it applicable across different datasets without requiring additional training.

4.3 Design Philosophy

The ability to train with a small number of training samples is an implicit requirement of every practical OffSV system. One convenient approach to build an effective feature extractor for the signature images is to design a Writer-Independent (WI) learning scheme [44]. Thus, the feature extraction stage learns how to efficiently encode the structure of the signature image. This approach is also followed when the Deep Learning models are utilized. In that case though, a large Offline signature dataset is necessary (e.g. GPDS [190]) for training the CNN models which are used to provide the feature representations of the input signature images. In this work, we demonstrate an alternative way to train deep architectures for learning the features, in order to disentangle the development of OffSV systems from the need of large signature databases, since -among other problems- privacy issues and legislation have lately made even harder to find such data publicly available. Thereby, our core idea is the exploitation of auxiliary data with large availability as substitute to the limited signature data.

The signature depicts wealthy personal information of the signatory, encompassing not only the representation of the person's name but also both the physiological writing system (hand, arm, etc.) and the psychophysical state [49]. Each person possesses a distinctive style of handwriting, whether it is the everyday writing text or the signatures [234]. The handwritten text data are far more easily available in large volumes. Therefore, handwritten text can be an appropriate source of data for the initial training of the Deep Learning systems, which then can transfer the knowledge in the target problem of signature verification.

In this work, the handwritten text data are processed suitably aspiring to emulate shapes and forms that resemble signatures. The goal is to manipulate the auxiliary data in order to simulate the target data. We are performing this by employing a properly designed processing procedure of the text data, which exposes the underlying personal information of handwriting. The proposed technique analyzes documents of handwritten text, extracts text images and uses them as the training data of a CNN that solves a writer identification problem. This initial training process leads to a baseline CNN, which is specialized in encoding handwritten signal. This training is demonstrated in Figure 4-2 – Top panel.

Following the training of the aforementioned model, we utilize it either as an out-of-the-box feature extractor or as an initialization for fine-tuning of another CNN for realization of the task of interest, incorporating a transfer learning strategy. Two of the most popular such strategies are the parameter reuse followed by fine-tuning and the learning of some kind of feature mapping. Both techniques are graphically summarized in Figure 4-2.

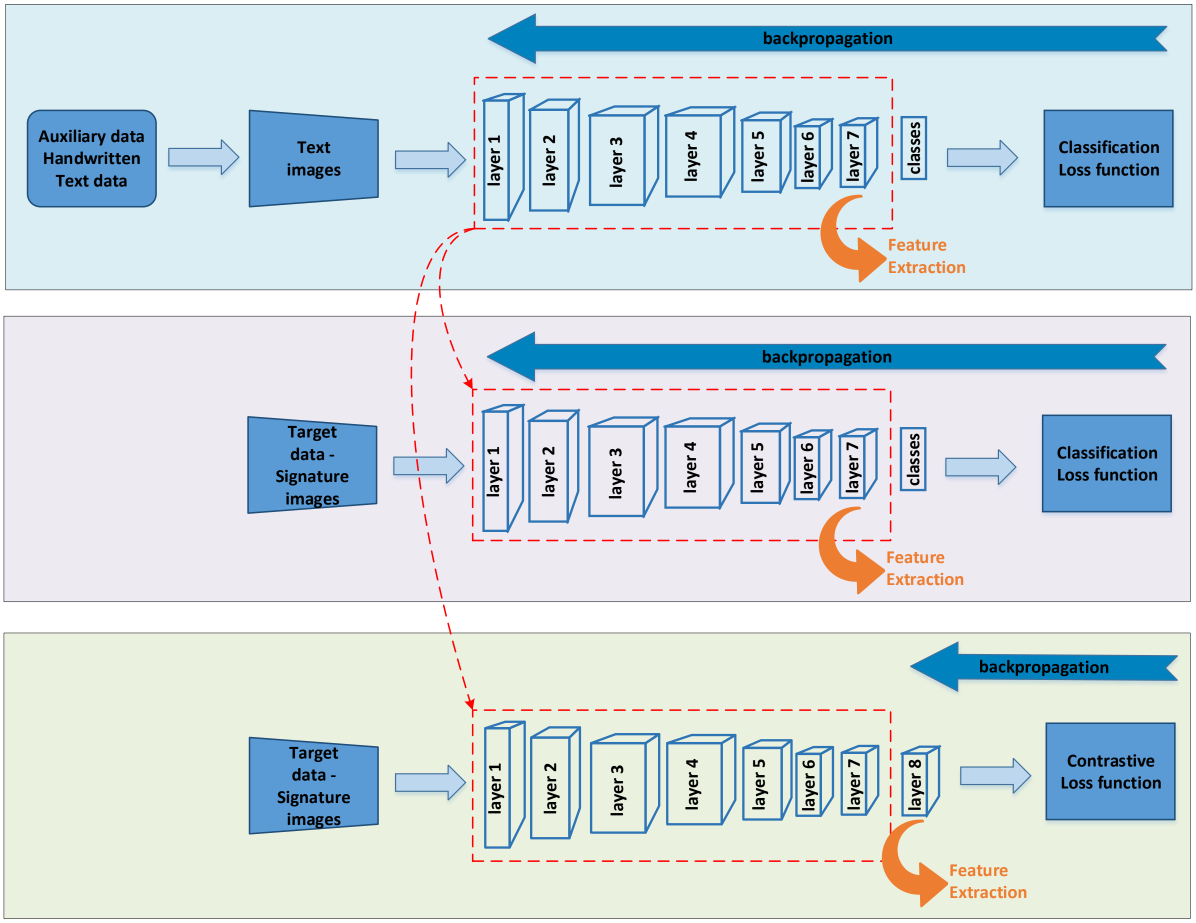

Figure 4-2: Different stages and techniques for transfer learning. Top panel: CNN is trained with the auxiliary data (text images), in the task of writer identification. Middle panel: The pretrained model is finetuned with the limited number of available signature images (target data). Ultimately, features are extracted from the penultimate layer of CNN. Bottom panel: features extracted by the pretrained model are used to learn a mapping function (Layer 8) via Contrastive Loss. In this scheme, the mapped features are discriminative but inherit metric properties tailored to the OffSV task.

In the first case, the weights of the baseline CNN (which in our case have learned to distinguish between persons' handwriting styles) are fine-tuned by end-to-end backpropagation in the new writer identification task, using signature images, as exposed in Figure 4-2 – Middle panel. This warm-starting approach essentially enables to start training the CNN from an already good initial (partial) solution and can reduce the number of signatures that are needed for accomplishing an efficient feature-extraction model for the target problem of signature verification. Still though, the performance scales with the amount of training data, since the entire CNN is trained end-to-end.

In the second direction, the CNN stripped from its final classification layer provides a feature representation of every input image, acting as a feature extractor. Given the fact that CNN learns to solve a writer identification problem using a text image as input, the model has already learned naturally discriminatory feature representations of the handwritten image information for the training set of writers. Nevertheless, the objective target of OffSV focuses basically on distinguishing between genuine and forgery signatures of a writer and not on distinguishing among writers. Therefore, a reorganization of the feature space driven by a similarity metric can be beneficial. The formulation of a metric learning problem using the extracted features contributes to this direction. Hence, the learned metric space and the function that maps the data to that space can be used as an additional module of the processing pipeline, following the main feature extraction step performed by the CNN. The metric learning module can be efficiently trained with less data for two reasons: a) the mapping function is itself a very small model (essentially a projection matrix) compared to a CNN, and b) is typically learned using pairs or triplets of images as the fundamental training datum, thus effectively increasing the number of available training examples for a given number of signature images. Therefore, the metric learning module can both address the limited sample availability and better encapsulate the relative similarities between signatures in the form of Euclidean distances between corresponding feature mappings, something advantageous in the OffSV task. This stage is illustrated on Figure 4-2 – Bottom panel. In this work, the mapping function is learnt via an optimization problem with Contrastive Loss [28] that utilizes pairs of features, labeled as similar or dissimilar. The objective of the optimization is to learn a function that maps the similar features close together in the latent space, while increasing the Euclidean distance of the mappings from dissimilar features. The similarity relationship (label) between the features of the pairs is determined from the writer's ownership of the corresponding images. So, all pairs of images from a single writer are considered similar, while pairs stemming from different writers are labeled as dissimilar. Since the mapping is obtained from the optimization of contrastive loss, the extracted features incorporate some sort of similarity metric. Thus, the mapped features can be essentially used to distinguish between different writers without them necessarily belonging to the utilized training set. Therefore, after learning the mapping function, it is then used for embedding the vectors generated by the CNN feature extractor for any new input image, to the final feature space.

In the final stage of the proposed processing pipeline, a classification stage implements the actual OffSV task, inferring on the validity of the processed signature. To that purpose, the vector representations of the signature images are processed by writer dependent (WD) SVM classifiers. Each of the WD classifiers is trained with the features stemming from genuine signatures of one registered writer, and some randomly selected genuine signatures from other writers, commonly called random forgeries. An important characteristic of this scheme is that there is no need for skilled forgery samples in order to train the WD models, with obvious practical advantages from an operational point of view. The different training stages of the proposed OffSV system are outlined in Figure 4-3.

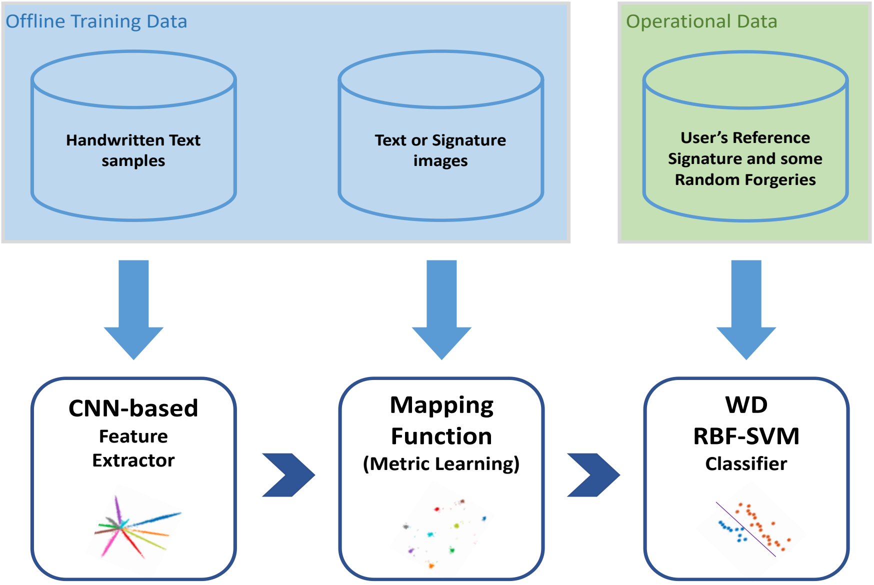
Offline Training Data
Handwritten Text samples
Text or Signature images
Operational Data
User's Reference Signature and some Random Forgeries
CNN-based Feature Extractor
Mapping Function (Metric Learning)
WD RBF-SVM Classifier

Figure 4-3: Overview of the different training stages of the proposed OffSV system with the respective data involved in each one.

4.4 Methodology

4.4.1 Preprocessing of handwritten text images

There are many sources of images with handwritten text in the public domain. An easily accessible source which was used in this work is the CVL dataset which is a public Offline handwritten text database [235] with numerous writers. The CVL-database consists of image-forms with cursively handwritten German and English texts. It contains 310 writers with 5 to 7 pages of text for each one. Each page consists of a form filled with pre-defined text, containing between 5 and 10 lines of text on average.

The goal is to extract multiple image samples from each form, which contains handwritten text. The extracted images should be in a format that can convey distinctive information of the writer's handwriting style, without necessarily including full words. Thus, there is no need for optical character recognition (OCR) or any similar language-dependent preprocessing. Therefore, we opted for a procedure of extracting Solid Stripes of Text (SSoT) from the handwritten text, which includes the following stages:

a. Convert the forms to grayscale.
b. Detect and extract horizontal stripes of text from the forms.
c. Removal of spaces between the handwritten words in each isolated horizontal stripe.

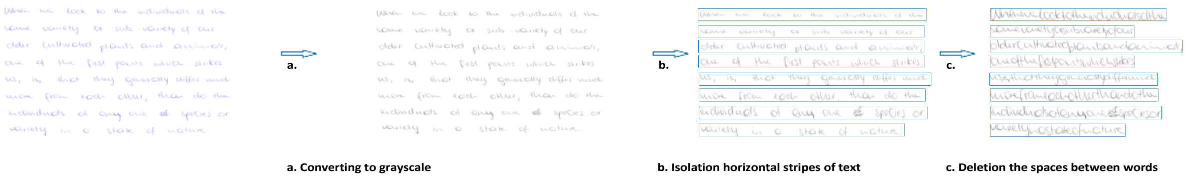
a.
b.
c.
a. Converting to grayscale
b. Isolation horizontal stripes of text
c. Deletion the spaces between words

Figure 4-4: Overview of the preprocessing of the text images. The extraction of Solid Stripes of Text (SSoT) from a page with handwritten text consists of three steps: a) conversion of the image into grayscale, b) detection and isolation of stripes of text following the horizontal direction, c) detection and deletion of empty spaces among handwritten words in each horizontal stripe in order to obtain Solid Stripes of Text.

A graphical summary of the preprocessing of text images is illustrated in Figure 4-4. In the first step, the RGB images-forms are converted into grayscale. This is necessary because the forms in the database are scanned in color, written with pens of various colors. Given the fact that the persons usually write across a generally horizontal direction, it is possible to isolate the horizontal stripes of text. With the form in grayscale, the relative intensities of the pixels are utilized for detecting the horizontal boundaries of the relevant areas, separating those from the empty ones across the document's area. In particular, the standard deviation (STD) of the pixels' intensity across every row of the image is calculated. The image then is segmented into horizontal stripes with text by detecting rows of pixels with STD value greater than 20% of the maximum document's overall intensity STD value, in order to filter out the rows with no text while accounting for noise and smudges. Additionally, the detected horizontal stripes with less than 20 pixels in height are discarded as noise-induced false positives. At the end of this process the horizontal stripes with text in each document are marked.

A procedure similar to the above is subsequently used in order to also detect the spaces between words, by finding the pixel columns with small intensity STD in each horizontal stripe. Finally, the empty spaces between words are deleted and a Solid Stripe of Text (SSoT) with continuous letters is stored as a separate image for each line of text in the dataset. The followed preprocessing is necessary in order to ensure that the training samples do not end up having crops with large amounts of white space and little/no text. There are some documents in the database where the lines of text are too close to each other for the text merging process to be accurate in this simplistic form. These samples are processed normally with space removal considering that the results are similar to random crops operation. The choice of not using entire words but rather Solid Stripes of Text (SSoT) is not negatively affecting the results as the task is to recognize the handwriting style and not its textual content. It is important to note here that no further modification (e.g., scaling, rotation, etc) is performed on the extracted SSoT.

4.4.2 Simulating signature images

The target domain of interest deals with signatures images, whereas the auxiliary data are handwritten text. The strategy for the selection of text crops to train the feature extraction CNN can significantly affect the quality of the final representation, since the data essentially guide the CNN to encode the most informative visual traits for the task. With this in mind, our purpose is to generate text crops that resemble signature images as much as possible, by proper handling the Solid Stripes of Text (SSoT). The signatures usually consist of a combination of allographs and letters (i.e., symbols), especially in Latin-based languages [50]. In this manner, the SSoT, as a block of consecutive letters, can be segmented into vertical intervals to produce samples with similar form. This cropping process does not actually modify the vertical size of the letters and thus, it preserves the handwriting style properties.

The aspect ratio is a common structural feature of Offline signatures [236] and it is the most reasonable tool to manipulate the cropping process. Three different strategies of cropping the SSoT are utilized in this study, relying on the aspect ratio of the final cropped segments. Therefore, the SSoTs are cropped using different values of aspect ratio selected in three different ways. Two of the cropping strategies consider aspect ratio to be a fixed parameter. In the first, the value of aspect ratio is associated with the size of the canvas -in which the images are centered before feed the CNN- as well as the size of the input to the CNN. The second cropping strategy applies the value of the aspect ratio of the signatures' trace, estimated from three public signature datasets. The third strategy produces crops of variable aspect ratio, by selecting random aspect ratio values lying within a fixed range. An example illustration of the three cropping strategies is presented in Figure 4-5. At the end of each process, several cropped segments are generated from every single SSoT. The set of cropped segments from each cropping strategy form a different set of sample text images.

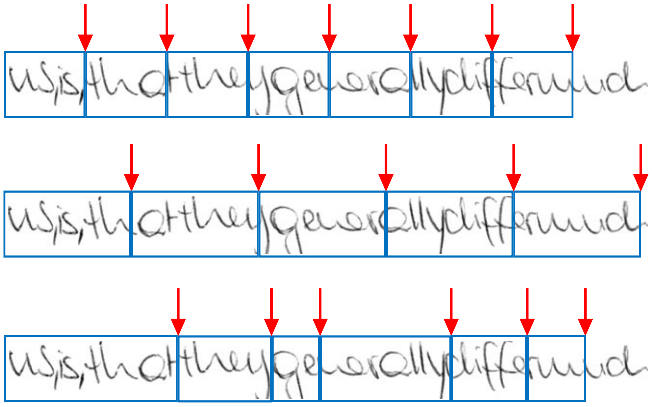

Figure 4-5: Three strategies of cropping a SSoT based on the aspect ratio value are demonstrated. The arrows indicate the position of cropping, and the boxes contain the cropped results, i.e. cropped segments. Top and Middle scheme have fixed value of aspect ratio, which is defined by the user, and so the width of each cropped segment equals to the multiplication of its height with the aspect ratio value. Bottom scheme shows the cropping process when random values of aspect ratio are utilized, and each cropped segment has a different width.

4.4.3 Geometrical normalization

The used signature datasets consist of grayscale signature images that are already extracted from the documents where they are written, so there is no need for signature extraction process. Nevertheless, some simple (pre)processing operations are always used to normalize images. The geometrical normalization steps are dedicated to noise removal and size normalization since scanned images may contain noise and the methods require the images of a fixed size. The noise is removed utilizing a combination of a gaussian filter along with OTSU thresholding [51]. The common fixed size of the images is obtained by centering each signature into a blank canvas of a predefined size, and then resize the canvas to the desired size, thus preserving each signature's original aspect ratio. The reason for adopting this implementation of centering-resizing is that it has shown to achieve better results in many OffSV systems [224], [237]. The geometrical normalization process shares exactly the same pipeline with previous works on OffSV [102], [103], [225] and the detailed steps are the following:

- Apply a gaussian filter to remove small components.
- Utilize the threshold obtained from OTSU to remove background noise.
- Center the image in a large canvas of predefined size by aligning the signatures' center of mass to the center of the canvas so as not to affect the width of strokes.
- Invert image to have black background and grayscale foreground by subtracting each pixel from the maximum brightness (i.e., white) once the background pixels are set to black and the foreground pixels are left in grayscale.
- Resize the image to the common fixed size.

The above geometrical normalization steps are implemented in every image input to CNN. Thus, both the images from the signature datasets as well as the text images emanating from the cropped segments of SSoT are processed through the geometrical normalization steps. The implementation of the same geometrical normalization for the text and signature images is intentional because the goal is to train CNN using auxiliary data of text that simulate the signatures, as an alternative of using the original signature images. The geometrical normalization has two parameters which are defined by the user: a) the size $H_{canvas} \times W_{canvas}$ of canvas and b) the common size $H_{input} \times W_{input}$ of the final images. The canvas size is a hyperparameter under study during the training of the models while the common size is determined by the input size of the CNN, as in the work of [102]. Examples of text and signature images after geometrical normalization with different canvas sizes are illustrated in Figure 4-6.

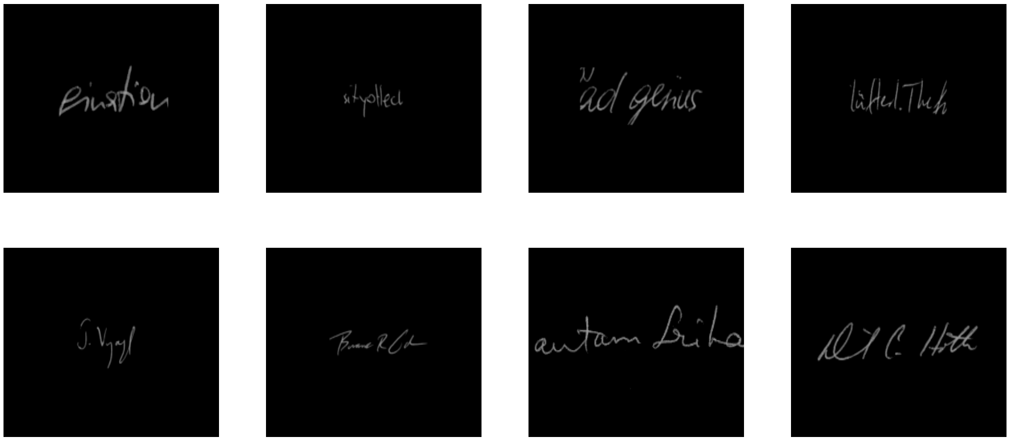

Figure 4-6: Examples of text and signature images after geometrical normalization. The top row includes processed text images, and the bottom row contains processed signature images, when different sizes of canvas are utilized.

4.4.4 SigNet CNN architecture

The SigNet CNN architecture utilized in this work is inspired by the work of [7] and is modified [102], [223], [224] in order to address the Offline signature recognition problem. The SigNet primarily is designed for solving the writer identification task. Given as input a grayscale image with handwriting, it predicts the identity of the writer among a predefined set of writers, essentially optimized for classification task. Subsequently, the SigNet model is utilized for feature extraction providing a vectorial representation for each input image. In previous works [82], [102], [103], [224], [227] the SigNet was trained using the signatures from various users therefore, it learns to distinguish between signatures from different writers in the dataset. Provided a large collection of signatures from many writers is available, the SigNet proved to be an efficient feature extractor for the signature verification problem. In this setting, the SigNet implicitly learns feature representations in a Writer-Independent manner and the representations are subsequently used by a classifier that is trained in a Writer-Dependent way.

We employ similar concept in this work, but by using text data for training the CNN. The manipulation of the text data to simulate the signatures images makes us anticipate that training the SigNet in the writer identification problem of the handwritten text images can lead SigNet to learn features that are relevant to the problem of interest, i.e., the signature verification. The proposed methodology benefits from the large availability of text data and the simple image manipulation process that simulates the signatures' form, thus eliminating the need for large-scale signature data which are nevertheless of limited availability.

The utilized CNN follows the SigNet architecture, which is summarized in the Table 4-1. SigNet takes as input a grayscale image of size 150 × 220 pixels and outputs the probabilities for the known writers' identities via a softmax operation. Following the work of [102], after every layer a batch normalization [8] is applied, followed by the ReLU non-linearity [238]. The feature extraction is incurred from layer 7 (Fully Connected layer) and the feature's dimension equals to 2048. The CNN is trained using simple translational augmentations, by taking crops of resolution 150 × 220 pixels randomly positioned inside the 170 × 242 pixels images used for training. All experiments used the same set of optimization hyper-parameters, minimizing the classification loss with Stochastic Gradient Descent with mini-batch size of 64, Nesterov Momentum factor of 0.9, while the L2-penalty with weight decay of 0.0001 is used for regularization.

Table 4-1: Overview of the SigNet CNN architecture.

SigNet architecture				
Layers			**Dimensions**	**Other parameters**
#	input	Grayscale image with handwriting	1 × 150 × 220	
1	conv	Convolution	96 × 11 × 11	stride = 4, padding = 0
	pool	Max Pooling	96 × 3 × 3	stride = 2
2	conv	Convolution	256 × 5 × 5	stride = 1, padding = 2
	pool	Max Pooling	256 × 3 × 3	stride = 2
3	conv	Convolution	384 × 3 × 3	stride = 1, padding = 1
4	conv	Convolution	384 × 3 × 3	stride = 1, padding = 1
5	conv	Convolution	256 × 3 × 3	stride = 1, padding = 1
	pool	Max Pooling	256 × 3 × 3	stride = 2
6	fc (dense)	Fully Connected	2048	
7	fc (dense)	Fully Connected	2048	
	output	Softmax	classes - number of writers	

4.4.5 Learning a feature mapping function (CoLL)

CNN addresses the classification problem of writer identification therefore, it ultimately learns to construct features that are as linearly separable as possible, in order to better facilitate the final classification layer. Therefore, such features are not necessarily equipped with a metric that reflects the similarity of the auxiliary data [239]–[242].For this purpose, the feature learning has to incorporate a ranking loss function. These type of loss functions require a similarity score between data points, such as a binary score of similar and dissimilar points. In the user identification task such notion is inherit, because the images that belong to the same person are similar and all others are dissimilar to them. Hence, the exploitation of a ranking loss during feature learning, can lead to discriminative features which in their turn, can distinguish between –in principle– any different writers (even out-of-sample writers) on any two (or more) data points. Thus, the model tries to rearrange the feature space, by learning representations with a small distance between similar data and greater distance for dissimilar ones.

There are different forms of ranking losses, distinguished by the setup of the training problem. The most popular is the Contrastive Loss or Pairwise Loss [28] which utilize pairs of data samples. Its aim is to gradually (i.e., during training) decrease the distance between similar pairs and make that larger than a margin m for the dissimilar pairs. The Contrastive Loss Layer (CoLL) is the selected implementation and therefore is applied to the extracted features (obtained by the fc 7 layer of the CNN), in order to learn a mapping function that incorporates the metric learning. Summarizing, the CNN is used as a fixed feature extractor, and it is not trained end-to-end with the Contrastive loss. This decision was made in order to accommodate fair comparisons to the baseline SigNet features in the task of OffSV. The CoLL is thus used as an individual component applied on the SigNet's features and works as a transformation layer producing discriminative features in a metric space designed to express the similarity of the data.

Therefore, the CoLL can be trained independently using pairs of features from the previously trained CNN. The similar pairs are comprised of features stemming out of two images which belong to the same writer, whilst the dissimilar pairs comprised of two features that originated from two images which appertain to different writers. It is important to note here that when a signature dataset is utilized for training the CoLL, all training pairs are constructed from genuine signatures, hence skilled forgeries are not required. Thus, a similar pair is made up with genuine - genuine for a given writer and a dissimilar pair is a genuine - random (unskilled) forgery pair. The dimensionality of the new output feature (output space) is selected to be the same as the size of the input feature (input space), i.e., a vector of 2048 elements, for as much as possible fairness in comparisons with the baseline SigNet. The parameterized measure of similarity in the output embedded space is defined as the Euclidean distance since it is simple and fast. Hence, the Contrastive loss is formulated as follows:

$$L_C = Y \cdot L_s + (1 - Y) \cdot L_D \qquad \textit{eq. 4.1}$$

where L_s is the partial loss function for a pair of similar vectors and L_D is the partial loss function for a pair of dissimilar vectors given by the relations:

$$L_s = \frac{1}{2}\left\|G(s_i) - G(s_j)\right\|^2 = \frac{1}{2}\left\|D_{ij}\right\|^2 \quad eq.\ 4.2$$

$$L_D = \frac{1}{2}\left(max\{0, m - \left\|G(s_i) - G(s_j)\right\|\}\right)^2 = \frac{1}{2}\left(max\{0, m - \left\|D_{ij}\right\|\}\right)^2 \quad eq.\ 4.3$$

with $G(\cdot)$ the CNN feature extractor, s_x the input image (in the current implementation $G(s_x)$ is a feature vector of 2048 dimensions), and m the margin of Euclidean distance in the embedded space, while Y is the label of each pair with:

$$Y = \begin{cases} 1\ , for\ \ similar\ pairs\ (same\ writer's\ data) \\ 0\ , for\ dissimilar\ pairs\ (different\ writers'\ data) \end{cases}$$

It is obvious that the Contrastive Loss is equal to the Euclidean distance between the two input features for a similar pair, otherwise is equivalent to hinge loss. The CoLL is minimized using adaptive moment estimation (Adam) method with mini-batch [243]. At each iteration, a subset of 32 similar pairs and 32 dissimilar pairs are randomly selected to create the mini-batch of size 64 and along with a learning rate of 0.0001, a gradient decay factor of 0.9, and a squared gradient decay factor of 0.99, the learnable parameters of the transformation layer are updated. The margin m outlines a radius around the point in the embedded space and the dissimilar pairs contribute to the loss only if their distance is inside this radius. The value of margin m was set to 0.1 after a grid search. The CoLL is trained using the feature representations either of the processed text images or the genuine signature images from one dataset and then, it can be applied in any feature vector from any input image utilized as a standard mapping function.

The CoLL maps the features extracted by SigNet into an output embedded space permeating the metric qualities that original features were lacking. In particular, this last layer forces the attraction of the samples owned by each writer into form clusters via the projection of the feature vectors to the new latent space. Simultaneously, the new space enforces greater distancing between features from different writers. Thus, the simple Euclidean distance in the latent space reflect the neighboring relationships in the input space according to the samples' ownership, and as a linear projection function, CoLL provides a mapping which is smoother and more coherent in the output space [28]. This essentially results into a reorganization of the feature space which is in-principle more suitable for the verification task, since the initial CNN-generated features are optimized for a specific identification task without any explicit motivation for exhibiting metric traits. An indicative 2D visualization (t-SNE projection) of the feature spaces is provided in Figure 4-7, comparing the four different feature extraction schemes descripted in Figure 4-2 evaluated for all the signatures of MCYT75 dataset. It can be easily observed that the representations produced by CoLL, especially when it is trained with signature data (Figure 4-7 (d)), provide a more uniform distribution of the different signatures overall, while maintaining very good intra-class compactness and separability between both different writers and imitators (skilled forgeries). It is important to note that the signatures used for the training of CoLL and finetune of the CNN (Figure 4-7 (b) and (d)) are different from the samples of MYCT which are mapped here, thus the latter being completely unseen data to every compared scheme. The 2D projection of the features from the CNN trained solely with text data (Figure 4-7 (a)) provides a distribution with visibly worse characteristics in terms of both inter-class separability and intra-class compactness and shape. Nevertheless, it is still remarkable that the features have far better characteristics than similar features from CNNs pre-trained in external classification tasks, as previously reported in literature [102]. This can be attributed to the special design and preprocessing of the text-based identification task that resulted into training the CNN to a truly similar task thus generating inherently more appropriate features for the OffSV. The other two schemes (Figure 4-7 (b) and (c) lie in between the previous two cases, delivering relatively good separability and distribution, but slightly inferior to that of Figure 4-7 (d). A noteworthy observation though, is that the utilization of CoLL-even with text data- improves the resulting representation. This signifies both the importance of engaging a metric-learning stage to the overall pipeline, and the affinity of the specially pre-processed text data to the signature data, since learning a metric for text clearly improves signatures' representation.

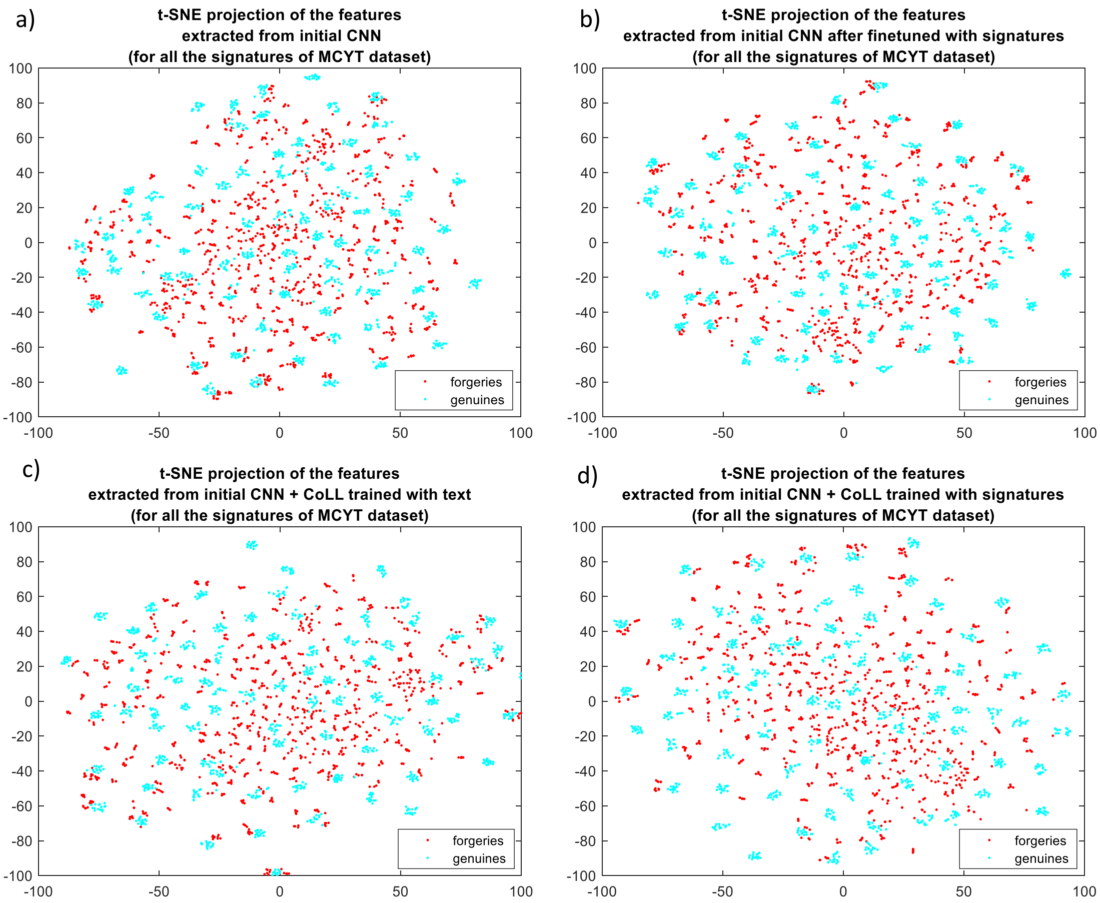

Figure 4-7: 2D projections using t-SNE of feature vectors, which are provided from the four feature extractors related to our work. The signature images are fed into the feature extractors schemes and the vectorial representations (features) are provided. Next, the vectors of 2048-dimensions are mapped into 2-dimensions through the t-SNE dimensionality reduction method. Thus, the signatures of MCYT75 dataset are represented as points on the 2D embedded space. The cyan points correspond to genuine signature while the red points correspond to skilled forgery signatures of MCYT75 dataset for all the writers. The 2D projections in a) result from features extracted from a CNN trained with text images while in b) the same CNN is finetuned with the genuine signatures of CEDAR dataset. The points in c) came from the CoLL module -placed at the top of the initial CNN of case a)- when the same text images are utilized for training both CNN and CoLL. Finally, in d) the representations are produced by CoLL, which is fed with the features from the initial CNN of case a) and CoLL is trained with the genuine signatures of CEDAR dataset, the same images that used for finetuning the CNN in case b).

4.4.6 Employing Writer-Dependent (WD) classifiers

Since a vectorized representation is constructed for every signature image via the feature extraction and mapping process, the feature is fed into a classifier that infers on the validity of the signature. In this study, the Writer-Dependent (WD) approach is followed, where one classification model is trained for each one of the writers. The signature verification problem is addressed through the respective classifier that answers the question “is the writer really who he/she claims to be?”. Consequently, the classifier tries to separate the genuine signatures of the corresponding writer from forgery signatures and thus, it works as a binary classifier among the two populations.

The SVM (Support Vectors Machine) classifier with a Radial Basis Function (RBF) kernel is utilized for constructing the classification model of each writer. The SVM is trained with a positive class ω^+ consisting of a number of genuine signature features by the writer and a negative class ω^- composed of features from genuine signatures by other writers (also called random forgeries), since the skilled forgeries of the writer are not available in a practical setting. The number of the used genuine signature features of the writer is denoted as N_{REF} and it is a measure for comparisons between OffSV systems because the smaller the reference set has needed the more preferable is the system in an everyday application. The number of the genuine signatures features of other writers is set to be the twice of N_{REF} in order to populate the negative class with more samples than the positive class during the SVM training. The reason behind this decision is to better cover the space of the negative class, since the trained model is required to reject skilled forgeries, even if such samples are not present during training.

A radial basis SVM classifier has two hyper-parameters, the γ (gamma) and *c* (cost). The parameter γ (gamma) defines how far the influence of a single training sample reaches and can be seen as the inverse of the radius of influence of support vectors. The regularization parameter *c* trades off the correct classification of training samples against the maximization of the decision function's margin. In our implementation, a holdout cross validation procedure returns the optimal writer's parameters γ and *c* minimizing the misclassification rate (loss) in the training set for every writer.

4.4.7 Accuracy metrics

Many metrics have been used in order to test the efficiency of a OffSV system, such as the False Rejection Rate (FRR), which is referred to the misclassification of a genuine signature as being a forgery, the False Acceptance Rate (FAR), which is mentioned to the misclassification of a forgery as genuine signature, and the Area Under Curve (AUC) considering the Receiver Operating Characteristic (ROC) curve drew for each writer [44]. The point where the FRR and FAR are equals (FRR=FAR) is known as the Equal Error Rate (EER). The EER describes the overall performance of a biometric system with only one demonstrated value and for that it is a very popular metric in the evaluation of OffSV systems too [44], [47], [49], [50], [195], [244]. Some researchers address the signature verification problem incorporating both skilled and random forgeries [245] in the negative population of the classifier or evaluate the performance based on each type of forgery, i.e. only using Random forgeries and only using Skilled forgery signatures in the negative class. This has an impact on the calculation of the EER value since the FAR is related to the evaluated forgeries samples. Additionally, the EER can be calculated employing user-specific decision thresholds or global decision threshold.

In this work, due to the plethora of experimental results we opted for focusing only on the Equal Error Rate (EER), obtained using optimal user-specific decision thresholds with the genuine signatures of the user and the corresponding skilled forgeries. Thus, the EER is calculated when

FRR = FAR$_{skilled}$ using user-specific decision thresholds. After training the feature extraction schemes, the vector representations of the signatures are processed by the Writer-Dependent (WD) classifiers. The training of every SVM WD classifier has been repeated 10 times with the feature representations of randomly selected Reference genuine samples. The EER results are obtained in the terms of the average and standard deviation values across these 10 experiments for the test set of signatures, i.e., the rest genuine and the skilled forgeries signatures of the user.

4.5 Experimental Setup

4.5.1 Handwritten Text Dataset

The CVL-database is a public dataset of digitized documents with hand-filled forms of text, suitable for writer identification as well as optical character recognition tasks [235]. The dataset includes 310 writers with a varying number of documents for each writer spanning from 5 to 7. First, the forms were split into a training set and a validation set, with 3 of the forms by each writer placed into the training set and 1 kept for validation. The forms were selected randomly from the available set of each writer, as some writers have more forms than others.

4.5.2 Handwritten Signature Datasets

Three popular datasets of Offline signatures are utilized in this work to assess the efficiency of the presented scheme. All the corpuses belong to Western scripts and are Latin-based. The signatures have been digitized -by means of scanning- after acquisition and they are available as grayscale images.

The first signature dataset is the publicly available CEDAR (Centre of Excellence for Document Analysis and Recognition) [56]. It consists of 55 enrolled writers with 24 genuine and 24 forgeries signatures per writer. The forgeries are a mixture of random, simple, and skilled simulated signatures. Each person signed in a square box of 50 mm by 50 mm and the forms are scanned at 300 dpi in grayscale.

The second signature dataset is the Offline version of the MCYT (Ministerio de Ciencia Y Tecnologia, Spanish Ministry of Science and Technology), known as the MCYT75 Offline Signature Baseline Corpus ("Database") and it is publicly available [57], [198]. The MCYT75 includes 75 writers with 15 genuine and 15 forgeries signatures per writer. The forgeries contributed by 3 different user-specific forgers and thus, they are skilled simulated signatures. The signatures are captured in a paper template within a 17.5 mm by 37.5 mm (height by width) frame and are digitized by means of scanning at 600 dpi in grayscale.

The third signature dataset is the Offline handwritten signature GPDS (Digital Signal Processing Group) database, which is no longer publicly available due to the General Data Protection Regulation (EU) 2016/679 ("GDPR") [189], [190], [246]. The GPDS-960 corpus began with 960 enrolled writers, having 24 genuine and 30 forgeries signatures per writer. The forgeries signatures marked as skilled since they made by 10 forgers from 10 different genuine specimens. The signatures were collected using black or blue ink on white paper in two different bounding boxes evenly distributed, one box is 18 mm height by 50 mm width and the other is 25 mm height by 45 mm width. There are two versions of the dataset based on the image type, the grayscale version (GPDS960GRAY), which is scanned at 600 dpi, and the black-and-white version (GPDS-160, GPDS-300 with 160 and 300 users respectively), which is scanned at 300 dpi. During the move to the grayscale version of the dataset though, 79 users and 143 imitations of the remaining signers were lost. Thus, the GPDS960GRAY signature database consists of 881 users. The standard practice for evaluation with GPDS though [44], [195], is to use a subset with the first 300 users of the GPDS960GRAY called GPDS300GRAY, which is what we utilized in this work for compatibility with previously published results.

4.5.3 Constructing Training Sets

As already mentioned, three different strategies are evaluated for cropping SSoT into text samples. For the first case, the aspect ratio is set in the value of 1.4 since this is the aspect ratio of the input images in the CNN, as defined in the standard SigNet architecture. In the second case, the aspect ratio arises from the mean aspect ratio of the signatures' trace in the three used signature datasets and is set to 2.2. In the third strategy, the aspect ratio takes a random value in each cropped SSoT, with the restriction that the width of the final crop should be between 350 pixels and 50 pixels. Finally, three corresponding sets of text images are formed by applying the above settings, having about seventy thousand training and twenty-five thousand validation images for the first and third set, and about forty-five thousand training and fifteen thousand validation images from the second set.

The geometrical normalization is controlled mainly by two parameters, the common final size of the images and the size of the canvas in which the images are centered. The final size of the images determined from the input of the CNN. The CNN takes as input a grayscale image with 150 pixels width and 220 pixels height. Nonetheless, the images are resized to resolution of 170 × 242 pixels in the end so that to apply random crops of size 150 × 220 as data augmentations during training of the CNN. The canvas size specifies the area where the image's center of mass is aligned to. The centering of the image in a large canvas before resizing serves the persistence of the stokes' width but poses the problem that an image which is larger than the canvas should be scaled down and also, some details can be lost in the very small images. Empirically, the conjunction of centering and resizing as opposed to only resizing results in superior performance of OffSV systems [224], [237]. Thus, the canvas size is a parameter of crucial importance for the performance of the system. In this work, different sizes of canvas were investigated covering a large range of values, though all with the same aspect ratio, which is the same as the CNN's input image, equal to W/H=1.4. Specifically, the tested canvases are of dimensions 300×430, 400×560, 500×710, 600×850 and 730×1042 pixels. Since our study relies on the exploitation of auxiliary data for efficient CNN learning schemes, the utilization of different canvas sizes also allows the generation of multiple training images from the same original set of text images. This enables us to investigate the effects of the relationship between the spatial distribution of the signals in the target and auxiliary domains, and whether this should be taken into consideration when preparing the external data for knowledge learning or it can be addressed via more general guidelines.

In this study, we tried multiple combinations of cropping and geometrical normalization settings to reveal the influence of image preprocessing to the accuracy of an OffSV system, and also indicate best practices for future research efforts. First, 15 different training sets are constructed based on the three text sets (from the three cropping strategies) and five canvas sizes, as presented on Table 4-2. Additional training sets for the CNN can be created by merging the existing sets. Therefore, the union of text images from all cropping strategies can form a new training set, as also images from the first and the second cropping strategy. Finally, the union of sets from each individual cropping can form new training sets (using all the canvas sizes), as demonstrated on Table 4-3. Overall, 20 different training sets of text images are investigated for their efficiency in the training of CNN models. The same procedure is executed for the validation images with the difference that the final 150 × 220 pixels samples are cropped from the center of the 170 × 220 images.

Furthermore, the genuine signatures from the CEDAR or MCYT75 datasets are used in the same spirit, creating 12 (6 with CEDAR + 6 with MCYT75) signature training sets. These sets are utilized either for finetuning the CNN after its training with text data or training the CoLL module to learn the mapping function, and they also constitute external data (of the same nature though) to the final verification task. The combinations for creating the 12 signature training sets are summarized in Table 4-4 From the genuine signatures of each signer, one genuine signature is used for the validation set and the rest constitute the training set in every single set. Once again, after the centering step, the training images of size 170 × 242 pixels are cropped randomly in size of 150 × 220 and the validation images are center cropped to obtain the final 150 × 220 images.

For better clarity regarding the evaluation protocol, it is important to note that the signatures used for the target test verification task in each experiment, are processed with only one specific canvas size that corresponds to the respective dataset, as proposed in the works of [102], [103]. These canvases are related to specific features of each dataset which are linked to the acquisition techniques followed in each case and are closely followed here for the sake of fair comparisons. Hence, the signatures of CEDAR utilize a canvas size of 730 × 1042 pixels, the signatures of MCYT75 use a canvas with resolution of 600 × 850, and the signatures of GPDS300GRAY are processed with a canvas of 952 × 1360 pixels. Finally, all images are center cropped with resolution 150 × 220 pixels in order to be processed by the trained CNN.

Table 4-2: Text Sets generated with single canvas sizes.

# Text sets	Cropping scenario based on aspect ratio	Canvas size (Height × Width)	# Text sets	Cropping scenario based on aspect ratio	Canvas size (Height × Width)	# Text sets	Cropping scenario based on aspect ratio	Canvas size (Height × Width)
1.	1.4	300 × 430	**6.**	2.2	300 × 430	**11.**	random	300 × 430
2.		400 × 560	**7.**		400 × 560	**12.**		400 × 560
3.		500 × 710	**8.**		500 × 710	**13.**		500 × 710
4.		600 × 850	**9.**		600 × 850	**14.**		600 × 850
5.		730 × 1042	**10.**		730 × 1042	**15.**		730 × 1042

Table 4-3: Text Sets generated with multi canvas sizes by merging the Text Sets that generated with single canvas sizes.

# Text sets	Cropping scenario based on aspect ratio	Canvas sizes (Height × Width)	merge Text sets
16.	1.4	300 × 430, 400 × 560, 500 × 710, 600 × 850, 730 × 1042	1 – 5
17.	2.2		6 – 10
18.	1.4 & 2.2		1 – 10
19.	random		11 – 15
20.	all		1 – 15

Table 4-4: Sign Sets based on the hyperparameter of canvas size using the genuine signatures of CEDAR or MCYT75 datasets.

# Signature sets	Canvas size(s) (Height × Width)	merge Sign sets
I.	300 × 430	-
II.	400 × 560	-
III.	500 × 710	-
IV.	600 × 850	-
V.	730 × 1042	-
VI.	300 × 430, 400 × 560, 500 × 710, 600 × 850, 730 × 1042	I – V

4.5.4 Assessing Different Mechanisms of Feature Learning

As mentioned earlier and summarized in, there are several ways to obtain the feature-level representation of the signature images using the trained CNN. In the spirit of a thorough evaluation, we opted for assessing all levels of possible feature learning schemes that lie in the described framework. Thus, in addition to the fully-trained pipeline with CoLL, we also evaluated the effectiveness of the representations produced directly by the trained (with text) CNN without any modifications, as also the representations produced if the CNN is further fine-tuned with signatures in the traditional way. Finally, since CoLL can be trained both with signatures and text data, we evaluated and compared both strategies in the respective experimental settings.

4.5.5 Training WD classifiers

After the feature extractors of CNN and CoLL are trained, Writer-Dependent (WD) classifiers are also trained with the feature representations of the signatures. Thus, feedforward propagation is performed for every training image until the feature extraction layer of each experimental case. The extracted feature vectors of 2048 dimensions are used as input to the classifiers. The WD binary classifiers are Radial Basis Function Support Vectors Machines (RBF SVM). The RBF SVM is trained for each writer using a number N_{REF} of Reference signatures' features of the writer along with twice this number of Random forgeries signatures' features, picked randomly from the genuine signature pool of other writers in the dataset. Finally, the SVM (trained) model is evaluated using feature vectors from the remaining genuine writers' signatures and from the skilled forgeries signatures of the writer. The features are used either as is, or normalized and centered to zero mean and unit variance along each dimension using the global mean and standard deviation. This is pronounced in the corresponding results in the "sd" column (True or False).

The evaluation of the signature verification systems in the WD manner is quantified using the Equal Error Rate (EER). The metric of EER using user-specific decision thresholds is calculated when the False Acceptance Rate (FAR) is equals to the False Rejection Rate (FRR) for each user, considering the respective genuine and skilled forgeries signatures of the user. For every trained feature extractor, the SVM WD classifier of the user is trained 10 times with different Reference genuine signatures. Finally, the average EER value as well as the standard deviation across the 10 runs are reported.

4.6 Experimental Results

4.6.1 Training CNN only with Text images

The first experimental setting involves the features generated by the trained CNN without any modification to better suit the target task. In all experiments, the CNN was initialized using He-

Normal [247], and trained from scratch using the text image sets 1-20 (Table 4-2 and Table 4-3), obtaining 20 trained models. In each CNN, the writer's identity is inferred from the text image via a typical classification task of 310 classes, which is the number of the writers in the text dataset. The accuracy obtained for the 20 different training sessions is demonstrated in Figure 4-8. It is important to note that the accuracy is calculated in the level of individual text -generated- images and is not averaged across whole documents, as it is the usual approach for text-based identification systems [235]. It is evident that the size that the text strip occupies in the final image plays a crucial role in the obtained accuracy, with the smaller canvases (e.g., sets 1, 6, 11, 20) that have a larger portion of text inside the image bearing the best performance. In line with that observation is the fact that if the text cutouts are resized to the full input image's dimensions, the accuracy gets above 90% (however in that case the performance is unsatisfactory at signature verification task). The writer identification task using text is secondary and out of the scope of this work though and thus, we did not perform a thorough analysis of the obtained performance since the sole objective of this phase is to generate CNNs that are effective in the OffSV task. In the subsequent stage and for each configuration, the final layer of the respective model is removed, and the CNN is used as a fixed function that generates a global feature vector for each input signature image. In order to quickly assess the quality of the learned representations, WD classifiers are trained on each of the three signature datasets with the extracted features, and the EER values are presented in the next error bar diagrams of Figure 4-9, Figure 4-10, and Figure 4-11.

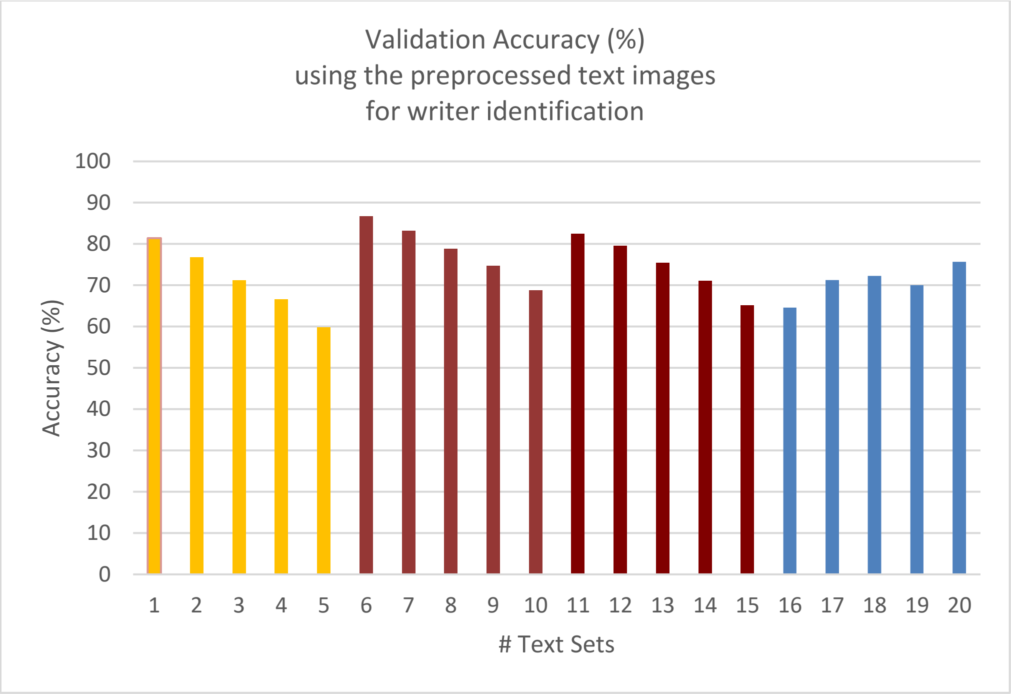

Figure 4-8: Validation Accuracy (%) for the 20 generated Text Sets. The geometrical normalization steps are applied to the preprocessed text images of the CVL-database, and the CNN predicts the writer considering only one validation image (individual predictions are not consolidated into document-level predictions).

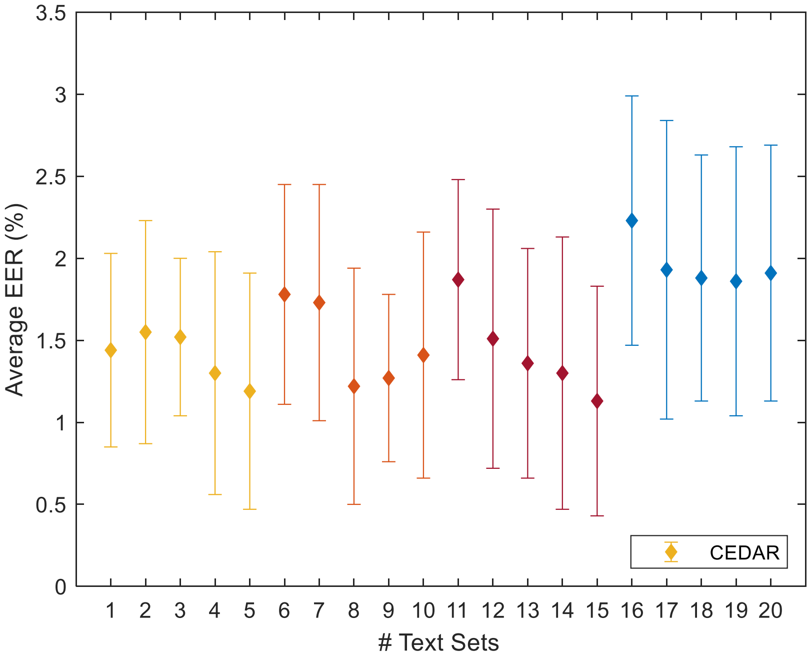

Figure 4-9: Error bar diagram of EER (%) for the CEDAR dataset using the 20 different CNN models, with N_{REF}=10 and 10 iterations with random reference genuine signatures for every experiment.

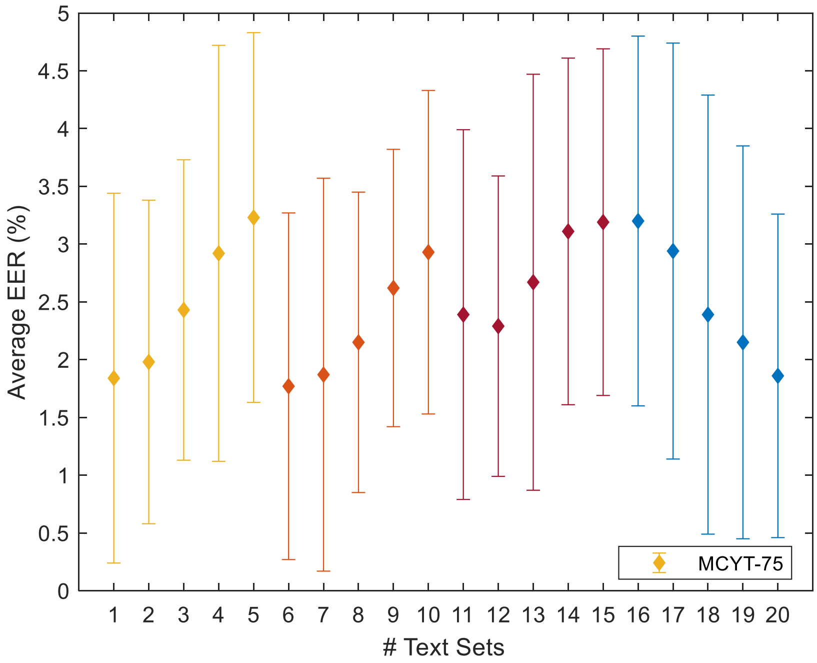

Figure 4-10: Error bar diagram of EER (%) for the MCYT75 dataset using the 20 different CNN models, with N_{REF}=10 and 10 iterations with random reference genuine signatures for every experiment.

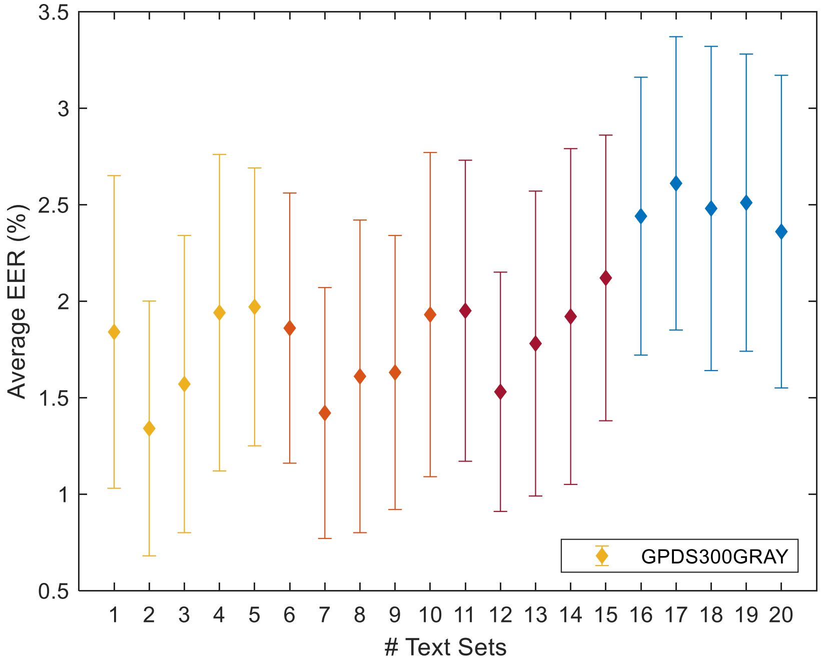

Figure 4-11: Error bar diagram of EER (%) for the GPDS300GRAY dataset using the 20 different CNN models, with N_{REF}=12 and 10 iterations with random reference genuine signatures for every experiment.

From the signature verification results some interesting observations can be made. First, there are instances where slightly better performance can be obtained using single-canvas Text sets (i.e., sets 1-15), compared to mixed-canvas sets 16-20. It is known that the signing procedure depends on many parameters, including both the signer's behavior and the conditions during the act of signing. Even though the behavioral state of each signer cannot be regulated, the acquisition conditions under the recording of each dataset like the type of paper, the available pens, the signature boxes, the signers' posture and even the environmental conditions, can have an effect on the signatures, reflected as dataset-level characteristics. Thus, such implicit dataset-specific traits could be coincidentally matched by a CNN trained via one specific canvas size and one cropping strategy that better fits with the dataset, but such a mechanism has limited practical importance since it requires prior knowledge of the reference dataset at training time.

The second and most important observation is that somewhat better results are obtained when all cropping strategies are utilized together (i.e., in the Text set 20). In that case, the training set is larger than any other and most importantly, it includes all the types of crops, thus priming the trained CNN to generate features that express more general visual cues of the handwritten signal. In the same manner, set 18, which is essentially a merge of 16 and 17, is more effective than each of them. This remark extends to the superior performance obtained when utilizing random aspect ratio values instead of a single aspect ratio value, which again can be justified due to the greater generalization of cases that the Text set 19 includes against both Text sets 16 and 17. Therefore, it seems that the CNN models that learn from more general Text sets, have the potential to consistently perform well in all three datasets.

From the above results, we can point out the more efficient baseline CNN models for the final target task of signature verification. In order to keep the number of experiments manageable, only these CNN models are used for the next sections that we investigate the following stages of the proposed pipeline. Thus, for the CEDAR and MCYT75 datasets, which have about the same number of signatures (and they are much smaller than GPDS300GRAY), only one CNN model from each cropping strategy is selected, while the last five (16-20) CNN models are selected for all three datasets. These five last models serve our purpose of designing an OffSV system that can be sufficient across datasets. The selected trained CNN models -that we'll utilize in the next experiments- as well as the corresponding EERs (for the first experiment) are summarized in Table 4-5.

Table 4-5: The Selected Initial CNNs.

Test Signature dataset		CNN (trained with text)	WD classifiers		trained CNN models
db name	**canvas size**	**#Text set**	**sd**	**EER**	**name**
CEDAR	730 × 1042	5.	False	1.19 (± 0.72)	M5
		8.	False	1.22 (± 0.72)	M8
		15.	False	1.13 (± 0.70)	M15
		16.	False	2.23 (± 0.76)	M16
		17.	False	1.93 (± 0.91)	M17
		18.	False	1.88 (± 0.75)	M18
		19.	False	1.86 (± 0.82)	M19
		20.	False	1.91 (± 0.78)	M20
MCYT75	600 × 850	1.	False	1.84 (± 1.60)	M1
		6.	False	1.77 (± 1.50)	M6
		12.	False	2.29 (± 1.30)	M12
		16.	False	3.20 (± 1.60)	M16
		17.	False	2.94 (± 1.90)	M17
		18.	False	2.39 (± 1.80)	M18
		19.	False	2.15 (± 1.70)	M19
		20.	False	1.86 (± 1.40)	M20
GPDS300GRAY	952 × 1360	16.	False	2.44 (± 0.72)	M16
		17.	False	2.61 (± 0.76)	M17
		18.	False	2.48 (± 0.84)	M18
		19.	False	2.51 (± 0.77)	M19
		20.	False	2.36 (± 0.81)	M20

4.6.2 Finetuning CNN with Signature images

As a next step, the selected initial CNN models are finetuned with the Signature sets obtained applying the parameters of Table 4-4. Since the signatures used for finetune are considered as external data from different signers than those that engage with the target OffSV task, the data configuration in the experiments that involve external signature data is as follows: In one setting, the Signature sets obtained using the CEDAR dataset and utilized for finetuning, while the evaluation is performed in the datasets of MCYT75 and GPDS300GRAY. In a separate setting, the Signature sets based on the MCYT75 dataset are used for finetuning and the systems are evaluated on CEDAR and GPDS300GRAY datasets. The finetuning is performed for 20 epochs and the freezing of the initial layers is utilized for the first epochs considering the best performance in each case. The optimization was achieved with a learning policy of decreasing learning rate by a factor of 10 after 10 epochs with initial value of 0.001, along with Nesterov Momentum factor of 0.9, weight decay of 0.0001, and batch-size of 16. The results are reported for each dataset in the following Table 4-6, Table 4-7, Table 4-8. The column of initial CNN, in the Tables, indicates the CNN model, which is used as the initial pre-trained model (with the text data) for the finetuning using the signature data.

Table 4-6: EER results for CEDAR (finetuning with Signature the initial CNNs).

Test Signature dataset		**initial CNN (trained with text)**	**CNN (finetuned with sign)**	**WD classifiers with $\mathcal{N}_{REF}$ = 10**	
db name	**canvas size**	**#Text set model**	**#Sign MCYT set**	**sd**	**EER**
CEDAR	730 × 1042	M5.	I.	False	2.60 (± 0.82)
			II.	False	2.40 (± 0.82)
			III.	False	2.39 (± 0.85)
			IV.	False	2.42 (± 1.00)
			V.	False	2.15 (± 0.95)
		M8.	I.	False	2.20 (± 0.90)
			II.	False	2.24 (± 0.87)
			III.	False	1.58 (± 0.74)
			IV.	False	1.51 (± 0.76)
			V.	False	1.44 (± 0.83)
		M15.	I.	False	2.50 (± 0.85)
			II.	False	2.50 (± 0.64)
			III.	False	2.32 (± 0.68)
			IV.	False	2.41 (± 0.88)
			V.	False	2.20 (± 0.83)
		M16.	VI.	False	2.26 (± 0.66)
		M17.	VI.	False	2.15 (± 0.91)
		M18.	VI.	False	2.41 (± 0.80)
		M19.	VI.	False	1.95 (± 0.68)
		M20.	VI.	False	2.05 (± 0.86)

Table 4-7: EER results for MCYT75 (finetuning with Signatures the initial CNNs).

Test Signature dataset		**initial CNN (trained with text)**	**CNN (finetuned with sign)**	**WD classifiers with $\mathcal{N}_{REF}$ = *10***	
db name	**canvas size**	**#Text set model**	**#Sign CEDAR set**	**sd**	**EER**
MCYT75	600 × 850	M1.	I.	False	1.83 (± 1.20)
			II.	False	1.74 (± 1.20)
			III.	False	1.91 (± 1.50)
			IV.	False	1.99 (± 1.40)
			V.	False	2.03 (± 1.30)
		M6.	I.	False	1.65 (± 1.30)
			II.	False	1.68 (± 1.40)
			III.	False	1.94 (± 1.40)
			IV.	False	2.12 (± 1.30)
			V.	False	2.33 (± 1.50)
		M12.	I.	False	1.52 (± 1.30)
			II.	False	1.80 (± 1.40)
			III.	False	1.97 (± 1.50)
			IV.	False	2.00 (± 1.50)
			V.	False	2.38 (± 1.50)
		M16.	VI.	False	2.20 (± 1.50)
		M17.	VI.	False	2.54 (± 1.40)
		M18.	VI.	False	2.19 (± 1.50)
		M19.	VI.	False	2.08 (± 1.50)
		M20.	VI.	False	1.77 (± 1.60)

Table 4-8: EER results for GPDS300GRAY (finetuning with Signatures the initial CNNs).

Test Signature dataset		**initial CNN (trained with text)**	**CNN (finetuned with sign)**		**WD classifiers with $\mathcal{N}_{REF} = 12$**	
db name	**canvas size**	**#Text set model**	**#Sign set**	**#Sign db**	**sd**	**EER**
GPDS300GRAY	952 × 1360	M16.	VI.	CEDAR	False	2.64 (± 0.76)
		M17.	VI.		False	2.72 (± 0.66)
		M18.	VI.		False	2.31 (± 0.78)
		M19.	VI.		False	2.52 (± 0.82)
		M20.	VI.		False	2.21 (± 0.68)
		M16.	VI.	MCYT75	False	3.01 (± 0.90)
		M17.	VI.		False	3.07 (± 0.84)
		M18.	VI.		False	2.69 (± 0.80)
		M19.	VI.		False	3.18 (± 0.83)
		M20.	VI.		False	2.86 (± 0.96)

The finetuning with about one thousand signature images improves the performance in most of the cases, as it is expected. Each Signature set consists of about one thousand signature images since there are 55·24=1320 and 75·14=1050 genuine signatures in CEDAR and MCYT75 respectively. Exceptions are the Sign sets VI that they have quintuple number of images because they are obtained as a merger of the others sets. The performance of the initial model is crucial for the performance of the finetuned model, meaning that, in general, an initial model providing good results leads also to good results after the finetuning. Ultimately, the finetuning procedure leads to an increase of the performance even though the rise cannot be characterized as significant.

4.6.3 Training CoLL with Text images

Next, alternatively to traditional finetuning, the CoLL module is employed in order to apply a feature mapping on the extracted CNN features. In this scheme, the CNN models are trained with text data (presented at Table 4-5) and then, they are used as a fixed feature extractor. The CoLL module is fed with the CNN features and trained with pairs of features using contrastive loss in order to learn the mapping function. The first option to train the CoLL module is to also utilize text images. In this context, one Text set (from the 1-20) is utilized with the selected CNN model and the extracted features are used for creating the feature pairs and for training the CoLL. The column of initial CNN indicates the selected CNN model (Table 4-5), which is used for feature extraction before CoLL. The Text sets that are used rely on the selected CNN model in the basis of having the same cropping strategy, so as again limit the number of experimental cases. For example, when the selected CNN model is trained from the Text set 1, the relevant Text sets for training the CoLL are the sets 1-5 because only these originated from the same cropping strategy. The EER is computed and the results for the three signature datasets are provided in the Table 4-9 and Table 4-10, while Table 4-11 demonstrates the difference of using a CNN scheme versus a CNN-CoLL scheme (CoLL is added after fixed CNN) when both schemes share the same training text sets. The addendum of CoLL module at the top of CNN feature extractor increases the performance of the OffSV systems and it appears to have more significant impact than the previous finetuning strategy, although signature images are not utilized at all during training.

Table 4-9: EER results for CEDAR (CoLL trained with Text).

Test Signature dataset		**CNN (trained with text)**	**CoLL (trained with text)**	**WD classifiers with $\mathcal{N}_{REF}$ = *10***	
db name	**canvas size**	**#Text set model**	**#Text set**	**sd**	**EER**
CEDAR	730 × 1042	M5.	1.	True	1.06 (± 0.62)
			2.	True	1.10 (± 0.54)
			3.	True	0.99 (± 0.74)
			4.	True	1.19 (± 0.66)
			5.	True	1.15 (± 0.63)
		M8.	6.	True	1.17 (± 0.84)
			7.	True	1.18 (± 0.76)
			8.	True	1.12 (± 0.84)
			9.	True	1.20 (± 0.73)
			10.	True	1.21 (± 0.86)
		M15.	11.	True	1.27 (± 0.84)
			12.	True	1.18 (± 0.79)
			13.	True	1.23 (± 0.85)
			14.	True	1.12 (± 0.73)
			15.	True	1.13 (± 0.59)

Table 4-10: EER results for MCYT75 (CoLL trained with Text).

Test Signature dataset		CNN (trained with text)	CoLL (trained with text)	WD classifiers with N_{REF} = 10	
db name	canvas size	#Text set model	#Text set	sd	EER
MCYT75	600 × 850	M1.	1.	True	1.62 (± 1.20)
			2.	True	1.69 (± 1.30)
			3.	True	1.47 (± 1.30)
			4.	True	1.66 (± 1.40)
			5.	True	1.60 (± 1.30)
		M6.	6.	True	1.54 (± 1.30)
			7.	True	1.64 (± 1.40)
			8.	True	1.47 (± 1.50)
			9.	True	1.48 (± 1.30)
			10.	True	1.71 (± 1.50)
		M12.	11.	True	2.05 (± 1.30)
			12.	True	1.86 (± 1.30)
			13.	True	1.82 (± 1.50)
			14.	True	1.88 (± 1.40)
			15.	True	1.99 (± 1.10)

Table 4-11: EER results for CEDAR and MCYT75 with N_{REF}=10 as well as GPDS300GRAY with N_{REF}=12 for CNN and CoLL trained with the same Text sets.

Test Signature dataset		Train Set	CNN (trained with text)	CoLL (trained with text)
db name	canvas size	#Text set	EER (WD)	EER (WD)
CEDAR	730 × 1042	16.	2.23 (± 0.76)	1.86 (± 0.72)
		17.	1.93 (± 0.91)	1.61 (± 0.65)
		18.	1.88 (± 0.75)	1.49 (± 0.76)
		19.	1.86 (± 0.82)	1.51 (± 0.81)
		20.	1.91 (± 0.78)	1.65 (± 0.78)
MCYT75	600 × 850	16.	3.20 (± 1.60)	2.26 (± 1.60)
		17.	2.94 (± 1.90)	2.21 (± 1.60)
		18.	2.39 (± 1.80)	2.06 (± 1.50)
		19.	2.15 (± 1.70)	1.54 (± 1.70)
		20.	1.86 (± 1.40)	1.65 (± 1.60)

Test Signature dataset		Train Set	CNN (trained with text)	CoLL (trained with text)
db name	canvas size	#Text set	EER (WD)	EER (WD)
GPDS300GRAY	952 × 1360	16.	2.44 (± 0.72)	2.09 (± 0.82)
		17.	2.61 (± 0.76)	2.23 (± 0.64)
		18.	2.48 (± 0.84)	2.17 (± 0.88)
		19.	2.51 (± 0.77)	2.25 (± 0.71)
		20.	2.36 (± 0.81)	2.30 (± 0.76)

Table 4-11 reflects the effectiveness of CoLL module in the system, since EER values are lower in every case using the same training data and regardless of the canvas size. To support this claim, we apply a statistical analysis of the experimental results based on common omnibus tests in order to confirm whether the considered models significantly outperform the baseline models. Following the work of [248], the popular non-parametric Friedman test and the parametric repeated measures ANOVA (Analysis of Variance) are executed for calculating the p-value [249] for the ten repetitions of each WD classifier, using the same permutations of reference/test samples. The p-values (both ANOVA and Friedman results) lie in orders of magnitude between 1E-6 and 1E-2 for all 15 cases of Table 4-11, indicating that the obtained difference in performance is statistically significant. As an example, ANOVA for the results corresponding to Text set 20 have p-values equal to 4.5E-3, 6.3E-6, and 1.7E-2 for CEDAR, MCYT75, and GPDS300GRAY respectively, while for the case of Text set 17 the p-values of Friedman tests are 1.8E-3, 3.7E-2 and 5.6E-3 for the same datasets. The important finding of the current experiments here is that by simply employing CoLL, using exactly the same training images, leads to superior results due to the more favorable distribution of the features in the latent space. This behavior comes in contrast to the regular finetuning, which can deliver a performance improvement only in specific combinations of text and signature datasets. It is important to note again that the dimensionality of the features after the CoLL was intentionally kept the same (i.e., 2048-dim feature), so as to highlight the role of the learned mapping regardless of any dimensionality reduction that can be incorporated to the mapping function if needed. This way, the comparisons are fair and can better justify the effectiveness of CoLL in the overall framework.

4.6.4 Traininng CoLL with Signature images

In the last series of experiments, the CoLL is trained using the features from signature images. In that case, signature images from the sets of Table 4-4 are processed by one CNN model from Table 4-5 and the obtained representations are utilized for training a CoLL module. The CEDAR or MCYT75 signature datasets are utilized for training and in each case the other two signature datasets are used for evaluation, following the same rationale as in section 4.6.2 for the selection of the signature training sets. The experimental results in terms of EER are presented in the next Table 4-12, Table 4-13, and Table 4-14 for the three test signature datasets.

Table 4-12: EER results for CEDAR (CoLL trained with Sign).

Test Signature dataset		**CNN (trained with text)**	**CoLL (trained with sign)**	**WD classifiers with $\mathcal{N}_{REF}$ *= 10***	
db name	**canvas size**	**#Text set models**	**#Sign MCYT set**	**sd**	**EER**
CEDAR	730 ×1042	M5.	I.	True	1.23 (± 0.75)
			II.	True	1.27 (± 0.76)
			III.	True	1.13 (± 0.65)
			IV.	True	1.20 (± 0.75)
			V.	True	1.12 (± 0.68)
		M8.	I.	True	1.23 (± 0.78)
			II.	True	1.35 (± 0.64)
			III.	True	1.32 (± 0.52)
			IV.	True	1.21 (± 0.61)
			V.	True	1.09 (± 0.58)
		M15.	I.	True	1.15 (± 0.73)
			II.	True	1.20 (± 0.71)
			III.	True	1.08 (± 0.71)
			IV.	True	1.10 (± 0.75)
			V.	True	1.15 (± 0.54)
		M16.	VI.	True	2.03 (± 0.75)
		M17.	VI.	True	1.71 (± 0.68)
		M18.	VI.	True	1.57 (± 0.59)
		M19.	VI.	True	1.56 (± 0.72)
		M20.	VI.	True	1.66 (± 0.74)

Table 4-13: EER results for MCYT75 (CoLL trained with Sign).

Test Signature dataset		CNN (trained with text)	CoLL (trained with sign)	WD classifiers with $\mathcal{N}_{REF}$ = 10	
db name	**canvas size**	**#Text set models**	**#Sign CEDAR set**	**sd**	**EER**
MCYT75	600 × 850	M1.	I.	True	1.43 (± 1.30)
			II.	True	1.46 (± 1.30)
			III.	True	1.39 (± 1.40)
			IV.	True	1.63 (± 1.20)
			V.	True	1.62 (± 1.40)
		M6.	I.	True	1.39 (± 1.20)
			II.	True	1.40 (± 1.20)
			III.	True	1.26 (± 1.10)
			IV.	True	1.38 (± 1.20)
			V.	True	1.48 (± 1.40)
		M12.	I.	True	1.53 (± 1.10)
			II.	True	1.88 (± 1.30)
			III.	True	1.97 (± 1.30)
			IV.	True	1.89 (± 1.30)
			V.	True	2.07 (± 1.30)
		M16.	VI.	True	2.18 (± 1.40)
		M17.	VI.	True	2.13 (± 1.60)
		M18.	VI.	True	1.94 (± 1.50)
		M19.	VI.	True	1.64 (± 1.40)
		M20.	VI.	True	1.62 (± 1.30)

Table 4-14: EER results for GPDS300GRAY (CoLL with Sign).

Test Signature dataset		CNN (trained with text)	CoLL (trained with sign)		WD classifiers with $\mathcal{N}_{REF}$ = *12*	
db name	**canvas size**	**#Text set**	**#Sign set**	**#Sign db**	**sd**	**EER**
GPDS300GRAY	952 × 1360	M16.	VI.	CEDAR	True	2.11 (± 0.79)
		M17.	VI.		True	2.20 (± 0.75)
		M18.	VI.		True	2.19 (± 0.84)
		M19.	VI.		True	2.23 (± 0.75)
		M20.	VI.		True	2.22 (± 0.74)
		M16.	VI.	MCYT75	True	1.98 (± 0.81)
		M17.	VI.		True	2.26 (± 0.75)
		M18.	VI.		True	2.04 (± 0.86)
		M19.	VI.		True	2.16 (± 0.75)
		M20.	VI.		True	2.12 (± 0.76)

Given that the addition of CoLL in the framework exhibits superior performance, even if is trained only with text images (for instance Table 4-11), the utilization of external signature images is advantageous. Therefore, the use of signatures for learning the CoLL leads to mostly superior (or at least comparable) results against all the previous experiments. Only in the case of CEDAR dataset where the signatures of MCYT75 were utilized for the training of CoLL module, the obtained EER values were a little bit worse. However, the deterioration is still less than 0.1% compared to the results of Table 4-9 and thus, cannot be considered significant. Thus, the combination of a CNN that learns features from a large amount of -readily available- text images along with a CoLL that learns the feature mapping through a limited number of signature images results in an efficient feature learning scheme for the OffSV task. In addition, another observation can be made about the normalization ("sd" parameter) of the final extracted features. When the CNN features are used to train the SVMs, there is no need for any normalization since the CNN has a batch normalization layer before its output. On the contrary, the normalization to zero mean and unit variance is beneficial when the CoLL module is used to produce the final features because the feature mapping has not provided normalization controls.

4.6.5 Comparison with SigNet trained with Signature images

In this section, we perform a fair comparison of the proposed feature extraction process with CoLL, to the original SigNet feature extractor proposed by Hafemann et al. in [102]. This SigNet model utilized only genuine signatures and no skilled forgeries during its training, similar to our scheme. The two compared feature extraction methods are applied to the same signature images -after applied the same geometrical normalization steps- and their output features are processed by the same classifiers. Thus, the comparison focuses only on the feature extraction stage and the quality of the generated features. The original SigNet was trained with the genuine signature images of 531 writers from GPDS-960 corpus and the trained model was downloaded from the Official repository[2].

The error bar diagrams of Figure 4-12 represent the EER values of all the proposed CNN-CoLL variations (based on the used training sets) for the three datasets, along with the corresponding EER and error margins derived using the SigNet as feature extractor. Similar to all previous results, the experiments are repeated ten times by randomly selecting the reference signatures, as is the standard practice in the OffSV literature. Additionally, Table 4-15 contains the results of our proposed method as well as the EER values in the case of our implementation with the downloaded SigNet model. This Table provides the direct comparison with SigNet and summarizes the multitude of previous experimental results. The various tested models are divided into single and multi-canvas preprocessed text and signatures, based on the used training set. For the models that trained with single-canvas images, the table is organized such that for each model (identified by the set used for its training) the top row includes signature sets with the same canvas size with the selected CNN model, the middle row incudes the signature set that provide the best performance using the sign-trained CoLL, and the bottom row includes the set with the best result for the text-trained CoLL.

As it is clear from Figure 4-12, the error margins of the reported average EERs between the proposed OffSV systems and the original SigNet CNN in all three signature datasets, i.e. CEDAR, MCYT75, and GPDS300GRAY, are highly overlapping. In order to strengthen the validity of our finding we perform a statistical analysis [248] of the results across the different experimental setting and dataset permutations. Once again, pairwise statistical comparisons between the original SigNet and every investigated setting for training a CNN-CoLL model are implemented using the Friedman's test and ANOVA (Analysis of Variance) for the ten repetitions of classifiers (with the same permutations of reference and test signatures). For most tested settings the p-values have large values (> 0.1), indicating that the models produced via the proposed technique are able to produce results which are statistically equivalent to those of Signet, even if they are trained with limited signature data. Especially important is the fact that for settings that utilize random or multiple canvas sizes (five rightmost settings in all plots of Figure 4-12), the p-values for all three datasets range between 0.2 and 0.97 for ANOVA and 0.11 to 1.0 for Friedman tests, signifying that these approaches are a safe option for replicating the performance of Signet.

[2] https://github.com/luizgh/sigver/tree/master/sigver/featurelearning/models

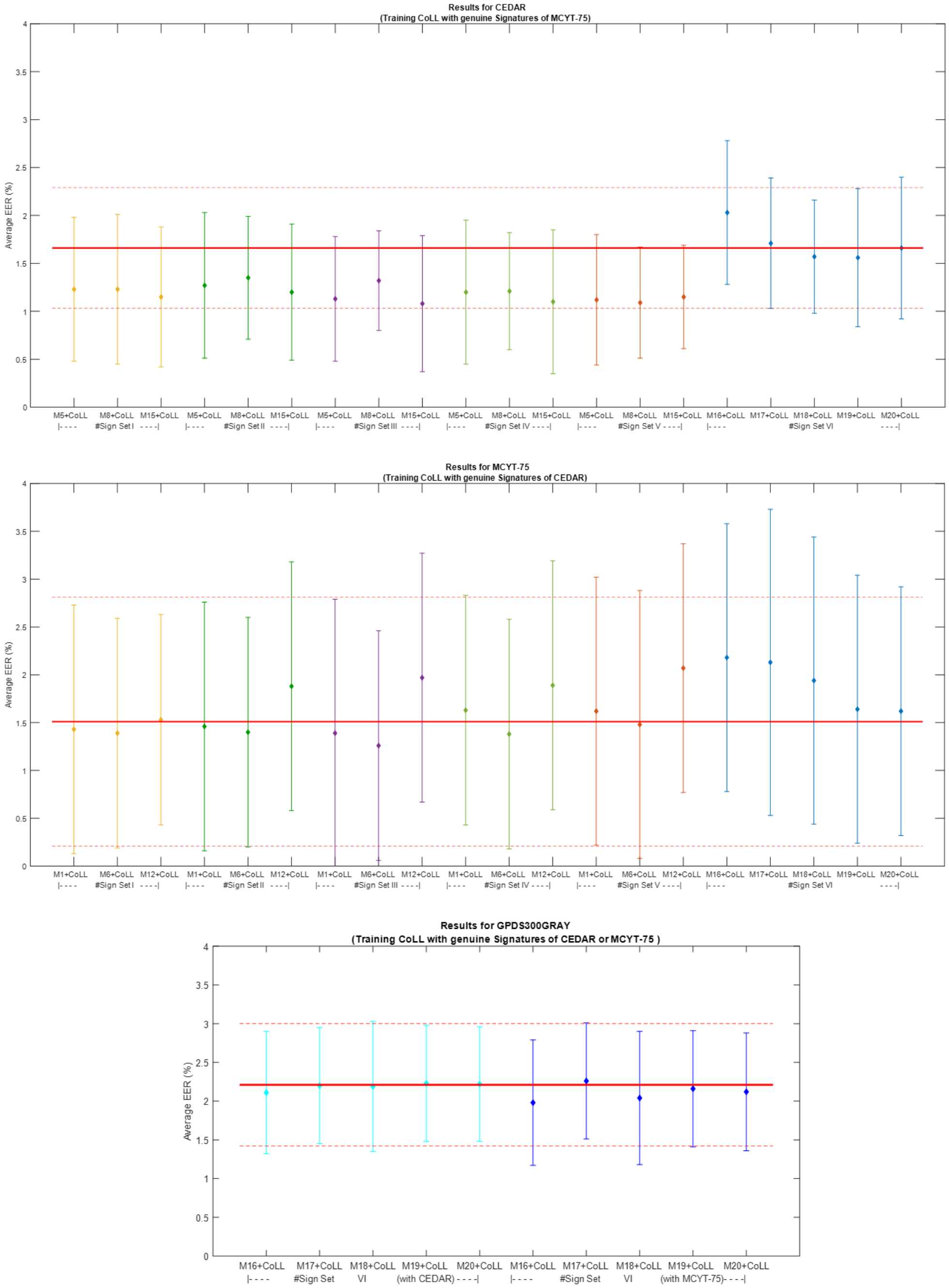

Figure 4-12: Error bar diagrams of EER (%) for the CEDAR, MCYT75, and GPDS300GRAY datasets using the different CNN-CoLL models from Table 4-12, Table 4-13, and Table 4-14, and comparison with the results of original SigNet model. The red lines represent the results from our implementation of original SigNet feature extractor proposed by Hafemann et al. in [102] with the solid red line indicating the average EER and the dashed red lines the respective error margins.

In some of the other investigated settings, the observed variations in the average EER were found to be statistically significant. For example, in some extreme cases (where the compared average EERs seems to differ), like M15–CoLLIII and M8–CoLLV in the CEDAR dataset the corresponding models achieved better performance than original SigNet with p-values of 6E-4 and 4E-4 (ANOVA) respectively. Similarly, models for M6–CoLLIII and M12–CoLLV in the MCYT75 are slightly better or worse than original SigNet, with p-values (ANOVA) of 2E-2 and 2E-4 respectively. The p-values values of Friedman test are very similar to those of ANOVA at every tested setting. The results of Figure 4-12 however, are presented in the spirit of an ablation study on the effects of canvas size to the overall performance of the feature extraction CNN, and they do not Offer any particular insight to the problem of how to train an efficient feature extraction CNN with less signature data. They can rather be attributed to circumstantial conditions that may benefit the classifiers for a particular database, which cannot be easily translated in a real-life situations, especially when considering that the fluctuation of results (i.e. variation of EER) from different CNN-CoLL settings (due to the different preprocessing parameters for generating the training sets) are considerable smaller than the variation that arises from the writer's signature variability, based on the selected reference signatures (via the ten repetitions of the experiments).

On the other hand, the statistical analysis of the results suggests that by using the proposed CNN-CoLL technique it is feasible to train an effective feature extraction model, using less signature images by taking advantage of the metric learning via the Contrastive Loss Layer (CoLL) and the pre-training with properly processed handwritten text images. The original SigNet is trained with about 531·24=12744 signature images (GPDS-960) whilst the proposed feature extraction system can be trained with about 55·24=1320 (CEDAR) or 75·25=1125 (MCYT75) signature images, providing statistically equivalent results. Hence, the presented technique can use one order of magnitude fewer training signatures than the SigNet, delivering similar level of performance. Most importantly, achieving such performance using random canvas sizes and arbitrary cropping ratios (such as Text set 20 and Signature Sets VI) in all datasets highlights the robustness and versatility of the proposed approach. The utilization of using the most general setting for the selection of these parameters, combined with the effective use of CoLL, eliminates the requirement of selecting a specific training set for each dataset. This level of flexibility enables the method to be easily adapted to various datasets without the need for extensive customization. Consequently, the proposed approach offers a practical and efficient solution for OffSV, demonstrating promising potential for real-world applications.

Table 4-15: Overview of our results for CEDAR and MCYT75 with N_{REF}=10 as well as for GPDS300GRAY with N_{REF}=12.

Test dataset			Proposed method								
Signature		SigNet [3][102]	Training Canvas	initial CNN (trained with text)		CNN (finetuned with sign)		CoLL (trained with text)		CoLL (trained with sign)	
	Signature Preprocessing canvas size	EER (WD)	Canvas type	#Text Set	EER (WD)	#Sign Set	EER (WD)	#Text Set	EER (WD)	#Sign Set	EER (WD)
CEDAR	730 × 1042	1.66 (±0.63)	Single canvas	M5.	1.19 (±0.72)	V.	2.15 (±0.95)	5.	1.15 (± 0.63)	V.	1.12 (±0.68)
						IV.	2.42 (±1.00)	4.	1.19 (±0.66)	IV.	1.20 (±0.75)
						III.	2.39 (±0.85)	3.	0.99 (±0.74)	III.	1.13 (±0.65)
				M8.	1.22 (±0.72)	III.	1.58 (±0.74)	8.	1.12 (±0.84)	III.	1.32 (±0.52)
						V.	1.44 (±0.83)	10.	1.21 (±0.86)	V.	1.09 (±0.58)
						I.	2.20 (±0.90)	6.	1.17 (±0.84)	I.	1.23 (±0.78)
				M15.	1.13 (±0.70)	V.	2.20 (±0.83)	15.	1.13 (±0.59)	V.	1.15 (±0.54)
						III.	2.32 (±0.68)	13.	1.23 (±0.85)	III.	1.08 (±0.71)
						IV.	2.41 (±0.88)	14.	1.12 (±0.73)	IV.	1.10 (±0.75)
			Multi canvas	M18.	1.88 (±0.75)	18.	2.41 (±0.80)	18.	1.49 (±0.76)	18.	1.57 (±0.59)
				M19.	1.86 (±0.82)	19.	1.95 (±0.68)	19.	1.51 (±0.81)	19.	1.56 (±0.72)
				M20.	1.91 (±0.78)	20.	2.05 (±0.86)	20.	1.65 (±0.78)	20.	1.66 (±0.74)
MCYT75	600 × 850	1.51 (±1.30)	Single canvas	M1.	1.84 (±1.60)	I.	1.83 (±1.20)	1.	1.62 (±1.20)	I.	1.43 (±1.30)
						III.	1.91 (±1.50)	3.	1.47 (±1.30)	III.	1.39 (±1.40)
						V.	2.03 (±1.30)	5.	1.60 (±1.30)	V.	1.62 (±1.40)
				M6.	1.77 (±1.50)	I.	1.65 (±1.30)	8.	1.54 (±1.30)	I.	1.39 (±1.20)
						III.	1.94 (±1.40)	10.	1.47 (±1.50)	III.	1.26 (±1.10)
						IV.	2.12 (±1.30)	11.	1.48 (±1.30)	IV.	1.38 (±1.20)

Test dataset		SigNet [3][102]	Proposed method								
Signature			Training Canvas	initial CNN (trained with text)		CNN (finetuned with sign)		CoLL (trained with text)		CoLL (trained with sign)	
	Signature Preprocessing canvas size	EER (WD)	Canvas type	#Text Set	EER (WD)	#Sign Set	EER (WD)	#Text Set	EER (WD)	#Sign Set	EER (WD)
				M12.	2.29 (±1.30)	II.	1.80 (±1.40)	12.	1.86 (±1.30)	II.	1.88 (±1.30)
						I.	1.52 (±1.30)	11.	2.05 (±1.30)	I.	1.53 (±1.10)
						III.	1.97 (±1.50)	13.	1.82 (±1.50)	III.	1.97 (±1.30)
			Multi canvas	M18.	2.39 (±1.80)	18.	2.19 (±1.50)	18.	2.06 (±1.50)	18.	1.94 (±1.50)
				M19.	2.15 (±1.70)	19.	2.08 (±1.50)	19.	1.54 (±1.70)	19.	1.64 (±1.40)
				M20.	1.86 (±1.40)	20.	1.77 (±1.60)	20.	1.65 (±1.60)	20.	1.62 (±1.30)
GPDS300GRAY	952 × 1360	2.21 (±0.79)	Multi canvas	M16.	2.44 (±0.72)	16.	3.01 (±0.90)	16.	2.09 (±0.82)	16.	1.98 (±0.81)
				M17.	2.61 (±0.76)	17.	3.07 (±0.84)	17.	2.23 (±0.64)	17.	2.26 (±0.75)
				M18.	2.48 (±0.84)	18.	2.69 (±0.80)	18.	2.17 (±0.88)	18.	2.04 (±0.86)
				M19.	2.51 (±0.77)	19.	3.18 (±0.83)	19.	2.25 (±0.71)	19.	2.16 (±0.75)
				M20.	2.36 (±0.81)	20.	2.86 (±0.96)	20.	2.30 (±0.76)	20.	2.12 (±0.76)

4.6.6 Summary of Performance in WD OffSV field

Table 4-16 provides an overview of the OffSV field, summarizing the most important results from various methods and evaluation protocols reported in the Writer-Dependent (WD) OffSV literature during the last 15 years, using the three most popular datasets CEDAR, MCYT75, and GPDS. It is obvious that a fair comparison between all methods is a strenuous task due to the many different protocols and technicalities that impact the performance. (e.g., number of reference signatures, use of skilled forgery training samples etc.). Therefore, the particular table serves the purpose of providing a general outlook of the WD OffSV research, emphasizing in the recent advances. In this context, a quick look to state-of-the-art systems can be useful. At the work of (Maruyama et al., 2021), the WD SVM classifier is populated with more points in the training stage using feature replicas extracted from a signature duplication process and thus, the improvement is stemming from these classifier scheme and is not attributed to a better feature extraction mechanism. Also, a variant of SigNet [102], named SigNet-SPP [103], utilizes spatial pyramid pooling for variable input image sizes, while another variant of SigNet, the SigNet-F [102], uses forged signatures along with the genuine signatures of GPDS-960 corpus for training. However, none of SigNet's variants is consistently better in all three datasets. It is worth noting that the difference in EER values between our implementation of SigNet and the published values in the work of [102] is associated with the different way of utilizing the WD classifiers. In our experiments the hyperparameters of RBF SVM are optimized through a cross-validation procedure for every writer, while at the work of [102] the same hyperparameters were used for all the writers. Finally, research conducted by Zois et al. [70], [91] utilizing the spatial pyramid pooling of sparse features and visibility motif features achieved a good tradeOff between learning-based and hand-crafted components in the model that fits OffSV task. Ultimately, we argue that the proposed approach proves the feasibility of achieving a low verification error, which is at least comparable to the state-of-the-art methods in all three datasets, despite following a fully learning-based approach with limited training samples. Therefore, it can provide a pathway to develop more complex deep learning based OffSV systems with the current data availability.

Table 4-16: Summary of state-of-the-art OffSV Systems in terms of EER, for the CEDAR, MCYT75, and GPDS300GRAY datasets.

Signature		**OffSV approach**		**WD classifiers**
db name	N_{REF}	**Reference**	**Method**	**EER**
CEDAR	12	[63]	Chain Code	7.84
	16	[250]	Chord moments	6.02
	16	[61]	Gradient LBP+LRF	3.54
	5	[69]	Archetypes	2.07
	12	[102]	SigNet-F	4.63
	12	[102]	SigNet	4.76
	10	[103]	SigNet-SPP	3.60
	5	[95]	Deep SC	2.82
	16	[212]	VLAD with KAZE	1.00
	10	[91]	SR –KSVD/OMP	0.79
	16 (10)	[251]	Hybrid Texture	1.64 (6.66)
	10	[77]	CNN-Triplet and Graph edit distance	5.91
	12	[100]	HOCCNN	4.94
	10	[70]	Visibility Motif profiles	0.51
	3	[227]	SigNet-F and classifier with replicas	0.82

Signature		OffSV approach		WD classifiers
db name	N_{REF}	Reference	Method	EER
	3	[102]*	SigNet	2.83
	5	[102]*	SigNet	2.14
	10	[102]*	SigNet	1.66
	3	**Proposed**	CNN-CoLL	2.50
	5	**Proposed**	CNN-CoLL	2.03
	10	**Proposed**	CNN-CoLL	1.66
MCYT75	10	[68]	Contours	6.44
	5	[252]	Ring Peripheral	15.02
	10	[62]	LBP	7.08
	10	[64]	Radon Transform	9.87
	10	[233]	HOG + DMML	9.86
	10	[253]	HOT	10.60
	8	[188]	Duplicator	9.12
	5	[217]	Archetypes	3.97
	10	[102]	SigNet-F	3.00
	10	[102]	SigNet	2.87
	10	[103]	SigNet-SPP	3.64
	10	[254]	FV with KAZE	5.47
	10	[229]	ResNet trained with text	3.98
	10	[101]	MLSE	2.93
	10	[91]	SR – KSVD/OMP	1.37
	14 (10)	[251]	Hybrid Texture	6.10 (9.26)
	10	[77]	CNN-Triplet and Graph edit distance	3.91
	12	[100]	HOCCNN	5.46
	10	[70]	Visibility Motif profiles	1.54
	3	[227]	SigNet-F and classifier with replicas	0.01
	3	[102]*	SigNet	3.28
	5	[102]*	SigNet	2.52
	10	[102]*	SigNet	1.51
	3	**Proposed**	CNN-CoLL	3.33
	5	**Proposed**	CNN-CoLL	2.61
	10	**Proposed**	CNN-CoLL	1.62
GPDS160GRAY	16	[255]	Geometric	9.64
	12	[256]	MDF, Energy, Maxima	17.25
	12	[215]	HOG-LBP	15.41
	10	[209]	Pseudo-dynamic	7.66

Signature		**OffSV approach**		**WD classifiers**
db name	N_{REF}	**Reference**	**Method**	**EER**
	12	[257]	HOG-LBP-SIFT	6.97
	12	[71]	LBP	11.74
	12	[228]	2-channel SigNet-F	2.08 (0.88)
	12	[219]	RBP	0.57
GPDS300GRAY	13	[258]	Circular Grid	4.21
	12	[259]	Cosine similarity	7.20
	6	[260]	Optical flow	4.60
	12	[60]	Poset-oriented grid	3.24
	14	[222]	DCGANs	12.57
	10	[233]	LBP + DMML	20.94
	10	[253]	HOT	9.30
	8	[188]	Duplicator	14.58
	12	[102]	SigNet-F	1.69
	12	[102]	SigNet	3.15
	12	[103]	SigNet-SPP-F	0.41
	10	[261]	HOT + AIRS	11.35
	12	[91]	SR – KSVD/OMP	0.70
	12	[251]	Hybrid Texture	8.03
	3	[227]	SigNet-F and classifier with replicas	0.20
	3	[102]*	SigNet	3.44
	5	[102]*	SigNet	2.84
	12	[102]*	SigNet	2.21
	3	**Proposed**	CNN-CoLL	3.69
	5	**Proposed**	CNN-CoLL	2.91
	12	**Proposed**	CNN-CoLL	2.12

*The trained SigNet model from [102] is used along with our SVM configuration.

4.7 Conclusions

The aim of this work is to present a methodology of efficient feature learning for the Offline Signature Verification task using Convolutional Neural Networks, designed to overcome the limitations in availability of signature images following the withdrawal of large datasets from the public domain due to privacy legislation. The proposed CNN-CoLL scheme is taking advantage of handwriting data in a more general sense. The handwritten style arises both in handwritten texts and signatures. The relevancy of writing and signing let us pre-train the CNN in an exterior task of identifying the author of an input image that contains text and then, use the trained CNN as a good initial baseline model for feature extraction. For validating our claim, we followed the most established evaluation methods in the related literature, ensuring that the results are directly comparable to the most popular deep-learning approach for OffSV task - the SigNet CNN architecture. We incorporated a series of simple processing steps for the raw text data, designed to simulate the signature images without the incorporation of sophisticated OCR or similar techniques, thus enabling a fast and efficient text manipulation, well-suited to large-scale data processing. This choice was made to allow harnessing information from the large abundance of available handwritten text data to develop better learning-based OffSV systems, and ultimately encourage further research towards the direction of incorporating modern deep-learning techniques in OffSV even though a large signature dataset is currently unavailable.

The addition of a feature mapping stage aiming to reorganize the feature space, based on metric learning with pairwise contrastive loss, boosted the performance of the presented OffSV system. The WI training of CNN-CoLL framework provides a feature extraction mechanism which is efficient for any query signature image of unseen writers (from other datasets or tasks). The CNN is trained solely with text images while the training of CoLL was evaluated with either text or genuine signatures (from irrelevant writers) as training examples.

A point of significant practical importance is that the presented scheme does not require skilled forgeries at any stage of the training pipeline. In this spirit, the WD SVM classifiers are also trained with samples of genuine against random forgeries but evaluated with the remaining genuine signatures as well as the skilled forgery signatures for each writer. Results indicate that the proposed CNN-CoLL scheme manages to successfully learn informative features with about one thousand signature images, while other CNN-based methods utilize over an order of magnitude more signature images in order to achieve similar performance in the OffSV task. The efficiency of the system is demonstrated with experiments in the most popular signature datasets, achieving better average EER than several state-of-the-art OffSV systems and statistically equivalent results to the original SigNet model, despite the latter being trained on the GPDS dataset with one order of magnitude more signature images compared to the presented scheme. Comparisons were focused to SigNet since this is the only GPDS-trained model with only genuine signatures and reproducible results, allowing a fair comparison using the most popular protocol in WD-OffSV literature.

Evaluation results also indicated that the variability of the EER due to the random selection of reference sets across iterations, is greater than the variability induced by the selection of the specific combinations of canvas sizes for the normalization of text and signatures during the training of CNN and CoLL Thus, although the preprocessing is of crucial importance, the comparable results when different models are utilized show that the different preprocessing parameters have lower effect than the writer’s natural variability as expressed in its reference signatures. Through a meticulous experimental study on the effects of cropping and canvas dimensions of the external text and signature data, we demonstrated that even with random choice of parameters for generating the training sets (i.e., Text Set 20 and Signature Set VI) the proposed pipeline can reliably train a model that learns efficient features across all tested datasets. Therefore, as long as those parameters lie inside a reasonable margin as the ones tested in this study, it is needless to seek for specific qualities in the external data which are tuned to the target domain. This finding is of particular practical importance, since it enables to train the feature extraction stage without any knowledge of the reference dataset, thus avoiding the need of retraining the CNNs as the reference set grows through the lifetime of an OffSV system. This last observation supports our core idea that transferring knowledge from the handwriting text data to the signature problem, even with a simple and fast preprocessing procedure that involves random selection of cropping strategy and canvas sizes for the generation of the training images based on text and signature data, can deliver state-of-the-art performance even compared to methods trained with 10X the amount of currently available data.

Chapter 5

Deep learning with knowledge distillation

5.1 Introduction

An extrinsic limitation in OffSV problem is introduced from the absence of large offline datasets [195]. Until recently, the GPDS-960 corpus offline database [262], with more than half a thousand writers having 24 genuine along with 30 forgeries signatures per writer, allowed the training of deep models into the similar task of writer identification [44]. Even though these CNN models are not specialized to the task of signature verification, the large size of GPDS-960 dataset enabled CNNs to be good universal functions for producing image-level feature descriptors for signature images, surpassing the expressiveness of hand-crafted features [46]. Unfortunately, this dataset, is no longer available due to the General Data Protection Regulation (EU) 2016/679 ("GDPR"), thus hindering the efforts of research community to investigate new models and design elaborate methods that require more training data.

Motivated by the data-intensive nature of CNNs' training, many OffSV systems pursue designing methodologies to address the lack of adequate signature training data. These approaches follow two main directions, the generation of synthetic signature images using geometrical transformations [73], [191] or generative learning models [75], [192]–[194] and the utilization of images from a relative domain such as the handwritten text documents [263], [264]. For completeness, there are also developed feature space augmentation methods that artificially populate samples for improving the classifier' performance, yet they rely on feature vector representations and do not create signature images for training [74], [76], [265]. Finally, considering the fully synthetic nature of generated signatures and the contingent unreal identities, the signature duplications would be prone to diverge from the realistic intra-subject variability criterion and thus, their use as the training images of an end-to-end deep learning model could be problematic, requiring special manipulations for their beneficial usage output [83], [84].

Despite the capabilities of the above approaches to cope with signature verification problems with small sample size, these methods ignored the knowledge of older benchmark models in the OffSV task. In situations where an effective CNN model is available, a popular approach is to transfer knowledge from this (expert) model to facilitate learning of another (new) model. This forms the case of Knowledge Distillation (KD), where the knowledge is transferred between the models that assume the role of teacher (expert) and student (new) [45]. When the teacher is pre-trained and fixed during training, it is called offline KD. Additionally, the condition when an effective teacher model exists but there is no access to its training data constitutes the data-free KD, where the distillation process uses only external or artificial data to perform the knowledge transfer from the teacher to the student model [24]. Finally, another branch which is relevant to this work is Feature-based KD (FKD), which involves distilling knowledge from the intermediate layers of the teacher model in order for the student model to learn feature representations that are a good approximation of the teacher's intermediate representations.

To the best of our knowledge, we consider this work to be the first one that introduces the data-free KD approach into the OffSV domain. Here we propose a novel KD method to transfer the knowledge from a teacher CNN into a new CNN student model with different architecture. This allows the new model to leverage the knowledge learned by the teacher model, even though the original training data are not available anymore. Furthermore, the new model is able to achieve improved performance on the OffSV task compared to the teacher. The KD scheme consists of 1) the teacher CNN supervising the training process, 2) the training data used to transfer the knowledge, and 3) the KD method that defines knowledge features, distillation loss, strategy, and connections. The ultimate goal is to express the learned information inasmuch as it is helpful for building up a well-performing student CNN. Therefore, to address OffSV using offline data-free Feature-based KD, the above components are realized as follows:

1) An appropriate teacher is one of the benchmark CNN models in the field, such as SigNet [102] which is trained with the genuine signature images from 531 writers using GPDS-960 corpus and the trained model is publicly available[3]. In this occasion, the teacher model can provide valuable feature representations for any input image, but not a meaningful classification response, since the training classes are person IDs which are irrelevant outside the specific identification task.
2) The data which act as information carriers for the distillation, can be either synthetic or external. In our work, we opt to utilize images of handwritten text because they possess a similar structure to signatures (thin pen strokes on a piece of paper) and most

[3] https://github. com/luizgh/sigver/tree/master/sigver/featurelearning/models

importantly, there is an abundance of data available from public sources. The option of using synthetically generated signatures was dismissed in light of evidence indicating that the currently available synthetic signature datasets can deteriorate the effectiveness of OffSV systems if used to train the feature extractor model [83], [84], [193].

3) The Feature-based Knowledge Distillation (FKD) is applied for guiding the activations at the intermediate layers of teacher and student models. Here, we utilize computationally efficient loss functions aiming to transfer the geometry of activations from intermediate layers of the teacher CNN to the activations of the student CNN model at matching spatial resolutions. The employed loss functions emerge from manifold-manifold distance functions, formulating the problem of FKD as a problem of learning similar manifolds of local activations in corresponding layers of teacher and student models. Furthermore, the training of the student model incorporates KD attained by an additional regularization loss that is based on the global feature, generated at the penultimate layers of the teacher and student models respectively. Under this direction an efficient loss function is designed to fit with the KD scope, inspired from the Self-Supervised Learning method of Barlow Twins [266]. Ultimately, the proposed KD method utilizes both geometric FKD and global FKD, thus integrating local information via manifold-to-manifold comparison as well as global information via metrics that range from typical temperature-scaled cross entropy to KD-oriented cross-correlation losses.
4) The requirements for the student CNN model architecture, utilized in the FKD scheme are: (i) matching of intermediate activations for at least some of spatial resolutions and (ii) for the global feature to share equal dimensions with that of the teacher model. The popular ResNet-18 CNN was selected as the student architecture, given its efficiency and modern topology [4].

The training of a feature extraction model for OffSV is a learning task different than the main verification task, since the identity and data of the users involved in the operational phase are not always available during the model's training. Following feature extraction, the decision stage analyses the feature representation of a signature image and decides upon its validity. Since the goal of this work is to demonstrate the value of the proposed FKD in designing an efficient OffSV feature extractor, at the final decision stage we follow the most straightforward WD approach, using WD Support Vector Machine classifiers to evaluate our method at the operational phase. Results indicate that our system achieves top-tier performance on three popular Latin offline signature datasets without requiring any signature images during Student-Teacher training. The verification error is in par with state-of-the-art models trained with thousands of signature images, obtained by only exploiting knowledge via the proposed FKD scheme. Also, the training of the OffSV system does not require any skilled forgery signatures because the final decision stage with the WD classifiers uses only genuine signatures and particularly, a few signatures of the writer along with some signatures of other writers, also known as random forgeries.

The contributions of the proposed work for OffSV could be summarized as follows:

- We demonstrate that FKD enables the efficient training of any new architectures that inherit the prior knowledge of benchmark models whose training data are unavailable, using external data of similar nature.
- The knowledge transfer between expert CNN model and new CNN model is accomplished without the use of signature images, employing only handwritten text data processed using a specialized yet simple pre-processing scheme.
- We propose a method for KD that combines information from both local features' geometry and global feature distribution.
- A novel global feature-level loss function is designed in the basis of H. Barlow's redundancy-reduction principle, enhancing the similarity between the compared features while minimizing the redundancy between the remaining components of these vectors, accommodating the utilization of two different architecture in the S-T KD scheme.

The rest of the paper is organized as follows. Section 2 presents an overview of the recent deep learning methods related to OffSV problem. Section 3 describes thoroughly the proposed FKD method through the Student-Teacher architecture. Section 4 presents the experimental results investigating many different KD schemes and finally Section 5 provides discussion and conclusions.

5.2 Related Work

Training a feature extraction model for Offline Signature Verification (OffSV) is typically a separate learning task from the main verification task, as the identity and data of users involved in the operational phase are not available during the model's training phase. This provides flexibility in designing an efficient and practical OffSV system, resulting in a multitude of developed methods. Deep learning schemes have demonstrated effective performance in OffSV, mainly as feature extractors [46]. The main points of any deep learning based OffSV system could be summarized on the CNN architecture, the design strategy, and the multi-task learning mechanism. To achieve feature learning, a variety of CNN architectures are employed using either a WI or WD approach. The choice of architecture depends on both the user's requirements and the available signature data. Several OffSV datasets are available, as detailed in a recent survey by Diaz et al. [195]. In addition, a plethora of strategies have been developed to effectively capture the underlying signature information. The Siamese concept has a prominent position among these strategies since it is well-suited to the verification problem, having two inputs to compare two patterns and one output whose state value corresponds to the similarity between the two patterns [267]. Finally, the multi-task learning enjoys high popularity in the OffSV field due to its easy and effective implementation. The multi-task approach begins with a first task that acts as primary learning, while additional learning task(s) fine-tuned specific characteristics on the feature representations of signatures [44]. A taxonomy of deep OffSV methods according to the involved CNN architecture is attempted on Table 5-1, including information about the respective strategies and presence of multi-task stages.

The SigNet architecture, based on the AlexNet CNN topology, is dominant in the OffSV field [7]. It was initially designed for writer identification, with the aim of distinguishing between signatures of different writers using only genuine signatures [102]. However, the architecture has since undergone various modifications, resulting in several versions of SigNet that differ mainly in the dimensions of the extracted features, such as the so-called thin SigNet [102], R-SigNet [268], and SigNet-SPP [103]. Also, different multi-loss settings have been employed as objective functions during training of SigNet. In these settings, the primary loss is responsible for associating signatures with their respective users, while additional loss terms are used either for detecting forgeries, resulting in the SigNet-F version [102], [268], or combined with other metric learning functions to form the Multi-Loss Snapshot Ensemble (MLSE) method [101]. To leverage the benefits of the SigNet-F feature extractor, post feature management methods are applied, such as the Dichotomy Transformation in the dissimilarity space [82] and the feature augmentation techniques to enhance the performance of the classifiers [74]. The Siamese scheme is also formulated using the SigNet's architecture in its identical subnetworks [223]. Building upon this, SigNet is utilized in multi-stage frameworks, either when it is initially trained to distinguish between signatures of different writers and subsequently re-trained using the contrastive loss function [84] or when it is initially trained with handwritten text data and then is used as the baseline model for training an additional contrastive loss layer at the top of the net [264]. The contrastive loss is the most common similarity ranking function in Siamese schemes and its objective is to learn such an embedding space in which similar sample pairs are pulled together while dissimilar ones are pushed apart [28]. An extension of the Siamese concept is the triplet loss, which is composed of: an anchor, a positive sample from the same class, and a negative sample from a different class. In this case, the goal is to minimize the distance between the anchor and positive sample while maximizing the distance between the anchor and negative sample in the embedding space [29], [269]. Beyond that, the dual triplets (or quadruplet) can also be used, which include two negative samples in addition to the anchor and positive samples [270]. In the case of OffSV using SigNet model, the first negative sample is a random forgery and the second negative sample is a skilled forgery signature [271].

Many CNN architectures for OffSV systems are utilized as identical streams of joint embedding (i.e., Siamese) frameworks that rely on contrastive representation learning. One such architecture is the DenseNet [272] including squeeze-and-excitation blocks (SE) [273] with [274] and without [275] spatial pyramid pooling (SPP) for the calculation of the global feature. The SPP layer is also compiled with the custom CNN architecture, named Position-Dependent Siamese Network (PDSN), to model the local similarity between signatures [276], while a custom CNN equipped with an inception layer [277], named Siamese Convolutional Inception Neural Network

(SCINN), is utilized to capture signature details [278]. In the work of Parcham et al. [279], the Capsule Neural Network (CapsNet) produces the final feature embeddings of the overall Siamese scheme and the resulting composite backbone architecture with the hybrid CNN-CapsNet models is named CBCapsNet. In this study, a plethora of CNN architectures are evaluated along with the CapsNet and thus, networks from the families of VGG [159], In/Xception [277], [280], ResNet [4], [281], [282], DenseNet [272], MobileNet [283], and NASNet [284] were investigated under the proposed Siamese set up. The triplet-based learning is also performed in the OffSV problem using the VGG-16 model [232] as well as the ResNet-18 and DenseNet-121 models [77]. In more complex configurations with custom CNN architectures for OffSV, the use of signature pairs can take many forms. For example, the work of Lu et al. [285] proposed the use of a smooth double-margin loss as an inventive extension of the contrastive loss while the work of Zhu et al. [79] proposed a CNN equipped with fractional max pooling function as long as the contrastive and triplet losses are formulated with the novel point-to-set (P2S) similarity metric. The Deep Multitask Metric Learning (DMML), created by Soleimani et al. [233] as a multi-task learning version of Discriminative Deep Metric Learning (DDML), has a shared layer for all the writers that is followed by separated layers which belong to each writer independently and the overall topology is optimized using the relevant signature pairs. In addition, both the Inverse Discriminative Network (IDN) and the Multiple Siamese Net (MSN) utilize the original image (i.e., with white background and gray signature strokes) and the inverse version (i.e., with black background and gray signature strokes) of each signature of the input pair and through pairwise connection of its four different streams providing three [88] or four [286] verification scores that combined for the final decision. Finally, in an altered direction, a pair of grayscale signatures is fed into a custom CNN architecture as a two-channel input image to incorporate the similarity between the two signatures implicitly in the encoding process [228].

Before their final use as feature extractors for signature images, popular CNN architectures are often trained following a different strategy, specifically in writer identification tasks, rather than the Siamese concept. The multi-task approach is commonly adopted in many cases, either as a multi-stage process where pretraining serves as coarse initialization for the network before the main training process specific to the method, or as a multi-loss implementation where multiple loss functions are optimized together to balance multiple objectives. For the OffSV problem, the ResNet-8 is pretrained with auxiliary Persian handwritten text images in the writer identification task and next either is used as a fixed feature extractor or is fine-tuned with signature images of the target domain [263]. Following a similar rationale, the CNNs are initially pretrained on the general imagenet dataset with millions of training images and then fine-tuning is performed on a single signature dataset under the writer identification problem to harness the effectiveness of the extracted vectors from the models such as VGG-16 [287], ResNet-50 [287], [288] and GoogLeNet [289]. Likewise, the pretrained (on imagenet dataset) models of VGG-16, VGG-19, ResNet-50, and DenseNet-121 feed with feature representations the CapsuleNet that operates as the final verification classifier and the whole system is trained in an end-to-end manner using signatures [193]. Also, the custom topology combining two streams of convolutional processes in the work of Zheng et al. [72] is pretrained on the signature writer identification task and subsequently the CNN is trained to capture micro deformations. Unlike the previous multi-stage approaches, the Shariatmadari et al. [100] trained their deep architecture utilizing a multi-loss approach combining two losses emerged from three CNN streams in different sizes of convolutional layers, while the proposed approach is based on Hierarchical One-Class CNN (HOCCNN) that trained only with genuine signatures from different feature levels. Furthermore, the CNN structure named Large-Scale Signature Network (LS2Net) with the class-center based classifier addresses the writer identification problem using the class centers -by averaging the extracted features of each class- and the 1-Nearest Neighbor classifier [290].

The Recurrent Neural Networks (RNNs) are a type of neural network that are well-suited for processing sequential data. In the context of signature verification, RNNs can be used to analyze signature images by dividing them into segments and treating each segment as a separate time step in a sequence. There are several ways that RNNs can be applied to this task. One approach is to simply design geometrical windows on the pixel domain, as described in [291]. Another approach is to use Local Binary Patterns (LBP) coded image windows, as described in [292]. These windows can be processed by a Bidirectional Long Short-Term Memory (BiLSTM) network, which is a type of RNN that is able to analyze the input data in both forward and backward directions. In a simpler implementation, a CNN can be used for feature extraction, with the output of the CNN being fed into a BiLSTM to classify the signature as genuine or forged [81]. The Static-Dynamic Interaction Network (SDINet) is another method for incorporating sequential information into static signature images by assuming pseudo dynamic processes in the static image [293]. It does this by uniformly dividing the feature maps of the signature into rows and columns, with each row or column representing a dynamic unit in the signing process. Thus, the static feature maps are converted into sequences based on the part-by-part nature of the signing process.

In the field of signature verification, there have been alternative proposed approaches that deviate from the usual line of research that was described above. One such approach is to use an autoencoder to generate forgery signatures from the genuine ones, where the encoder model is utilized to extract features from signature images [294]. In the same vein, another approach is to use an Adversarial Variation Network (AVN), as proposed in the work of Li et al. [295]. The AVN exploits a variation consistency mechanism to train a discriminative model for signature authentication that is more robust than a typical Generative Adversarial Network (GAN). The AVN’s feature extractor and discriminator are equipped with a variator that slightly perturbs the colors or intensities of the signature images to produce variants that should not affect the verification decision. Additionally, adversarial examples, which are intentionally designed to mislead a classifier, can pose a challenge for OffSV systems, as they can cause misclassification [225], [296]. Finally, Graph Neural Networks (GNNs) have been applied to the problem of OffSV for the first time in the work of Roy et al. [297] and the transformer structure has been introduced as a feature extractor for signature images by Ren et al. [298]. Both of these approaches show promising results.

Recently, the Self-Supervised Learning (SSL) approach has been introduced for the OffSV domain. Two SSL approaches have been developed for this task. The first approach involves pretraining a ResNet-18 model by minimizing the cross-correlation matrix between compared features. The resulting model is then used as a fixed feature extractor [299]. The second approach involves pretraining an image reconstruction network with an encoder-decoder topology. The encoder (ResNet-18) is then finetuned using a dual triplet loss, and the resulting model is used as a feature extractor [300]. Differently from these SSL approaches where supervisory signals are obtained from the data itself, we propose a KD method for the training of the OffSV feature extractor where the process is supervised from a teacher model. Hence, we leverage prior knowledge by having the student model use an existing efficient CNN model for signature encoding. Additionally, we utilize handwritten text data to transfer the knowledge from the teacher to the student and not signatures, contrary to the aforementioned SSL works that rely on signature samples from the same datasets for achieving descent performance. Although both methods, SSL and FKD, utilize loss functions based on the cross-correlation matrix of global features, in the proposed scheme the FKD loss function is tailored to the KD concept instead of feature similarity.

Table 5-1: A taxonomy of recent deep learning-based OffSV systems.

CNN involved Architecture		**Strategy**	**Multi-task**	**Authors, Year**	**References**
SigNet (AlexNet)	SigNet, thin SigNet	Identification	-	Hafemann et al., 2017	[102]
	SigNet-F	2-term Loss	Multi-Loss	Hafemann et al., 2017	[102]
	SigNet	Siamese	-	Dey et al., 2017	[223]
	SigNet-SPP, fine-tuned	2-term Loss & fine-tuning	Multi-Loss	Hafemann et al., 2018	[103]
	SigNet (MLSE)	3-term Loss	Multi-Loss	Masoudnia et al., 2019	[101]
	SigNet-F	Dichotomy Transformation	Multi-stage	Souza et al., 2020	[82]
	SigNet-F	Feature Augmentation	Multi-stage	Maruyama et al., 2020	[74]
	R-Signet-F	2-term Loss	Multi-Loss	Avola et al., 2021	[268]
	SigCNN	Dual Triplets	-	Wan and Zou, 2021	[271]

CNN involved Architecture		Strategy	Multi-task	Authors, Year	References
	SigNet-CoLL	Contrastive Layer (CoLL)	Multi-stage	Tsourounis et al., 2022	[264]
	SigNet	Multi-task Contrastive Learning	Multi-stage	Viana et al., 2022 Viana et al., 2023	[83], [84]
ResNet	ResNet-8	Auxiliary data	Multi-stage	Mersa et al., 2019	[263]
	ResNet-18	Pretraining Identification + Triplets	Multi-stage	Maergner et al., 2019	[77]
	ResNet-50	Pretraining Imagenet + Active Learning	Multi-stage	Younesian et al., 2019	[288]
	ResNet-50	Pretraining Imagenet + Identification	Multi-stage	Engin et al., 2020	[287]
	ResNet-18	SWIS: Self-Supervised Pretraining + Contrastive	Multi-stage	Manna et al., 2022	[299]
	ResNet-18	SURDS: Self-Supervised Pretraining + Dual Triplets	Multi-stage	Chattopadhyay et al., 2022	[300]
VGG	VGG-16 (reduced)	Pretraining Identification + Triplets	Multi-stage	Rantzsch et al., 2016	[232]
	VGG-16	Pretraining Imagenet + Identification	Multi-stage	Engin et al., 2020	[287]
DenseNet	DenseNet-36	Multi-region + Siamese	-	Liu et al., 2018	[275]
	DenseNet-121 (MCS)	Pretraining Identification + Triplets	Multi-stage	Maergner et al., 2019	[77]
	Mutual Signature DenseNet-36 (MSDN)	Multi-region, SPP + Siamese	-	Liu et al., 2021	[274]
InceptionNet	Convolutional Inception NN (SCINN)	Signature Synthesis + Siamese	-	Ruiz et al., 2020	[278]
	GoogLeNet	Pretraining Imagenet + Identification	Multi-loss/ -stage	Jain et al., 2021	[289]
CapsuleNet	VGG-16/19, DenseNet-121, ResNet-50 + CapsNet	Signature Augmentation + Pretraining Imagenet + end-to-end Verification	Multi-stage	Yapici et al., 2021	[193]
	VGG-16/19, ResNet-50/101/152, In/Xception, InceptionResNet, MobileNet, NASNet + CBCapsNet	Siamese	-	Parcham et al., 2021	[279]
Custom CNN	Shared layers followed by separated layers (DMML)	DDML with User-specific layer + Pairs	Multi-stage	Soleimani et al., 2016	[233]
	2-channel CNN	2-channel input Pair	-	Yilmaz and Öztürk, 2018	[228]

CNN involved Architecture		Strategy	Multi-task	Authors, Year	References
	CNN + PSDN	Siamese	Multi-Loss	Lai and Jin, 2018	[276]
	4-stream CNN	Multi-Path + Pairs (IDN) / (MSN)	-	Wei et al., 2019 / Xiong et al., 2021	[88], [286]
	HOCCNN	Hierarchical one-class Learning	Multi-Loss	Shariatmadari et al., 2019	[100]
	LS2Net	1-Nearest Neighbor (1-NN) classification task by using the class-centers	-	Çalik et al., 2019	[290]
	CNN with fraction max pooling	Point-to-Set (P2S) Similarity	-	Zhu et al., 2020	[79]
	2-stream combined CNN	Pretraining Identification + micro-Deformations Learning	Multi-stage	Zheng et al., 2021	[72]
	cut-and-compare Net	Segmentation, Comparison + Pairs	-	Lu et al., 2021	[285]
	SDINet	conversion of static feature maps into sequences	-	Li et al., 2021	[293]
RNN	LSTM/ BiLSTM	Spatial segments + Identification	-	Ghosh et al., 2020	[291]
	Recurrent Binary Pattern – BiLSTM	LBP coded windows + Identification	-	Yilmaz and Öztürk, 2020	[292]
	CNN with BiLSTM	Hybrid CNN-BiLSTM verification	-	Longjam et al., 2023	[81]
Autoencoder	custom 6-layer CNN	Utilization of encoder model	-	Prajapati et al., 2021	[294]
AVN	VGG (inspired from)	Variation consistency mechanism	-	Li et al., 2021	[295]
GraphNN	GLCM-GNN	Node Classification	-	Roy et al., 2021	[297]
Transformer	two-channel and two-stream (2C2S) transformer	squeeze-and-excitation (SE) operation between two standard Swin Transformer blocks + Pairs	-	Ren et al., 2023	[298]
Adversarial attack	adversarial examples	adversarial characterization / adversarial perturbations	-	Hafemann et al., 2019 / Li et al., 2021	[225], [296]

5.3 Proposed Method

5.3.1 Harnessing Knowledge through Distillation

The efficiency of CNNs in the modern Deep Learning era is founded on large and annotated training datasets and thus, the amount and quality of both data and labels is mission-critical. The most popular approach for reducing the amount of labeled training data without affecting the performance too much, is by employing prior knowledge from a source domain with an abundance of training data on a similar task. Then, transfer knowledge is performed from the source task to enable the learning on the target task utilizing the same network sequentially [301]. Towards a similar goal but with on a slightly different line, Knowledge Distillation (KD) is a method for transferring information from one network to another network whilst training constructively [302], [303]. The most prominent setting of KD is a Student-Teacher (S-T) scheme, where the knowledge is transferring from a “Teacher (T)” model to a “Student (S)” model and in this manner, the teacher CNN is supervising the training of the student CNN. Since the knowledge from the teacher reflects a more general type of information that could be expressed through many representations, there is no commonly agreed rule as to how knowledge is transferred. Therefore, various forms of KD methods are developed covering different aspects, like the types of distillation, the quality measures of knowledge, the design of S-T architecture, etc.

In brief, the KD schemes are either offline or online. In offline distillation, the teacher model is pre-trained and fixed, and its knowledge is distilled to train the student model. On the other hand, in online distillation, both the teacher and student models are updated simultaneously in an end-to-end training procedure. Self-distillation is a special case where the teacher and student models are the same. Additionally, there are variations in the number of teachers used, including distillation from one teacher or multiple teachers, where the student learns from an ensemble of teachers. Also, the S-T framework has been extended handling various data formats, such as data-free or cross-modal KD, and different labeling requirements, including label-free or meta-data KD. Additionally, different learning metrics have been utilized, involving adversarial distillation and KD using attention maps. Thus, there is a wide range of S-T variations developed, each tailored to the specific characteristics of the problem at hand. A detailed survey on KD and S-T learning methods can be found in [24], [45].

Knowledge often refers to the learned weights and biases, although there is a diversity in the sources of knowledge in a CNN. Typically, the two principal sources of knowledge in a CNN model are, the output prediction score, known as logits, and the activations of intermediate layers, known as hints. Since the soft logits represent the class probability distribution, the knowledge from teacher’s model is shifted to the student’s model by learning the class distribution via softened softmax (also called “soft labels”), where each soft label’s contribution is controlled using a parameter defined as temperature [302], [303]. The main idea is that the student model will learn to mimic the responses of the teacher model and not only the hard class predictions. However, since CNNs are compositional models that organize the information hierarchically, they could learn multiple levels of feature representation with increasing abstraction [304] and thus, the knowledge derived from the intermediate layers of a teacher model could provide favorable information. Like so, the goal of this type of KD (Feature-based KD), is matching the internal representations between student and teacher models. Supplementary to the above sources, the knowledge that captures the relationship between different activations and neurons -from one or more locations of features along the network- can also be used to train a student model. Next, we present a detailed description of the used FKD, explaining how knowledge is measured and how the information is transferred from teacher to student through the proposed loss functions.

5.3.2 Geometric Regularization through Local Activation Features

The local features are the activations from intermediate layers of a CNN, meaning that they are the output of a hidden layer that constitute a Feature Map (FM). The FM is an intermediate representation generated from a convolution layer and thus, includes local information since each entry of FM highlights only a local neighborhood of input pixels. In an S-T FKD scheme, the teacher's intermediate representations supervise the training of the student model, so to learn feature representations that match some qualities of the respective teacher's predictions.

Given its spatial structure, a FM can be considered as a set of multidimensional vectors representing local features depth wise. The overall affinity between two sets of multidimensional data (feature vectors) can be measured through a similarity or dissimilarity function, formulated from either statistical or geometrical perspective. According to the statistical approach, the distance between two sets of feature vectors is related to the dissimilarity between the underlying distributions from which the vectors are derived. On the other hand, the geometrical approach assumes that the data from each set of vectors are lying on a low-dimensional manifold inside the feature space and thus, the distance can be defined as a measure of the dissimilarity between geometrical properties of the corresponding manifold structures. In [305], a manifold-to-manifold distance is introduced based on the notion of reordering efficiency of the neighborhood graphs representing the manifolds of local features. Following, in [306], [307] this distance was extended to an efficient Feature-based Knowledge Distillation (FKD) technique through a geometric regularization of local activations within an S-T framework. Consequently, the local manifold-based regularization incentivizes a student CNN to create local features that resemble, in overall geometry, to those of a teacher model at several layers with matching spatial resolutions. In this work, we employ a FKD approach which is based on the above-mentioned manifold-to-manifold distance, regularizing the activations in several intermediate layers of student model via the respective activations in the teacher network.

5.3.2.i Geometric Distillation

Let us consider a Feature Map (FM) of size $H \times W \ \times D$, where *H* and *W* correspond to its spatial size (Height and Width) while the Depth size *D* denotes the number of channels. It consists of *N=H·W* feature vectors, each one having dimensionality of *D* (i.e., $x_j \in \mathbb{R}^{1\times D}$ is a channel-wise feature vector with D elements, one for each pixel location j=1,…,N). Thus, a FM is a set of N feature vectors in a feature space of size D. Hence, the dissimilarity between two feature maps extracted from two CNNs could be measured via a manifold-to-manifold distance metric between the local activation manifolds, at corresponding layers of the two models with one-to-one correspondence between samples of the two compared sets. This happens in the case where the two compared feature maps have the same spatial size (H&W), independent of the feature dimensionality (i.e., the value of D). The neighboring relations within each feature set can be encoded by a Minimal Spanning Tree (MST), which in the form of a minimalistic backbone, connecting the nodes representing feature vectors [308]. In such case, neighborhoods can be defined via a geodesic radius around each node on the MST. The MST was used in such setting because it is less prone to topological short-circuits and thus, generating neighborhoods whose affinities are more indicative of the underlying manifolds' features [309]. Finally, a measurable quantity of local affinity for each FM's vector can be obtained with the Neighborhood Affinity Contrast (NAC) [310]. The NAC measures the ratio of the sum of square Euclidean distances of a sample to all its neighbors, to the sum of distances to all the other samples of the set. Thus, NAC is an atypical measure of compactness of the local neighborhood of each sample. The, the NAC ratio (i.e., intra distance to inter distance) is calculated using the following formula:

$$NAC_M^{FM} = \frac{\sum_{j=1}^{N} dist_{ij} \cdot m_{ij}}{\sum_{j=1}^{N} dist_{ij}} \in \mathbb{R}^{1\times N} \qquad \textit{eq. 5.1}$$

where the total number of FM vectors is $N = H \cdot W$, the pairwise vectors' normalized square Euclidean distance is $dist_{ij} = \frac{\|x_i - x_j\|_2^2}{\sum_{j=1}^{N}\|x_i - x_j\|_2^2}, x_i \in \mathbb{R}^{1\times D}$ and the neighborhood mask $M \in \{0,1\}^{N\times N}$ is based on the geodesic distance between the i-th and j-th nodes (i.e., feature vectors) on the MST with

$$m_{ij} = \begin{cases} 1, & Distance_{geodesic}^{MST}(i,j) \le r \\ 0, & Distance_{geodesic}^{MST}(i,j) > r \end{cases} \qquad \textit{eq. 5.2}$$

with *r* a geodesic radius indicating the number of hops that define the neighbors of each node on the MST.

In this work, the neighborhood mask $M \in \{0,1\}^{N\times N}$ is computed only on the teacher's side (Mt) and once for each datum, as proposed in [306], [307], [310], to force the student model's activations to mimic the neighboring relations, as expressed in the corresponding activations in the teacher's model. Therefore, the student's model is guided to produce local activation features with similar geometrical characteristics to those of the teacher model. The comparison between a feature map from the teacher CNN (FMt) and a feature map from the student CNN (FMs) -with equal spatial resolutions- is provided by the mean squared distance between the respective NAC vectors from the teacher's and student's FM, using the same neighborhood mask Mt. Therefore, the Geometrical Loss of the local features from the intermediate layers in the utilized S-T FKD scheme is defined as:

$$GeomL = \|NAC_{Mt}^{FMt} - NAC_{Mt}^{FMs}\|_2^2 \qquad eq.\ 5.3$$

By designing in this way, multiple supervision connections between layers of the teacher and the student can be simultaneously implemented, adding the respective regularization terms in the overall loss function. Also, the geometrical regularization of local activations could work synergistically with any other KD loss as well as task-dependent loss terms. In our implementation, we utilize two connections between intermediate layers on the S-T scheme for geometric regularization of local activation features. Thus, there are two geometrical regularization terms targeting different layers of the networks, one in an early layer and one after the middle of the teacher's topology, connecting with layers equivalent in terms of spatial size to the student's topology (Figure 5-1).

5.3.3 Response Regularization through Global Features

Typically, when a CNN model is utilized for feature extraction, the extracted feature is provided by the penultimate layer, just before the classification layer. Since this feature is the network's response, optimized so to facilitate a classification task, it reflects the discriminative qualities learned by the model during training. Hence, if the global feature of a CNN model exhibits preferable characteristics, it is reasonable to try to teach the student model to imitate this directly. Fortunately, in a S-T scheme, the teacher's knowledge could be transferred to the student through the feature information of the respective responses, regardless of the classification task solved by the teacher and the respective data.

Assuming that the global features from the two models (teacher and student) have the same dimensionality, it is easy to formulate a function to compare them. In this work, the FKD through global features is incurred by minimizing the difference between the teacher response and student response using either the cross-entropy loss of the two temperature-scaled features, or a loss utilizing the cross-correlation matrix between the two feature vectors.

5.3.3.i Response Distillation based on cross-entropy

The softmax function transforms the input vector into a probability distribution promoting the highest value against others. Accordingly, the output values are restricted to the range of [0,1] with their sum being equal to 1, whilst the larger values are intensified, and the lower values are denoted. One effective calibration technique for rescaling the output values to increase the sensitivity of low probability candidates is the temperature scaling [302], [303]. The softmax with temperature parameter softens the distribution by penalizing the larger logits more than the smaller logits and thus, more probability mass will be assigned to the smaller logits. This characteristic could be very beneficial for our case, where high-dimensional feature vectors (and not class predictions) are utilized directly in the loss function. The softmax with temperature parameter τ for an input vector $f \in \mathbb{R}^{1\times K}$, where K is the feature dimensionality, is calculated as:

$$p_i = \frac{e^{(f_i/\tau)}}{\sum_l^K e^{(f_l/\tau)}} \qquad \textit{eq. 5.4}$$

The imitation of teacher's output by the student model is driven by the minimization of cross-entropy between the temperature scaled features extracted from two CNNs respectively. Thus, the Response Loss of the final features with the temperature-scaled cross-entropy function (T-CE) in the S-T FKD scheme is defined as:

$$RespL\text{TCE} = -\sum_i^K (\ q_i^t \cdot \log(p_i^s)\ ) \qquad \textit{eq. 5.5}$$

where p_i^s and q_i^t are the final extracted features after temperature softmax from the student and the teacher respectively while K equals to the features' dimensionality.

In our implementation of S-T FKD framework, the above Response regularization term is evaluated either in conjunction with classification loss or other KD losses, or as a single loss term of the training procedure.

5.3.3.ii Response Distillation based on cross-correlation:

The cross-correlation between two different signals can be used as a technique for comparing two signals. A commonly used extension of the simple cross-correlation is normalized cross-correlation which can detect the correlation of two signals with different amplitudes. The cross-correlation matrix of two vectors in the $\mathbb{R}^K$ space is a matrix with elements the cross-correlations of all pairs of elements of the vectors. In particular, the cross-correlation matrix C between the normalized responses from the teacher ($z^t \in \mathbb{R}^{1\times K}$) and the student ($z^s \in \mathbb{R}^{1\times K}$) is computed by the following formula:

$$C_{ij} \triangleq \frac{\sum^K z_i^s \cdot z_j^t}{\sqrt{\sum^K (z_i^s)^2}\sqrt{\sum^K (z_j^t)^2}} \qquad \textit{eq. 5.6}$$

where C_{ij} is the correlation between i-th element (z_i^s) of normalized student's vector ($z^s \in \mathbb{R}^{1\times K}$) with the j-th element (z_j^t) of normalized teacher's vector ($z^t \in \mathbb{R}^{1\times K}$), while the above relation could be also applied to vectors normalized via z-score standardization across batch, during the training of the proposed S-T framework.

An objective function which tries to make two feature vectors similar while reducing the redundancy between their components, can be expressed by enforcing their cross-correlation matrix as close to the identity matrix as possible. For this purpose, the Barlow Twins loss function [266] has been proposed as a self-supervised learning approach, comparing the embeddings of two distorted versions of an input image by the same network. In our work, we exploit a similar logic, but comparing the features extracted from two different networks instead, (i.e., teacher and student) for the same input image. Thus, the cross-correlation matrix between the feature responses of the teacher and student CNNs is computed, creating an S-T FKD scheme where the Response Loss for the final features is defined as:

$$RespL\mathrm{BT} = \sum_{i}^{K} (1 - C_{ii})^2 + \lambda \sum_{i}^{K} \sum_{j \neq i} \left(C_{ij}\right)^2 \qquad \textit{eq. 5.7}$$

with a trade-off parameter $\lambda \geq 0$ that controls the importance between the two terms of the loss, with the first term trying to pull the diagonal elements of the cross-correlation matrix towards 1 and the second term trying to minimize the off-diagonal elements. The features are centered and normalized to unit variance along the batch dimension before the calculation of the cross-correlation matrix.

In addition to the aforementioned loss term, we also evaluate a more relaxed version that offers some additional degrees of freedom to the student model's response, and is more compatible with the context of S-T distillation. Whilst the Response loss in Barlow Twins tries to match the response in the i-th element of student's vector to the i-th element of teacher's vector, the relaxed version tries to match the i-th element of student's vector with the element of teacher's vector with which it has the maximum correlation. The rationale behind this option is that since the different architectures of the student and teacher networks do not share parameters as in a Barlow Twins self-supervision setting, the two models could produce similar responses but within an arbitrary permutation of their features' elements. Hence, the utilized loss should facilitate such behavior and not necessarily enforce element-wise correspondence between the student's and teacher responses. Therefore, we opted for a loss function that accentuates the maximum correlation for each feature component. Ultimately, the proposed Response Loss of the final features in the S-T FKD scheme is defined as:

$$RespL\mathrm{BC} = \sum_{i}^{K} \left(1 - \max_{j}(C_{ij}) \right)^2 + \lambda \sum_{i}^{K} \sum_{j \neq argmax_{j}(C_{ij})} \left(C_{ij}\right)^2 \qquad \textit{eq. 5.8}$$

As in the previous case, the parameter $\lambda \geq 0$ is a coefficient that adjusts the balance between the invariance term which enhances the maximum activations between the compared features, and the redundancy reduction term that decorrelates all the remaining components of the features. For this relaxation of the Barlow loss function we will use the term Barlow Colleagues (BC), in contrast to the unmodified Barlow Twins (BT).

5.3.4 Building the Student-Teacher Knowledge Distillation (S-T FKD) Architecture

Figure 5-1 presents the proposed S-T FKD scheme which distils the knowledge via an offline approach, where the teacher model is fixed, and the knowledge is transferred using both the feature maps of intermediate layers and the final features, to leverage the local and global information respectively.

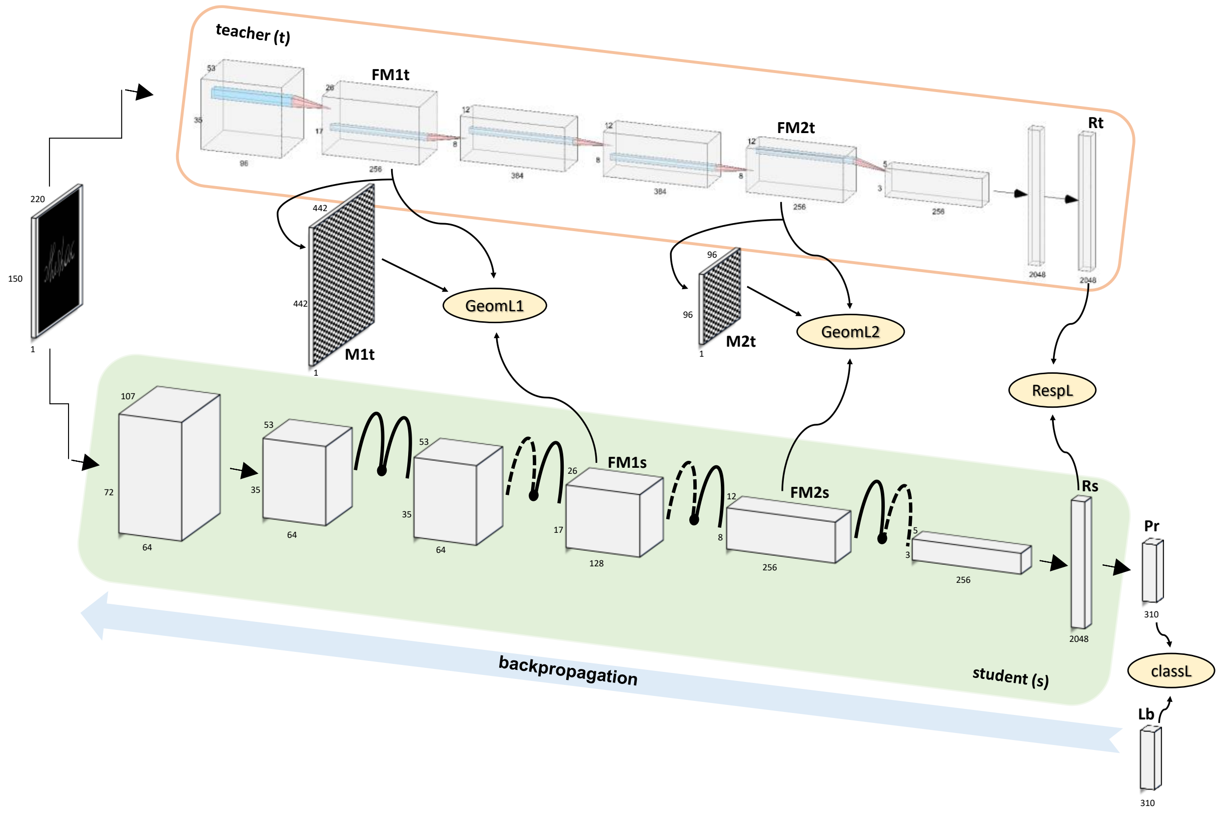

Figure 5-1: The S-T scheme presents the produced representations as data pass the sequential operations of student and teacher CNNs. The student CNN has a fully connected layer at the end of its topology that serves the classification task (classL) using the predicted classification scores (Pr). The teacher model provides the neighborhood masks (M1t and M2t) for any given Feature Map (FM1t and FM2t) as well as the final feature vector response with 2048 elements (Rt). Two Feature Maps (FM1s and FM2s) as well as the final extracted feature (Rs) of the student model participate in the FKD. The FKD is implemented utilizing the geometrical regularization terms from intermediate layers (GeomL1 and GeomL2) and the response regularization term from the final extracted features (RespL). Thus, the overall multi-loss function combines the FKD regularizations together with the classification loss to supervise the training of the student CNN.

First, to form the S-T architecture, we utilize as the teacher CNN the original SigNet feature extractor proposed by Hafemann et al. in [102]. The used SigNet model provides a feature representation for any input image and not a predicted classification result whereas a classification score depends on the number of classes of the corresponding training dataset. Thus, for any input image, the SigNet produces a feature vector of 2048-dimensions. While the teacher SigNet is built on the base of AlexNet architecture, the student CNN follows a more modern architecture based on the ResNet-18 topology. Taking advantage from its residual skip connections via the four residual blocks, the student model is much deeper than the teacher model even though they have approximately equal number of learned parameters given the addendum of one fully connected layer with 2048 neurons in the student CNN as its penultimate layer for feature extraction before the final classification layer. The inclusion of the fully connected layer for feature extraction has as input the feature map of the previous layer without using any spatial pyramid layer. The architectures and sizes of activations both for the teacher and student models are presented in Table 5-2.

Table 5-2: Architectures and activations' size for teacher (SigNet) and student (ResNet18) models.

SigNet		ResNet-18	
Layer/Block Name	**Activation size**	**Layer/Block Name**	**Activation size**
Input	150 × 220 × 1	Input	150 × 220 × 1
conv1	35 × 53 × 96	conv	72 × 107 × 64
pool1	17 × 26 × 256	pool	35 × 53 × 64
conv2 (FM1t)	17 × 26 × 256	block1	35 × 53 × 64
pool2	8 × 12 × 256	block2 (FM1s)	17 × 26 × 128
conv3	8 × 12 × 384	block3 (FM2s)	8 × 12 × 256
conv4	8 × 12 × 384	block4	3 × 5 × 256
conv5 (FM2t)	8 × 12 × 256	fc1 (Rs)	2048
pool5	3 × 5 × 256	fc2	#classes
fc6	2048		
fc7 (Rt)	2048		
fc8	#classes		

Secondly, the geometrical regularization requires to define the positions of distillation in the two models while the unique meeting condition is the same spatial resolution of the two connected feature maps. We select to utilize the final representation for each spatial resolution assuming that incorporates its optimal knowledge. For an input image of 150 × 220 pixels, the spatial pixel resolutions during the pass inside the teacher SigNet model are 35 × 53 (conv1-bn1-relu1), 17 × 26 (maxpool1 or conv2-bn2-relu2), 8 × 12 (maxpool2 or conv3-bn3-relu3, conv4-bn4-relu4, conv5-bn5-relu5), and 3 × 5 (maxpool5). Given the four different spatial resolutions founded along teacher SigNet's layers, the output from the first convolutional layer (i.e., 35 × 53) presumably captures primordial information and the smallest representation (i.e., 3 × 5) is degenerated; thus both they are deregistered. Hence, two spatial sizes, the 17 × 26 and 8 × 12 resolutions, are remaining. We opt to make use of the output after the consecutive operations of convolution, batch-normalization, and relu non-linearity supposing that this representation includes the best possible information as well as makes easier the correspondence with the output of a residual block in the student model. In this manner, the spatial resolutions lead to utilize the feature map representations produced as the output of the second and third residual block in the student's ResNet-18 model. Finally, the one geometrical regularization (*GeomL1*) is utilized the teacher's FM of size 17 × 26 × 256 and the student's FM of size 17 × 26 × 128 while the other geometrical regularization (*GeomL2*) uses the volumes with size 8 × 12 × 256 from the teacher and the student respectively. The response regularization (*respL*) is computed with the final extracted features, which are provided from the CNNs' layer when the classification output score is removed (i.e., if the classification task is not considered) and incorporate the total global knowledge of the network. Thus, the feature vector from the output of the feature extractor teacher (layer fc7 at SigNet model) and the feature vector from the fully connected layer (layer fc1 at ResNet-18 model) of the student CNN are utilized. Both features have 2048 elements and produced after the consecutive operations of fully convolution, batch-normalization, and relu non-linearity in the two models.

The Geometrical Loss (GeomL) of the intermediate feature maps (FM) from two different distillation positions (FM1t, FM1s and FM2t, FM2s) and the Response Loss (RespL) of the final extracted features (Rt, Rs) are implemented as regularization terms of the S-T FKD training. These are considered together with the typical (cross-entropy) Classification Loss (classL or CL) of the ground truth labels (Lb) and predicted classes (Pr), to form an overall multi-loss function that supervises the training of the student CNN. Thus, the overall multi-loss function of the S-T scheme is defined as:

$$\mathcal{L} = l_1 \cdot GeomL1 + l_2 \cdot GeomL2 + g \cdot RespL + c \cdot classL \qquad \textit{eq. 5.9}$$

with coefficients $l_1, l_2, g, and\ c$ representing the weights for each term that contributes to the overall loss, where $c = 1$ to allow compiling relative relations with the other terms.

5.4 Experimental Evaluation

5.4.1 Datasets

5.4.1.i Offline signature datasets for evaluation

The CNN models trained via the FKD processes are applied on three most popular OffSV datasets to assess their efficiency. As mentioned above, in OffSV setting, the CNN models are utilized for feature extraction, while WD classifiers undertake the final signature verification stage. The CEDAR, MCYT75, and GPDS300GRAY offline signature datasets are evaluated to measure the performance of the models as feature extractors. The three datasets include signatures scanned from documents as grayscale images.

The CEDAR dataset (Centre of Excellence for Document Analysis and Recognition) includes 55 writers with 24 genuine and 24 forgeries signatures per writer [56]. The forgeries are a mixture of random, simple, and skilled simulated signatures since they are contributed by some writers of the dataset that asked to forge three other writers' signatures, eight times per subject. The offline CEDAR dataset is publicly available.

The MCYT75 (Ministerio de Ciencia Y Tecnologia, Spanish Ministry of Science and Technology, MCYT75 Offline Signature Baseline Corpus ("Database")) has 75 enrolled writers with 15 genuine and 15 forgeries signatures per writer [57], [59]. The forgeries generated by 3 different user-specific forgers and thus, they are skilled simulated signatures. The offline handwritten signature dataset MCYT75 is publicly available.

The GPDS300GRAY is a standard subset of GPDS960 corpus (Digital Signal Processing Group) with 253 writers and each of them has 24 genuine and 30 forgeries signatures [246], [262], [311]. The forgeries signatures marked as skilled since they made by 10 forgers from 10 randomly selected genuine specimens and the forger was allowed to practice the signature without time limit. Although the GPDS960 database is no longer publicly available due to the General Data Protection Regulation (EU) 2016/679 ("GDPR"), we utilize the GPDS300GRAY dataset only for evaluation to measure the performance and to accomplish comparisons with other works that report results on this dataset [81], [83], [265].

Preprocessing signature images:

The grayscale signature images are subjected to some simple preprocessing steps dedicated to normalization, noise removal and size correction, since scanned images may contain noise and also the methods require the images in a fixed size. The normalization process shares the same steps as many previous works on OffSV [83], [84], [102], [103], something that also enables fair performance comparisons. The preprocessing includes the following steps: gaussian filtering and OTSU thresholding to remove background noise, centering into a large blank canvas of a predefined size of 952 × 1360 pixels (Height × Width) -common for all datasets- by aligning the signatures' center of mass to the center of the canvas so as not to affect the width of strokes and to present the original aspect ratio, inverting the images to have black background and grayscale foreground by subtracting the maximum brightness (i.e., white value of 255), and resizing the images to the input size of the CNNs that is 150 × 220 pixels (Height × Width).

5.4.1.ii Handwritten Text data for Training in S-T configuration

Training in a S-T configuration is realized using text data that work as information carrier to transfer the knowledge from the teacher into the student. Taking into consideration the biometric qualities of handwriting, the handwritten text data from the auxiliary domain are processed by a specially designed procedure to create an auxiliary task. The auxiliary task was designed so as its data to resemble more to those of the target domain of handwritten signature images. The text data come from the publicly available CVL-database, where 310 writers fill in 5-10 lines of predefined text on page-forms [312]. The text data are processed according to the procedure proposed by Tsourounis et al. in [264] to generate text images that resemble the distributions of signature images and use them as the training data of a CNN that solves a writer identification problem. In brief, the handwritten text documents are first converted to grayscale, then the lines of text are isolated as Solid Stripes of Text (SSoT), and finally, the SSoT are cropped into vertical intervals to generate the text images. Subsequently, the text images are preprocessed as signature images, following the same preprocessing steps detailed above. Given the findings in [264] about the effects of cropping and canvas dimensions of the text data, the random choice of parameters is more favorable offering better generalization. Hence, in this work the utilized training text set is obtained by each SSoT with random aspect ratio and using all available canvas sizes. In the context of this work though, the increased computational load needed for each training image -given the estimation of the Neighborhood Affinity Contrast (NAC) for some Feature Maps (FMs) as well as the calculation of the MST from the side of the teacher- led us to reduce the training set by sampling one text image for each canvas. Ultimately, about sixty thousand training and twenty-five thousand validation images are used for the training of S-T schemes (the training set could be downloaded from the official repository of our work[4]).

5.4.2 Experimental Setup and Protocols

In the context of this work, the trained CNN models are utilized as feature extractors, and the produced descriptors of the input signatures are consumed by binary classifiers that distinguish the genuine from the forgery signatures. Hence, the generalization performance of the CNN models is measured using the evaluation metrics obtained from the classifiers in the verification task. In this work, we follow the Writer Dependent (WD) approach and thus, a Support Vector Machine (SVM) classifier is trained for each writer. The implementation of WD classifiers is based on the work of Hafemann et al. [102]. In this manner and for fair comparisons, we utilized the implementation provided in the official repository[5] of [102] for the partition into training and test sets, the training of the classifiers, and the calculation of evaluation metrics.

5.4.2.i Experimental Protocol

Initially, a number of genuine signatures for every writer, denoted as the number of reference signatures N_{REF}, is selected to form the training set for the SVMs. In this manner, for each writer's SVM, the positive training class consists of the reference signatures from the writer while the negative training class is composed of the reference signatures from all other writers of the evaluated dataset (also called random forgeries). The test set for each writer includes the remaining genuine signatures from the writer and a number of the corresponding skilled forgeries. Following the works of [74], [84], [102], the number of training and testing samples used for evaluation on each dataset, is summarized in Table 5-3. Preferably, the number of genuine test samples would be equal to the test skilled forgeries (without diminishing the test set) for the two populations to have equal contributions to the error. The chosen number of the reference signatures per subject is in line with the most common experimental protocols in the literature [44]. For each experiment, ten (10) repetitions with WD classifiers trained using

[4] https://github.com/dimTsourounis/FKD
[5] https://github.com/luizgh/sigver

randomly selected splits of data are performed (with different reference signatures), and the results are presented in terms of the average and standard deviation values across these 10 iterations.

Table 5-3: The partition into training and test sets for the WD classifiers using the signature datasets.

OffSV Dataset				**Training set**		**Test set**	
Name	**Writers**	**Genuine**	**Skilled Forgeries**	**Genuine (writer's N_{REF})**	**Random Forgeries (others' N_{REF})**	**Genuine (Rest)**	**Skilled Forgeries**
CEDAR	55	24	24	12	12 × 54	10	10
MCYT75	75	15	15	10	10 × 74	5	15
GPDS300GRAY	253	24	30	12	12 × 252	10	10

5.4.2.ii Writer Dependent classifiers (WD SVM)

The WD classifiers were trained using soft margin binary Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel, while the two associated hyper-parameters (cost parameter C and scaling parameter gamma), were set to constant values of $C = 1\ and\ \gamma = 2^{-11}$. Also, more weight to the positive class is used in order to correct for the class imbalance, given that the positive training class consists of only a few genuine signatures and the negative training class has much more signatures due to the usage of samples from many writers. So, the weight for the negative class is set to 1 and the weight for the positive class is the ratio of the number of negative training examples to the number of positive training examples.

5.4.2.iii Evaluation Metrics

The number of reference signatures specifies the genuine signatures of a writer, used to construct the positive class during the training of its corresponding SVM, while the negative training class is created from the reference signatures of all other writers of the dataset. After the training of an SVM, a decision threshold should be defined to distinguish any query test signature as genuine or forgery. Mainly one of two approaches is followed to determine the decision threshold; either utilizing all the available training signatures of the datasets (from all the users) or using just the training signatures that correspond to each specific user, in order to select the threshold closest to FPR = 1 – TPR (i.e., False Positive Rate equals to one minus True Positive Rate). The first approach sets one optimum global decision threshold (a posteriori) that is common for all the writers' SVMs, and the second approach sets user-specific thresholds by using the optimal decision threshold for each writer's SVM individually. For calculating the False Acceptance Rate (FAR: misclassifying a forgery as being genuine) and False Rejection Rate ($FAR_{skilled}$: misclassifying a genuine as being skilled forgery), the global decision threshold is used.

The Equal Error Rate (EER) for each user is calculated considering only skilled forgeries (not random forgeries) when FRR equals to $\text{FAR}_{\text{skilled}}$ and using two forms, the global decision threshold to report $\text{EER}_{\text{globalthreshold}}$ and the user-specific threshold to report $\text{EER}_{\text{userthreshold}}$. In both cases, the reported EER is the average value from all the writers of the dataset and after the ten repetitions of experiment with different reference samples for every writer in each iteration. Finally, the mean Area Under the Curve (AUC) is often used as a metric measured on the ROC curves created for each user individually.

5.4.3 Implementation Details of S-T training

The only parameter of the utilized geometric regularization is the radius of r that defines the neighborhood size on each respective MST. In the following, we use a radius of $r = 5$, a value resulted as the most reliable setting for good performance in a small set of preliminary experiments with the selected teacher model. Also, following the observation made in [306], [307] that training is more stable when earlier layers have a smaller contribution in the overall loss than the deeper ones, we set the contribution coefficients $l_1 = 10 \, and \, l_2 = 100$ throughout the evaluation, since these values provided good results in the same preliminary experiments. The response regularization is implemented in three different ways, using the cross-entropy loss of temperature scaled features (T-CE), the cross-correlation matrix of normalized features based on Barlow Twins loss (BT), and the novel version of the latter named Barlow Colleagues (BC). For the first case, we ran a search for the temperature factor τ as well as the coefficient of contribution g and found the best results for $\tau = 10$ and $g = 0.001$. For the other two cases, the trade-off parameter $\lambda = 0.0001$ and coefficient $g = 0.0001$ are set after a grid search. Our observations suggest that a downscale (approximately two or three orders of magnitude) of the response distillation loss term relative to the classification loss, is beneficial to the overall performance. Additionally, the utilization of the regularization terms together with the classification term from the beginning of the training, produced better results than applying a warmup training with only the classification loss.

The S-T framework was trained using the Stochastic Gradient Descent (SGD) optimizer with initial learning rate of 0.01 which is reduced by a factor of 10 every 20 epochs for a total of 60 epochs, using Nesterov Momentum with a momentum factor of 0.9. The batch size was 64 according to the maximum capacity of the utilized GeForce RTX 2070 GPU. Each S-T training took approximately 30 hours. Implementation of the proposed FKD method is available for download at the official repository[6].

[6] https://github.com/dimTsourounis/FKD

5.4.4 Results and Analysis

5.4.4.i Proof-of-concept (SigNet-to-SigNet)

A goal of this work is to demonstrate teacher-to-student knowledge transfer using exclusively auxiliary data. The first setting we investigated as a proof-of-concept is when the teacher and student models follow the same architecture. In this manner, we investigate only if the transfer of knowledge is effective, without any performance contributions stemming from architectural differences. In this experiment, the teacher model is a SigNet, which is trained on signature images [102] and is not updated during S-T training, while the student model follows the SigNet's architecture with random parameters initialized using the Xavier sampling [313]. The student model is trained on the task of text-based writer identification, utilizing different combinations of distillation losses along with the main classification loss (CL). The evaluation of the trained student model was performed on the signature verification task, using WD classifiers that trained on the features extracted from layer fc7 of the student model and following the user threshold approach to calculate the performance metrics. Specifically, the Equal Error Rates (EER) are reported for the three evaluation datasets. Table 5-4 summarizes the obtained EER values from the CNN models trained in both standard identification task (CL), as on the various S-T training configurations. The results obtained using the teacher model as well as the student model at its initial conditions (i.e., with Random Weights) are also reported as baseline performance.

Table 5-4: Performance of the WD classifiers with user-threshold on the SigNet-to-SigNet FKD schemes.

Method	**Overall loss**				**EER (user threshold)**		
	l_1	l_2	g	c	**CEDAR** ***(N_{REF} = 12)***	**MCYT75** ***(N_{REF} = 10)***	**GPDS300GRAY** ***(N_{REF} = 12)***
Teacher – SigNet (fixed)	-	-	-	-	4.33±0.66	3.14±0.60	3.29±0.24
Student – SigNet (Random Weights)	-	-	-	-	11.95±0.81	10.71±1.17	9.53±0.39
CL (w/o KD)	0	0	0	1	3.91±0.60	8.27±0.67	4.28±0.17
CL + KD: GEOM	10	100	0	1	3.58±0.24	7.30±0.74	3.39±0.17
CL + KD: T-CE	0	0	0.001	1	2.94±0.34	8.26±1.42	4.11±0.19
CL + KD: BT	0	0	0.0001	1	2.94±0.34	4.76±0.97	4.46±0.38
CL + KD: BC	0	0	0.0001	1	3.78±0.56	7.92±0.81	3.78±0.24
CL + KD: GEOM & T-CE	10	100	0.001	1	3.02±0.34	7.49±1.20	3.37±0.19
CL + KD: GEOM & BT	10	100	0.0001	1	3.73±0.56	7.55±1.29	3.40±0.27
CL + KD: GEOM & BC	10	100	0.0001	1	3.51±0.32	7.02±1.26	3.17±0.16

As can be easily inferred by the results of Table 5-4, the utilization of any feature knowledge distillation (FKD) technique together with classification loss (CL) is beneficial in apposition to sole CL. Also, training only on the textual task alone (CL) produces a model with less discriminative features for the OffSV task, delivering inferior performance to the teacher model, but superior to randomly initialized model as expected. Since many experiments provide EER values with small differences, statistical tests for the ten repetitions in each setting using both Friedman and paired signed-rank Wilcoxon with $p<0.05$ were performed to clarify the comparisons against the teacher model. First, we can observe that the trained student model is statistically better in the CEDAR dataset for all the FKD schemes, while all variations are statistically on par with the teacher model in the GPDS dataset. Secondly, the results in the MCYT dataset are inconclusive both for the standard CL training and on the S-T schemes. An exception in the above is the case when classification loss and response distillation with BT loss (i.e., CL + KD: BT) is applied, where worse performance in GPDS dataset and better in MCYT are observed. An explanation behind this behavior could be that the architecture between teacher and student CNNs is the same. Finally, the large EERs in MCYT dataset meaning there is a trammel that degrade the performance, and this could be caused by the utilized text data that cannot adequately simulate the distribution of signatures on this dataset, with the limited capability of the student CNN having also a negative impact. In the following experiments, the same training regime is retained but the student CNN is changed from the AlexNet-based SigNet topology to the modern and efficient ResNet-18 architecture.

5.4.4.ii Model-to-Model Experiments (SigNet-to-ResNet)

Once the functionality of the proposed mechanism for knowledge transfer is established, our main goal is to train and evaluate new models with the ResNet-18 architecture. Additionally, we examine the effects of local and/or global distillation terms in conjunction with the baseline CL loss. The classification loss term CL is utilized throughout all experiments, since it is beneficial to the overall performance of the teacher model, as has also been indicated in several related studies in the literature [302], [306], [307].

The Table 5-5 includes the experimental results (EER with user threshold) for the three offline signature datasets following the Writer Dependent (WD) evaluation with the trained ResNet-18 models for feature extraction. In order to provide a baseline, in the same Table we also report the results obtained by the ResNet-18 model with randomly initialized weights using Xavier initialization [313]). For completeness, the model trained only with the classification objective (CL loss only) is also presented in the Table 5-5.

Table 5-5: Performance of the WD classifiers with user threshold on the SigNet-to-ResNet FKD schemes.

Method	**Overall loss**				**EER (user threshold)**		
	l_1	l_2	g	c	**CEDAR** **(*N_{REF} = 12*)**	**MCYT75** **(*N_{REF} = 10*)**	**GPDS300GRAY** **(*N_{REF} = 12*)**
Teacher (fixed)	-	-	-	-	4.33±0.66	3.14±0.60	3.29±0.24
Student (RW)	-	-	-	-	9.86±1.31	10.44±0.99	8.57±0.31
CL (w/o KD)	0	0	0	1	2.39±0.36	3.98±0.75	3.68±0.27
CL + KD: GEOM	10	100	0	1	2.24±0.45	3.67±0.73	2.95±0.24
CL + KD: T-CE	0	0	0.001	1	1.92±0.32	4.53±1.04	3.44±0.34
CL + KD: BT	0	0	0.0001	1	2.37±0.38	4.25±0.77	3.65±0.28
CL + KD: BC	0	0	0.0001	1	2.07±0.43	3.22±0.63	2.89±0.28
CL + KD: GEOM & T-CE	10	100	0.001	1	2.19±0.38	3.52±0.74	2.87±0.20
CL + KD: GEOM & BT	10	100	0.0001	1	1.85±0.32	4.31±0.94	2.97±0.29
CL + KD: GEOM & BC	10	100	0.0001	1	**2.25±0.24**	**3.29±0.62**	**2.74±0.28**

Comparison between KD and CL losses:

As can be easily inferred from the results, the exploitation of any FKD method together with the CL is advantageous for the student's performance, since the combined optimization of any of the KD terms along with CL is better than CL alone. Also, the combination of local and global KD along with the classification task is the most effective KD method (i.e., CL + KD: GEOM & RESP) considering the performance on all the three datasets. Furthermore, the settings where global KD is realized via our proposed adaptation of Barlow Colleagues (BC) loss achieves the best overall performance for the three signature databases, while Barlow Twins (BT) loss or temperature scaling (T-CE) loss exhibit better results in only one dataset, while degrading results in the others (e.g., CL + KD: T-CE or CL + KD: GEOM & BT).

These conclusions are verified with statistical tests (Friedman and Wilcoxon with p-value at 0.05 on the ten repetitions of classifiers' EERs), comparing the results of S-T training against the sole classification training. In this manner, the KD with BC either in combination with geometrical loss (CL + KD: GEOM & BC) or alone (CL + KD: BC) results to statistically significant improvement of EER on all three datasets while the geometrical regularization alone (CL + KD: GEOM) demonstrates significant difference only on GPDS dataset. The other two response regularization methods (CL + KD: BT, CL + KD: T-CE) as well as their combination with geometric regularization (CL + KD: GEOM & BT, CL + KD: GEOM & T-CE) produce results which are statistically equivalent to those achieved by using only CL loss. Finally, it is interesting to observe that the ResNet model trained only with CL loss, has better results than the SigNet architecture from the previous experiment (Table 5-4), proving the greater capability of ResNet architecture and confirming the need to utilize more contemporary architectures for OffSV. Weak evidence on that can also be derived from the comparison of the two architectures with random weights, where the ResNet (RW) is superior to SigNet (RW). Ultimately, the S-T scheme provides feature extraction models superior to those obtained by training only to the task of text classification, yet utilizing the same data sources but inducting the prior knowledge of teacher on the OffSV task.

Comparison between KD and Teacher's performance:

The student model resulted from S-T training with FKD via geometrical and response regularizations together with the CL loss, clearly outperforms the teacher in both GPDS and CEDAR datasets. Since the best results obtained when utilizing both local and global based KD, the teacher model is initially compared with these three student models. An in previous, Friedman's test and Wilcoxon paired signed-rank test were used again, with a 5% level of significance, for the ten repetitions of classifiers, using the same permutations of reference and test signatures for the comparisons. The exploitation of geometrical and global KD along with the CL (i.e., CL + KD: GEOM & RESP (of BC, BT, or T-CE)) achieves statistically better performance than the teacher SigNet model for the GPDS and CEDAR datasets while delivering statistical equivalent results in two out of the three cases for the MCYT dataset. For example, the CL + KD: GEOM & BT combination has a bad effect in the performance that can be justified from the different CNN architectures between student and teacher, similar to the comments on the proof-of-concept section above. For completeness, the teacher's performance is also compared to the each of the KD versions individually. The student exhibits statistical difference in performance for all the cases expect that of temperature scaling loss (CL + KD: T-CE) in the GPDS, and the Barlow Twins loss (CL + KD: BT) in the MCYT dataset. Thus, we can observe that the most efficient single KD schemes are those utilizing geometrical loss (CL + KD: GEOM) and BC loss (CL + KD: BC), where the EER values are either lower or not statistically different than those of the teacher. Finally, the ResNet-18 model trained only with CL loss (without KD) is statistically inferior to the teacher model for the GPDS and MCYT datasets and statistically superior to the teacher for the CEDAR dataset. For the case of CEDAR, it is notable that the teacher has degraded performance anyway, probably due to the large canvas size since used universally for all three datasets, in an aim to eliminate unrealistic dataset-dependent pre-processing parameters. At last, the S-T training and specifically the setting with CL + KD: GEOM & BC losses outperforms the teacher on the OffSV problem, without using any signature images for training the feature extraction model.

Comparison between KD methods:

According to the above, FKD methods using geometrical loss and/or BC response loss are the most beneficial for training feature extraction models. In this section we compare the different KD methods using additional statistical tests to characterize the differences among them. Seven different S-T KD versions in the three signature datasets are compared, and an overview of the findings is provided. The geometrical loss (CL + KD: GEOM) is statistically in tie with the BC response loss (CL + KD: BC) as well as with the two combinations of geometrical and response BT or T-CE losses (CL + KD: GEOM & BT/T-CE) for all the three datasets. Besides, the combination of geometrical and BC losses is statistically superior in the GPDS dataset and statistically equivalent to the other two datasets as compared to using solely geometrical loss. Between the response losses, the BC approach is statistically better than the other two losses in GPDS and MCYT, while it has statistically equivalent results with BT and worse than T-CE in CEDAR. Furthermore, the utilization of BC in the combination of local and global KD has statistically better performance in the GPDS and MCYT datasets as well as statistically equivalent results to T-CE and worse behavior than BT in CEDAR dataset. As a general conclusion, the local information seems to have significant importance for the final performance, and also the exploitation of BC regularization in the overall loss reflects a safe and efficient solution across datasets. The regularization of local features on earlier layers guides the training to a higher degree, avoiding the divergence of learning process such could be induced by KD methods relying only on global information where the regularization is applied deeper in the network. Nevertheless, an appropriate response regularization loss could conflate global and local information and capitalize on the joint power of local and global features, as in the case of the BC loss that cooperates efficiently with the geometrical regularization allowing multiple degrees of freedom during the student's training. After all, the greater performance of student over teacher in two out of three datasets (GPDS and CEDAR) and the tantamount of student and teacher results in the other dataset (MCYT) confirms the efficiency of the proposed S-T framework. Also, our choice to utilize a modern ResNet-based architecture (changing from AlexNet-based) has a good impact in order to exploit optimal the knowledge of the teacher. Ultimately, the proposed FKD methods enable the expert CNN (teacher) in signature signals to supervise the learning of the ResNet student without the need to utilize signatures during training and finally provide an effective CNN-based feature extractor for OffSV.

Label-free FKD:

Although the classification loss only requires the information that a text document from which we extract the text images is written by a specific writer, the unsupervised version of the S-T FKD framework is fully disengaged from the writers' IDs. Thus, the unsupervised FKD configuration eliminates the need for labeled data, making the multi-loss function solely consist of knowledge distillation (KD) terms without the classification loss (CL). This means that the specific writers of the handwritten texts do not need to be known, allowing for a large abundance of handwritten text data from various sources to be easily accessed and used for training the S-T scheme. This unsupervised S-T scheme aligns well with real-world conditions where handwritten text data can be utilized for training without the need to identify the writer. Consequently, the unsupervised S-T KD scheme facilitates the transfer of knowledge to a different CNN architecture for the OffSV task and also serves the most practical scenario. Figure 5-2 presents the box plots of the EER values following the Writer Dependent approach with user-specific thresholds for three signature datasets and three different cases where the geometrical loss and the response loss (KD: GEOM & RESP, using BC, BT, or T-CE) are utilized as the loss functions during S-T training.

The absence of a specific recognition problem, such as classification, and the use of feature regularizations as loss functions provide increased flexibility during training of the student model. The flexibility, coupled with the relaxed response loss, introduces the risk of training divergence. This could explain why the utilization of the BC loss yields the poorest results compared to the other two cases, which impose stricter constraints. In particular, the BT case demonstrates the best overall performance under the unsupervised FKD scheme. It is evident that the performance of the student model in the unsupervised case is inferior to that of the teacher model and the sole classification approach (CL (w/o KD)). However, the small difference (approximately 1% for the best unsupervised case) encourages further research in this direction. The practical advantage of the unsupervised approach lies in its ability to avoid the need for writer identification in the utilized handwritten text. However, exploring this aspect falls beyond the scope of the current work.

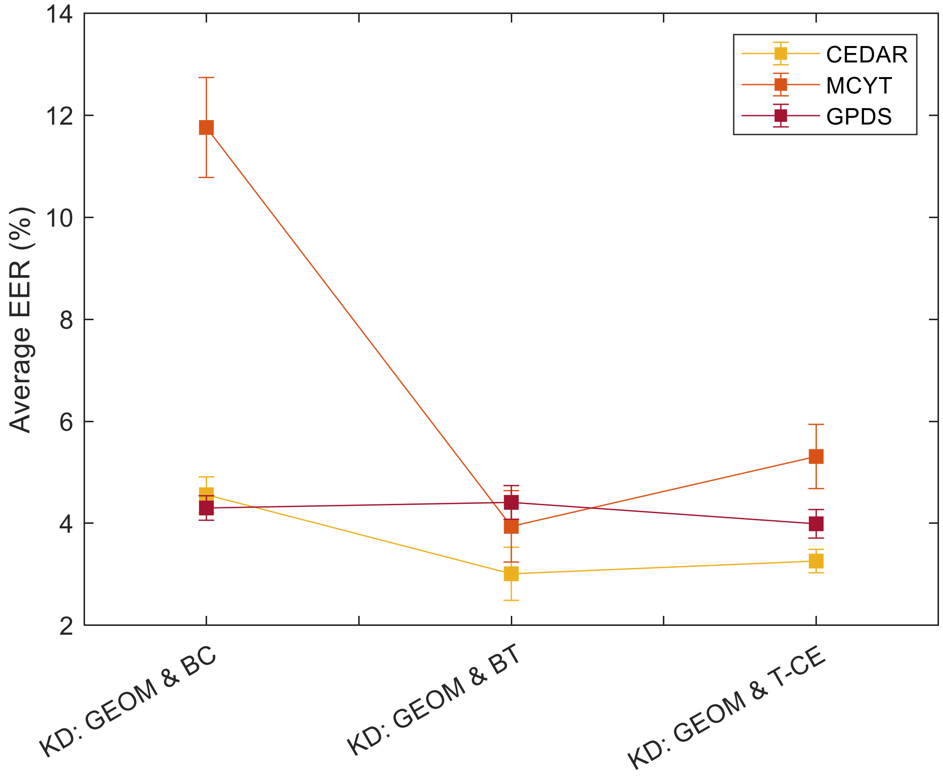

Figure 5-2: Comparison of the performance among three unsupervised FKD settings with different response losses in the S-T scheme. The models employ FKD with a combination of geometrical loss and one of the following response losses: BC loss, BT loss, or T-CE loss.

5.4.4.iii Summary of state-of-the-art WD OffSV systems

In this section we summarize the state-of-the-art (SoTA) methods for Writer Dependent (WD) OffSV systems, evaluating their performance to the proposed system. Given the inherent differences between various methods, there are many additional variables in the implementation of the systems' stages that renders the task of fairly comparing all of them very difficult. Hence, the purpose of the presentation of SoTA literature is to provide a general overview for the WD OffSV field, denoting the most important results in the three most popular datasets of CEDAR, MCYT75, and GPDS300GRAY.

The Table 5-6 presents a summary of the related SoTA works for WD OffSV task using the EER metric. Also, it includes the number of reference signatures used to form the positive class for training the WD classifiers. A common number of reference signatures could be found across methods for each dataset, despite the differences between methods as well as the different approaches for selecting these reference signatures.

Table 5-6: Summary of state-of-the-art OffSV Systems in terms of EER metric, for the CEDAR, MCYT75, and GPDS300GRAY datasets.

Refs	OffSV system		CEDAR		MCYT75		GPDS300GRAY	
	Authors, Year	**Method**	N_{REF}	**EER**	N_{REF}	**EER**	N_{REF}	**EER**
[233]	Soleimani et al., 2016	HOG + DMML	-	-	10	9.86	10	20.94
[253]	Serdouk et al., 2017	HOT	-	-	10	10.60	12	9.30
[73]	Diaz et al., 2017	Duplicator	-	-	12	9.12	12	14.58
[102]	Hafemann et al., 2017	SigNet	12	4.76	10	2.87	12	3.15
[102]	Hafemann et al., 2017	SigNet-F	12	4.63	10	3.00	12	1.69
[103]	Hafemann et al., 2018	SigNet-SPP	10	3.60	10	3.64	12	0.41
[276]	Lai and Jin, 2018	PDSN	10	4.37	10	3.78	-	-
[91]	Zois et al., 2019	SR – KSVD/OMP	10	0.79	10	1.37	12	0.70
[251]	Bhunia et al., 2019	Hybrid Texture	10	6.66	10	9.26	12	8.03
[77]	Maergner et al., 2019	CNN-Triplet and Graph edit distance	10	5.91	10	3.91	-	-
[100]	Shariatmadari et al., 2019	HOCCNN	12	4.94	12	5.46	-	-
[263]	Mersa et al., 2019	ResNet trained with text	-	-	10	3.98	-	-
[101]	Masoudnia et al., 2019	MLSE	-	-	10	2.93	-	-
[70]	Zois et al., 2020	Visibility Motif profiles	10	0.51	10	1.54	-	-

Refs	OffSV system		CEDAR		MCYT75		GPDS300GRAY	
	Authors, Year	**Method**	N_{REF}	**EER**	N_{REF}	**EER**	N_{REF}	**EER**
[74]	Maruyama et al., 2020	SigNet-F classifier gauss augments	3	0.82	3	0.01	3	0.20
[274]	Liu et al., 2021	MSDN	10	1.75	-	-	-	-
[193]	Yapici et al., 2021	Cycle-GAN	-	-	10	2.58	-	-
[72]	Zheng et al., 2021	micro deformations	12	2.76	-	-	-	-
[264]	Tsourounis et al., 2022	CNN-CoLL	10	1.66	10	1.62	12	2.12
[84]	Viana et al., 2023	MT-SigNet (Triplet)	12	3.50	10	2.71	-	-
[84]	Viana et al., 2023	MT-SigNet (NT-Xent)	12	3.32	10	3.22	-	-
	Proposed	S-T FKD (CL + KD: GEOM & BC)	12	2.25	10	3.29	12	2.74

The OffSV systems consist of three stages: the preprocessing, the feature extraction, and the classifier. These stages are designed according to the characteristics of each method and thus, many influential technicalities exist, like different preprocessing steps (e.g., such as different data preparation procedures of [91], [102], different input image size [72], [274], etc.), types of classifiers (e.g., like using SVM [84], [91], [101], [102], [253], one-class SVM [251], [276], Artificial Neural Networks (ANN) [100], [193], thresholding [77], [233], etc.) as well as major differences such as the type of training data (e.g., different training signature datasets [102], [251], private dataset [274], auxiliary data [263], [264], augmentation or synthetic data [73], [74], [84], [193], etc.). Although we have chosen so that the presented OffSV systems do not utilize skilled forgery signatures in the classifier's training and the number of reference signatures be common in many cases, the varying amount of the negative training class at the classifier has a major impact in the performance too (as also reported in the works of [102] and [74]). Additionally, the number of test samples differs since some methods utilize all the available signatures, considering the rest of the genuine and all the skilled forgeries (e.g., [91]), while other methods use equal number of test genuine and skilled forgery signatures -by selecting randomly the test skilled forgeries- (e.g., our implementation,[74], [84], [102]). Hence, easy comparisons between methods could be misleading and just a general outlook should be extracted. In this manner, we could argue that the proposed OffSV system proves the feasibility of achieving a low verification error, which is at least comparable to the state-of-the-art methods in all three datasets, despite nor using any signatures for training the feature extraction model.

5.4.4.iv Comparisons with SigNets

Finally, in this section we extensively compare the obtained models with two variants of SigNet: the teacher SigNet model and the SigNet-F model that used both genuine and skilled forgery signatures during its training (with the now defunct GPDS960 corpus) [102]. The comparisons are performed following identical preprocessing steps and utilizing the same partition of training and test signatures for the evaluated models to cast the comparisons as fair as possible. Thus, the respective classifiers are developed with common training and test sets across compared models. In this manner, the comparisons focus on the performance of the feature extraction stage and reveal the models' efficiency. Additionally, the calculation of multiple metrics provides a detailed performance analysis.

Table 5-7 provides a comprehensive evaluation of the performance achieved by the three compared feature extraction models: the SigNet-F, the SigNet (teacher) and our proposed model (ResNet from S-T KD with CL + KD: GEOM & BC) in the three signature datasets. The evaluation encompasses five different metrics, including the False Rejection Rate (FRR), the False Acceptance Rate (FAR) on skilled forged signatures, the EER values when global and user-specific thresholds are utilized, and the mean Area Under Curve (AUC) using the Receiver Operating Characteristic (ROC) curves. Additionally, the evaluation considers varying numbers of reference signatures (N_{REF}), ranging from 3, 5, 10, up to 12 signatures, to examine the model performance under different reference set sizes.

Table 5-7: Detailed comparison with SigNet and SigNet-F. All the reported metrics were obtained using WD classifiers while the FRR and FAR-skilled metrics are measured when the threshold is set to zero.

Dataset	Method	N_{REF}	FRR	FAR skilled	EER global threshold	EER user thresholds	AUC
CEDAR	SigNet-F	3	16.29 (± 0.68)	16.29 (± 0.79)	16.29 (± 0.73)	9.52 (± 0.95)	93.92 (± 0.88)
		5	14.02 (± 1.00)	14.13 (± 1.04)	14.07 (± 1.02)	8.78 (± 1.05)	94.76 (± 1.00)
		10	10.80 (± 0.78)	10.84 (± 0.91)	10.82 (± 0.85)	6.54 (± 0.97)	96.19 (± 0.67)
		12	10.60 (± 0.59)	10.60 (± 0.54)	10.60 (± 0.56)	5.99 (± 0.64)	96.66 (± 0.49)
	SigNet (Teacher)	3	13.51 (± 0.71)	13.47 (± 0.79)	13.49 (± 0.75)	6.75 (± 1.23)	96.27 (± 0.79)
		5	11.22 (± 0.68)	11.18 (± 0.71)	11.20 (± 0.68)	5.92 (± 0.47)	97.04 (± 0.28)
		10	8.33 (± 0.37)	8.36 (± 0.41)	8.35 (± 0.38)	4.34 (± 0.72)	97.84 (± 0.27)
		12	7.98 (± 0.55)	7.98 (± 0.58)	7.98 (± 0.56)	4.33 (± 0.66)	97.84 (± 0.39)

Dataset	Method	N_{REF}	FRR	FAR skilled	EER global threshold	EER user thresholds	AUC
	ResNet S-T FKD (Student)	3	7.85 (± 1.06)	7.85 (± 1.05)	7.85 (± 1.05)	3.39 (± 0.47)	98.50 (± 0.41)
		5	6.35 (± 0.90)	6.24 (± 0.81)	6.29 (± 0.85)	2.85 (± 0.27)	98.70 (± 0.29)
		10	4.78 (± 0.43)	4.71 (± 0.36)	4.75 (± 0.39)	2.20 (± 0.33)	99.11 (± 0.16)
		12	4.16 (± 0.36)	4.13 (± 0.37)	4.15 (± 0.36)	2.25 (± 0.24)	99.10 (± 0.19)
MCYT75	SigNet-F	3	10.05 (± 0.45)	10.09 (± 0.47)	10.07 (± 0.44)	5.99 (± 0.69)	97.16 (± 0.54)
		5	7.36 (± 0.68)	7.40 (± 0.65)	7.38 (± 0.66)	3.77 (± 0.71)	98.32 (± 0.39)
		10	6.32 (± 0.55)	6.30 (± 0.55)	6.31 (± 0.54)	3.19 (± 0.52)	98.52 (± 0.33)
		12	5.42 (± 0.52)	5.48 (± 0.69)	5.45 (± 0.58)	2.20 (± 0.58)	98.95 (± 0.35)
	SigNet (Teacher)	3	9.49 (± 0.77)	9.46 (± 0.75)	9.48 (± 0.76)	4.79 (± 0.87)	97.68 (± 0.45)
		5	7.15 (± 0.75)	7.05 (± 0.79)	7.10 (± 0.76)	3.86 (± 0.74)	98.21 (± 0.54)
		10	6.51 (± 0.40)	6.35 (± 0.38)	6.43 (± 0.38)	3.14 (± 0.60)	98.69 (± 0.21)
		12	6.09 (± 0.63)	6.19 (± 0.64)	6.14 (± 0.62)	2.73 (± 0.80)	98.80 (± 0.35)
	ResNet S-T FKD (Student)	3	13.31 (± 1.21)	13.32 (± 1.29)	13.32 (± 1.24)	7.94 (± 1.24)	94.70 (± 1.18)
		5	10.51 (± 0.46)	10.55 (± 0.46)	10.53 (± 0.46)	5.84 (± 0.86)	96.18 (± 1.08)
		10	7.12 (± 0.57)	6.03 (± 0.34)	6.49 (± 0.34)	3.29 (± 0.62)	98.43 (± 0.25)
		12	6.84 (± 0.92)	7.01 (± 1.07)	6.93 (± 0.99)	3.13 (± 0.66)	98.03 (± 0.52)
GPDS 300GRAY	SigNet-F	3	6.26 (± 0.26)	6.25 (± 0.27)	6.25 (± 0.26)	2.61 (± 0.30)	99.10 (± 0.12)
		5	5.01 (± 0.16)	5.01 (± 0.17)	5.01 (± 0.16)	2.04 (± 0.21)	99.37 (± 0.08)
		10	4.06 (± 0.16)	4.06 (± 0.17)	4.06 (± 0.16)	1.68 (± 0.08)	99.55 (± 0.05)
		12	3.93 (± 0.16)	3.92 (± 0.16)	3.92 (± 0.16)	1.53 (± 0.15)	99.57 (± 0.06)
	SigNet (Teacher)	3	9.29 (± 0.25)	9.29 (± 0.24)	9.29 (± 0.24)	4.79 (± 0.31)	97.92 (± 0.14)
		5	7.81 (± 0.24)	7.79 (± 0.25)	7.80 (± 0.25)	4.12 (± 0.26)	98.32 (± 0.12)

Dataset	Method	N_{REF}	FRR	FAR skilled	EER global threshold	EER user thresholds	AUC
		10	6.28 (± 0.19)	6.27 (± 0.18)	6.27 (± 0.18)	3.42 (± 0.21)	98.65 (± 0.13)
		12	6.02 (± 0.30)	6.05 (± 0.30)	6.04 (± 0.30)	3.29 (± 0.24)	98.72 (± 0.08)
	ResNet S-T FKD (Student)	3	8.76 (± 0.30)	8.77 (± 0.31)	8.76 (± 0.30)	4.76 (± 0.31)	98.05 (± 0.16)
		5	7.33 (± 0.23)	7.34 (± 0.24)	7.33 (± 0.23)	3.76 (± 0.33)	98.55 (± 0.14)
		10	5.49 (± 0.17)	5.51 (± 0.16)	5.50 (± 0.16)	2.86 (± 0.23)	98.93 (± 0.13)
		12	5.23 (± 0.23)	5.26 (± 0.20)	5.25 (± 0.22)	2.74 (± 0.28)	99.01 (± 0.10)

For each setting, Table 5-7 provides different qualities of the system's performance across the horizontal direction via the five calculated metrics while the varying number of reference signatures provides a different view of classifier's effectiveness. In this manner, we could derive a few important observations. First, the number of reference signatures has a critical impact on the performance of an OffSV system since decreasing the reference samples causes shrinkage on both the positive and the negative training class of the SVM. Hence, only a robust and discriminative feature extractor could assist the classifier to address the problem with a small amount of training samples. Secondly, the SigNet is clearly better than SigNet-F in CEDAR dataset and worse than SigNet-F in GPDS dataset. The inferior performance of SigNet in GPDS dataset though, is something totally reasonable because the knowledge of forged signatures from the same dataset is implicitly encoded in the SigNet-F offering a performance edge, given the later was trained with both genuine and skilled forgeries, even they originated from different writers of GPDS960 corpus (note that GPDS300 is a subset of GPDS960). For the MCYT dataset, the reported results are more complicated since the OffSV system using SigNet exhibits better performance on some metrics (e.g., global threshold for 3 or 5 references) and on the system based on SigNet-F on some other metrics (e.g., user-specific threshold for 3 references), while both systems are equivalent when 10 reference signatures are utilized. In the case of 12 reference signatures, only 3 genuine signatures remain for testing and consequently, the evaluation depends mainly on the test skilled forgery samples that is a less reliable indicator of performance (but it is included in the table for consistency with the other two datasets). In this manner, the performance disparity in MCYT dataset, given the amount of reference samples, relies on the classifier's effectiveness too, meaning that the SVM demands more training signatures to generalize well for all the cases. Ultimately, based on the above observations we can conclude that the SigNet exhibits greater generalization ability than SigNet-F across datasets and is actually a better choice as a single teacher in a S-T scheme.

In the light of the above, we finally compare the performance of ResNet to that of SigNet and SigNet-F on all datasets. The ResNet model trained with the proposed scheme is statistically superior to SigNet and statistically inferior to SigNet-F for the GPDS dataset for the same reasons mentioned above, while ResNet's higher performance is unambiguous for the CEDAR dataset. Thus, the student model achieves to surpass the teacher in both evaluated datasets. In fact, the utilization of a supplementary teacher, like SigNet-F, in a multiple-teachers FKD scheme maybe has positive impact and opens an interesting new path, but it is out of the scope of the current work. Lastly for the MCYT dataset, the ResNet has statistically equivalent performance with SigNet and SigNet-F for 10 reference signatures, whereas the system based on ResNet has inferior performance if the classifier needs to be trained with 3 or 5 genuine samples for the positive class. In conclusion, the experimental evidence from the three datasets support that the proposed FKD scheme comprises an adequate solution to transfer the knowledge from an expert CNN in the field of OffSV into a new (and modern) architecture, without the need fof any signature images, thus helping to create a new generation of models that could offer an excellent initial baseline for further research in Deep-Learning techniques for the OffSV problem.

5.5 Conclusions

In this work, we proposed a Feature-based Knowledge Distillation (FKD) learning framework applied to the OffSV problem. In the presented Student-Teacher (S-T) learning configuration, the knowledge is transferred from a benchmark CNN that provides efficient feature representations for signature images (acting as the teacher) into a new CNN model of different architecture (acting as the student). The only compatibility requirement between the teacher and student models are the spatial matching for at least some of intermediate activations and the common global feature dimensionality. We distilled knowledge through multiple layers via multiple connections among student and teacher topologies in order to incorporate both local and global information. We expressed the local information utilizing a manifold-to-manifold distance function that is designed to match the manifolds of local activations at the different layers of teacher and student models through geometric criteria of dissimilarity. Additionally, we promoted similarity in the feature responses of the student and teacher models by using loss functions that incorporate temperature-scaled cross-entropy or normalized cross-correlation to force the student's global features to imitate those of the teacher. The latter approach also included a novel loss function that leverages the cross-correlation matrix between the global features extracted from the student and teacher models respectively, considering the different architectures between the two CNNs. Hence, we presented, for the first time, a solution to inherit the prior knowledge of an effective CNN model into a new CNN model for signature representation learning through a KD method.

Since we did not have access to the signatures that the teacher model is trained on, we used auxiliary data from a related domain, such as images of handwritten text, as a source of information for KD. In this manner, we take advantage from the resemblance between handwriting in both texts and signatures, to overcome the lack of large amount of signature images that are required for the S-T training. The proposed S-T FKD framework is strengthened by utilizing knowledge from multiple layers and advanced relation-based distillation algorithms. These capture the correlations and higher-order output dependencies between the teacher and student models, enabling the student model to acquire the knowledge from a fixed teacher to a very large extent. Hence, when the response loss, calculated using the proposed Barlow-Colleagues (BC) cross-correlation function, is combined with the geometric loss, based on manifold-to-manifold distance, the student model becomes at least as efficient as the teacher, if not more so.

A significant motivation for this work was to enable the use of modern deep-learning models in OffSV, despite the current unavailability of a large signature dataset. Our FKD scheme serves as an excellent starting point for further research, as it incorporates knowledge from efficient OffSV models and addresses the lack of publicly available signature datasets. The presented OffSV system also shares common pre-processing and decision stages with other state-of-the-art methods, allowing for fair comparisons and emphasizing the unique contributions of each approach.

However, there are certain limitations in our study. The dependence on specific pre-processing steps for both signatures and text images restricts the flexibility of the system, as the teacher model has learned to encode information from pre-processed signatures. Nevertheless, we tried to mitigate the impact using common parameters during the signature pre-processing for all the evaluating datasets. We did not extensively analyze scale variations between raw text and signature images, limiting our understanding in that aspect since we addressed this implicitly by applying different canvas sizes during the generation of the training text images. Additionally, designing the student architecture with (spatially) matching intermediate activations and global feature dimension consistency with the teacher model, could pose some challenges for some architectures which may be a limiting factor from an architectural design perspective.

Ultimately, the main contribution of this work is the efficient knowledge transfer from a teacher model to any new architecture, as demonstrated with the use of a ResNet student. This allows for the leveraging of the efficiency of a basic and outdated teacher architecture, and the transfer of that knowledge into a deeper and more advanced student architecture for the efficient encoding of signatures. As a result, the student CNN can provide efficient feature representations for the OffSV task, even without utilizing any signature images during the FKD process.

Future work will include investigation of additional knowledge distillation methods and different distillation strategies. We believe that combining multiple teachers and multi-loss approaches could be promising for the OffSV task since they could improve the generalization ability of the student model and make it more effective across signature datasets and languages.

Additionally, future plans could incorporate synthetic signatures, using both existing synthetic datasets and generative methods to generate new signature images, during training a KD scheme. However, the main challenge in this research direction is the difference between the distributions of original and synthetic signatures, resulting in models that are only efficient on one or the other type of data. Another useful research approach is to develop CNN-based schemes that leverage diverse pre-processing steps, allowing for the decoupling of specialized preprocessing techniques tailored to specific datasets. This approach can be further enhanced by incorporating synthetic samples to augment the training process, particularly where a large number of real signatures is not available. All the aforementioned approaches could be also organized on a comprehensive study that involves a variety of KD methods for the OffSV problem, including both WI and WD evaluation phases for optimal deliberation. Finally, including few-shot learning techniques into a S-T framework would be an interesting area of future research.

Chapter 6

Summary and future directions

Various approaches have been investigated in this PhD thesis, aiming to address the challenges of small sample-size learning (SSSL) problem. The output of this research effort includes methods operating in the input domain, model domain, and feature domain of the overall learning tasks. By exploring different approaches, this research aspired to contribute to the existing body of knowledge on SSSL, offering some insights into effective strategies for handling limited data scenarios. While the majority of the proposed methods were developed in the context of the offline signature verification (OffSV) task, since it provided a useful ground for testing our methods in an intrinsically data-limited domain, it is important to note that the developed techniques are not inherently restricted to this specific application domain.

First, shallow representation models were investigated, employing traditional techniques aimed at addressing the challenge of limited sample size in offline signature verification task. Traditional methods heavily depend on hand-crafted features, demanding meticulous selections to align with the specific problem at hand. First and foremost, these methods necessitate expertise in feature engineering to select the most suitable technique for extracting features based on the characteristics of the underlying problem. Moreover, traditional approaches necessitate problem-specific preprocessing steps, often with writer-specific parametrization. Also, these methods often require the incorporation of an additional mechanism for encoding local descriptors, adding further complexity to the pipeline. However, shallow representation methods capture local image characteristics and provide low-level features which have shown effectiveness in recognition tasks, despite that their effectiveness often relies heavily on the careful selection of various design parameters by the user. In this PhD research, both hand-crafted and learned shallow (local) representations were studied in an effort to establish a performance baseline in the field of OffSV, producing noteworthy research outputs along the way.

Transitioning from shallow learning to deep learning, since this is the main domain of this dissertation, a hybrid approach was proposed that combines SIFT-descriptors and CNNs. This approach encompasses a hand-crafted feature generation process that transforms the pixel-wise image representation into relationships between statistical properties of the pixel regions resulting in a more compact representation of the information in the input image. The hypothesis is that the more informative representation can be beneficial to the training of CNNs with limited data. Results indicated that indeed this approach has a positive impact in effectively addressing the challenges posed by SSSL in various applications. However, it is important to acknowledge the limitations of this approach, especially when color plays a significant role in discrimination, since the utilized descriptor is essentially a gradient-encoding scheme. To overcome this limitation, a promising solution is the late fusion of global feature representations from both SIFT-CNN and pixel-CNN, treating them as two distinct streams within the framework. This fusion technique enables the combination of complementary information during the end-to-end training process, offering a potential direction for enhanced performance. By leveraging the strengths of both SIFT-CNN and pixel-CNN, this approach aims to exploit the advantages of each method and compensate for the limitations.

One of the significant practical challenges identified during the course of this PhD research was the retraction of the only publicly accessible large offline signature database, which was the only large enough dataset for training deep models. This event hindered the efforts of all research community around OffSV, to develop new deep architectures for addressing this problem. Such unique circumstances provided a useful testbed for this PhD research to develop methods that can circumnavigate such issues, thus we prioritized the design of efficient methods for the OffSV problem to address this current challenge.

The methods developed on that front were based in the idea of harnessing auxiliary, possibly unlabeled, data. Therefore, the primary task was to collect relevant training data from a similar domain with data abundance. Since handwritten text utilizes similar mechanics with signing, and there are ample data of digitized documents with handwriting in the wed, we focused on leveraging such data within a sophisticated training procedure, aimed at enhancing performance in the target problem of OffSV, using only limited training (signature) samples. To accomplish this, a specially designed preprocessing procedure was implemented, utilizing handwritten text documents to generate samples with similar signal characteristics to signature images. The handwriting information was extracted from the entire handwritten text document, enabling the generation of handwritten text training images without relying on word-oriented processing. This approach eliminates the need for word-level annotation and requires only the knowledge that the writer of the whole document is the same person. The preprocessing is simple and fast, making it highly suitable for large-scale data processing. While utilizing external handwritten text data during CNN training shows potential, additional domain adaptation techniques were employed to achieve comparable, if not superior, performance to models trained on the retracted datasets with thousands of signature images. Two approaches we developed: explicit domain adaptation and implicit domain adaptation. In the first case, the proposed method uses metric learning with a separate mapping layer - trained via contrastive loss- to embed the global features of the CNN in a new latent space. In the other case, the domain adaptation is achieved with the supervision of training by a teacher model, via a Feature-based Knowledge Distillation (FKD) scheme that utilizes both local and global information from intermediate representations. This method was proposed in an aim to demonstrate that public-domain trained models can be efficiently utilized into SSSL problems, even if their training data are not accessible. The proposed approaches effectively address the SSSL problem in the OffSV task, operating either on the feature space or the model space using auxiliary data in the input space. However, their main limitation lies in the reliance on specific preprocessing steps for input signatures, such as predefined parameters for canvas size in each dataset. In the model obtained from the FKD, an effort is made to alleviate this impact by using common parameters during the signature preprocessing for all evaluation datasets. The well performance model of the proposed method acts as proof of its generalization ability, highlighting its potential as a promising research direction not only in the current field but also in adapting to other domains. Therefore, our future plans include to assess the effectiveness of the proposed Knowledge Distillation loss functions on a large-scale benchmark.

In conclusion, during the course of this PhD research, we tried to approach the SSSL problem from multiple perspectives, developing methods that cover a wide range of circumstances. Although the proposed methods have been evaluated on specific applications, we consider them as contributions to the general SSSL research, and parts of which could serve as an initial steppingstone towards further research in the domain of generalized incremental few-shot learning (GIFSL), which represents one of the most cutting-edge and challenging learning scenarios within the field of machine learning. An exciting future research direction entails incorporating the proposed geometric criteria and loss functions of FKD developed in this thesis into GIFSL configurations, exploring new research directions and applications.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, Art. no. 7553, May 2015, doi: 10.1038/nature14539.

[2] A. Vaswani *et al.*, “Attention is All you Need,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Jul. 01, 2023. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[3] I. Goodfellow *et al.*, “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[4] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection.” arXiv, May 09, 2016. doi: 10.48550/arXiv.1506.02640.

[6] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, in NIPS’12. USA: Curran Associates Inc., 2012, pp. 1097–1105. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999134.2999257

[8] S. Ioffe and C. Szegedy, “Batch normalization: accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, in ICML’15. Lille, France: JMLR.org, Jul. 2015, pp. 448–456.

[9] R. Keshari, S. Ghosh, S. Chhabra, M. Vatsa, and R. Singh, “Unravelling Small Sample Size Problems in the Deep Learning World,” in *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, Sep. 2020, pp. 134–143. doi: 10.1109/BigMM50055.2020.00028.

[10] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *Journal of Big Data*, vol. 6, no. 1, p. 60, Jul. 2019, doi: 10.1186/s40537-019-0197-0.

[11] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen, “Image Data Augmentation for Deep Learning: A Survey.” arXiv, Apr. 18, 2022. doi: 10.48550/arXiv.2204.08610.

[12] L. Taylor and G. Nitschke, “Improving Deep Learning with Generic Data Augmentation,” in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, Nov. 2018, pp. 1542–1547. doi: 10.1109/SSCI.2018.8628742.

[13] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, “AutoAugment: Learning Augmentation Policies from Data,” 2019. Accessed: Jul. 05, 2023. [Online]. Available: https://arxiv.org/pdf/1805.09501.pdf

[14] S. Kingra, N. Aggarwal, and N. Kaur, “LBPNet: Exploiting texture descriptor for deepfake detection,” *Forensic Science International: Digital Investigation*, vol. 42–43, p. 301452, Oct. 2022, doi: 10.1016/j.fsidi.2022.301452.

[15] K. Man and J. Chahl, “A Review of Synthetic Image Data and Its Use in Computer Vision,” *Journal of Imaging*, vol. 8, no. 11, Art. no. 11, Nov. 2022, doi: 10.3390/jimaging8110310.

[16] B. Jena, G. K. Nayak, and S. Saxena, “Convolutional neural network and its pretrained models for image classification and object detection: A survey,” *Concurrency and Computation: Practice and Experience*, vol. 34, no. 6, p. e6767, 2022, doi: 10.1002/cpe.6767.

[17] L. Ruan and Q. Jin, “Survey: Transformer based video-language pre-training,” *AI Open*, vol. 3, pp. 1–13, Jan. 2022, doi: 10.1016/j.aiopen.2022.01.001.

[18] Z. Zeng *et al.*, “Knowledge Transfer via Pre-training for Recommendation: A Review and Prospect,” *Frontiers in Big Data*, vol. 4, 2021, Accessed: Jul. 05, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdata.2021.602071

[19] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?,” in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2014, pp. 3320–3328. [Online]. Available: http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks.pdf

[20] S. Tariyal, A. Majumdar, R. Singh, and M. Vatsa, “Deep Dictionary Learning,” *IEEE Access*, vol. 4, pp. 10096–10109, 2016, doi: 10.1109/ACCESS.2016.2611583.

[21] R. Keshari, M. Vatsa, R. Singh, and A. Noore, “Learning Structure and Strength of CNN Filters for Small Sample Size Training,” presented at the 2018 IEEE/CVF Conference on Computer Vision

and Pattern Recognition (CVPR), IEEE Computer Society, Jun. 2018, pp. 9349–9358. doi: 10.1109/CVPR.2018.00974.

[22] R. N. D'souza, P.-Y. Huang, and F.-C. Yeh, "Structural Analysis and Optimization of Convolutional Neural Networks with a Small Sample Size," *Sci Rep*, vol. 10, no. 1, Art. no. 1, Jan. 2020, doi: 10.1038/s41598-020-57866-2.

[23] J. Kukačka, V. Golkov, and D. Cremers, "Regularization for Deep Learning: A Taxonomy," Feb. 2018, Accessed: Jul. 05, 2023. [Online]. Available: https://openreview.net/forum?id=SkHkeixAW

[24] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," *Int J Comput Vis*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021, doi: 10.1007/s11263-021-01453-z.

[25] J. Lu, J. Hu, and J. Zhou, "Deep Metric Learning for Visual Understanding: An Overview of Recent Advances," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 76–84, Nov. 2017, doi: 10.1109/MSP.2017.2732900.

[26] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a Few Examples: A Survey on Few-shot Learning," *ACM Comput. Surv.*, vol. 53, no. 3, p. 63:1-63:34, Jun. 2020, doi: 10.1145/3386252.

[27] H. Xuan, A. Stylianou, X. Liu, and R. Pless, "Hard Negative Examples are Hard, but Useful," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 126–142. doi: 10.1007/978-3-030-58568-6_8.

[28] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality Reduction by Learning an Invariant Mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Jun. 2006, pp. 1735–1742. doi: 10.1109/CVPR.2006.100.

[29] K. Q. Weinberger and L. K. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Jun. 2009.

[30] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. M. Robertson, "Ranked List Loss for Deep Metric Learning," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5207–5216. Accessed: Jul. 05, 2023. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Wang_Ranked_List_Loss_for_Deep_Metric_Learning_CVPR_2019_paper.html

[31] A. K. Jain, A. A. Ross, and K. Nandakumar, *Introduction to Biometrics*. New York: Springer, 2011.

[32] S. Jabin and F. J. Zareen, “Biometric signature verification,” *International Journal of Biometrics*, vol. 7, no. 2, pp. 97–118, Jan. 2015, doi: 10.1504/IJBM.2015.070924.

[33] J. Espinal-Enríquez, R. A. Mejía-Pedroza, and E. Hernández-Lemus, “Chapter 13 - Computational Approaches in Precision Medicine,” in *Progress and Challenges in Precision Medicine*, M. Verma and D. Barh, Eds., Academic Press, 2017, pp. 233–250. doi: 10.1016/B978-0-12-809411-2.00013-1.

[34] X. Jiang *et al.*, “Biomedical Imaging: A Computer Vision Perspective,” in *Computer Analysis of Images and Patterns*, R. Wilson, E. Hancock, A. Bors, and W. Smith, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, pp. 1–19. doi: 10.1007/978-3-642-40261-6_1.

[35] A. S. Wiik, M. Høier-Madsen, J. Forslid, P. Charles, and J. Meyrowitsch, “Antinuclear antibodies: A contemporary nomenclature using HEp-2 cells,” *Journal of Autoimmunity*, vol. 35, no. 3, pp. 276–290, Nov. 2010, doi: 10.1016/j.jaut.2010.06.019.

[36] P. L. Meroni and P. H. Schur, “ANA screening: an old test with new recommendations,” *Annals of the Rheumatic Diseases*, vol. 69, no. 8, pp. 1420–1422, Aug. 2010, doi: 10.1136/ard.2009.127100.

[37] J. Iacovacci and L. Lacasa, “Visibility Graphs for Image Processing,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 974–987, Apr. 2020, doi: 10.1109/TPAMI.2019.2891742.

[38] Z. Gao, Y. Wu, M. Harandi, and Y. Jia, “A Robust Distance Measure for Similarity-Based Classification on the SPD Manifold,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3230–3244, Sep. 2020, doi: 10.1109/TNNLS.2019.2939177.

[39] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006, doi: 10.1109/TSP.2006.881199.

[40] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online Learning for Matrix Factorization and Sparse Coding,” *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Mar. 2010.

[41] Y. He, K. Kavukcuoglu, Y. Wang, A. Szlam, and Y. Qi, “Unsupervised Feature Learning by Deep Sparse Coding,” in *Proceedings of the 2014 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, Apr. 2014, pp. 902–910. doi: 10.1137/1.9781611973440.103.

[42] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:VISI.0000029664.99615.94.

[43] C. Liu, J. Yuen, and A. Torralba, "SIFT Flow: Dense Correspondence across Scenes and Its Applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 978–994, May 2011, doi: 10.1109/TPAMI.2010.147.

[44] L. G. Hafemann, R. Sabourin, and L. S. Oliveira, "Offline handwritten signature verification — Literature review," in *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Nov. 2017, pp. 1–8. doi: 10.1109/IPTA.2017.8310112.

[45] L. Wang and K.-J. Yoon, "Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3048–3068, Jun. 2022, doi: 10.1109/TPAMI.2021.3055564.

[46] M. M. Hameed, R. Ahmad, M. L. M. Kiah, and G. Murtaza, "Machine learning-based offline signature verification systems: A systematic review," *Signal Processing: Image Communication*, vol. 93, p. 116139, Apr. 2021, doi: 10.1016/j.image.2021.116139.

[47] R. Plamondon and G. Lorette, "Automatic signature verification and writer identification — the state of the art," *Pattern Recognition*, vol. 22, no. 2, pp. 107–131, Jan. 1989, doi: 10.1016/0031-3203(89)90059-9.

[48] R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63–84, 2000.

[49] D. Impedovo and G. Pirlo, "Automatic Signature Verification: The State of the Art," *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, vol. 38, no. 5, pp. 609–635, 2008, doi: 10.1109/TSMCC.2008.923866.

[50] S. Pal, M. Blumenstein, and U. Pal, "Off-line signature verification systems: a survey," presented at the Proceedings of the International Conference; Workshop on Emerging Trends in Technology, 1980163: ACM, 2011, pp. 652–657. doi: 10.1145/1980022.1980163.

[51] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan. 1979, doi: 10.1109/TSMC.1979.4310076.

[52] G. Peyré, "Manifold models for signals and images," *Computer Vision and Image Understanding*, vol. 113, no. 2, pp. 249–260, Feb. 2009, doi: 10.1016/j.cviu.2008.09.003.

[53] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J. C. Nuño, “From time series to complex networks: The visibility graph,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 13, pp. 4972–4975, Apr. 2008, doi: 10.1073/pnas.0709247105.

[54] B. Luque, L. Lacasa, F. Ballesteros, and J. Luque, “Horizontal visibility graphs: Exact results for random time series,” *Phys. Rev. E*, vol. 80, no. 4, p. 046103, Oct. 2009, doi: 10.1103/PhysRevE.80.046103.

[55] J. Iacovacci and L. Lacasa, “Sequential visibility-graph motifs,” *Phys. Rev. E*, vol. 93, no. 4, p. 042309, Apr. 2016, doi: 10.1103/PhysRevE.93.042309.

[56] M. K. Kalera, S. Srihari, and A. Xu, “Offline signature verification and identification using distance statistics,” *Int. J. Patt. Recogn. Artif. Intell.*, vol. 18, no. 07, pp. 1339–1360, Nov. 2004, doi: 10.1142/S0218001404003630.

[57] J. Ortega-Garcia *et al.*, “MCYT baseline corpus: a bimodal biometric database,” *IEE Proceedings Vision, Image and Signal Processing*, vol. 150, no. 6, pp. 395–401, 2003, doi: 10.1049/ip-vis:20031078.

[58] S. Chen and S. Srihari, “A New Off-line Signature Verification Method based on Graph,” in *18th International Conference on Pattern Recognition (ICPR’06)*, Aug. 2006, pp. 869–872. doi: 10.1109/ICPR.2006.125.

[59] J. Fierrez-Aguilar, N. Alonso-Hermira, G. Moreno-Marquez, and J. Ortega-Garcia, “An Off-line Signature Verification System Based on Fusion of Local and Global Information,” in *Biometric Authentication*, D. Maltoni and A. K. Jain, Eds., in Lecture Notes in Computer Science, vol. 3087. Springer Berlin Heidelberg, 2004, pp. 295–306. doi: 10.1007/978-3-540-25976-3_27.

[60] E. N. Zois, L. Alewijnse, and G. Economou, “Offline signature verification and quality characterization using poset-oriented grid features,” *Pattern Recognition*, vol. 54, pp. 162–177, Jun. 2016, doi: 10.1016/j.patcog.2016.01.009.

[61] Y. Serdouk, H. Nemmour, and Y. Chibani, “New off-line Handwritten Signature Verification method based on Artificial Immune Recognition System,” *Expert Systems with Applications*, vol. 51, pp. 186–194, Jun. 2016, doi: 10.1016/j.eswa.2016.01.001.

[62] J. F. Vargas, M. A. Ferrer, C. M. Travieso, and J. B. Alonso, “Off-line signature verification based on grey level information using texture features,” *Pattern Recognition*, vol. 44, no. 2, pp. 375–385, 2011, doi: 10.1016/j.patcog.2010.07.028.

[63] R. K. Bharathi and B. H. Shekar, “Off-line signature verification based on chain code histogram and Support Vector Machine,” in *2013 International Conference on Advances in Computing,*

Communications and Informatics (ICACCI), Aug. 2013, pp. 2063–2068. doi: 10.1109/ICACCI.2013.6637499.

[64] S. Y. Ooi, A. B. J. Teoh, Y. H. Pang, and B. Y. Hiew, “Image-based handwritten signature verification using hybrid methods of discrete Radon transform, principal component analysis and probabilistic neural network,” *Applied Soft Computing*, vol. 40, pp. 274–282, Mar. 2016, doi: 10.1016/j.asoc.2015.11.039.

[65] M. Okawa, “Offline Signature Verification Based on Bag-of-VisualWords Model Using KAZE Features and Weighting Schemes,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2016, pp. 252–258. doi: 10.1109/CVPRW.2016.38.

[66] M. Okawa, “From BoVW to VLAD with KAZE features: Offline signature verification considering cognitive processes of forensic experts,” *Pattern Recognition Letters*, vol. 113, pp. 75–82, Oct. 2018, doi: 10.1016/j.patrec.2018.05.019.

[67] G. Ganapathi and R. Nadarajan, “A Fuzzy Hybrid Framework for Offline Signature Verification,” in *Pattern Recognition and Machine Intelligence*, P. Maji, A. Ghosh, M. N. Murty, K. Ghosh, and S. K. Pal, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, pp. 121–127. doi: 10.1007/978-3-642-45062-4_16.

[68] A. Gilperez, F. Alonso-Fernandez, S. Pecharroman, J. Fierrez, and J. Ortega-Garcia, “Off-line Signature Verification Using Contour Features,” presented at the Proceedings 11th International Conference on Frontiers in Handwriting Recognition, Montreal, 2008.

[69] E. N. Zois, I. Theodorakopoulos, and G. Economou, “Offline Handwritten Signature Modeling and Verification Based on Archetypal Analysis,” presented at the 2017 IEEE International Conference on Computer Vision (ICCV), Oct. 2017, pp. 5515–5524. doi: 10.1109/ICCV.2017.588.

[70] E. N. Zois, E. Zervas, D. Tsourounis, and G. Economou, “Sequential Motif Profiles and Topological Plots for Offline Signature Verification,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13248–13258. Accessed: Dec. 21, 2020. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Zois_Sequential_Motif_Profiles_and_Topological_Plots_for_Offline_Signature_Verification_CVPR_2020_paper.html

[71] A. Alaei, S. Pal, U. Pal, and M. Blumenstein, “An Efficient Signature Verification Method Based on an Interval Symbolic Representation and a Fuzzy Similarity Measure,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 10, pp. 2360–2372, 2017, doi: 10.1109/TIFS.2017.2707332.

[72] Y. Zheng, B. K. Iwana, M. I. Malik, S. Ahmed, W. Ohyama, and S. Uchida, “Learning the Micro Deformations by Max-pooling for Offline Signature Verification,” *Pattern Recognition*, p. 108008, May 2021, doi: 10.1016/j.patcog.2021.108008.

[73] M. Diaz, M. A. Ferrer, G. S. Eskander, and R. Sabourin, “Generation of Duplicated Off-Line Signature Images for Verification Systems,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 5, pp. 951–964, May 2017, doi: 10.1109/TPAMI.2016.2560810.

[74] T. M. Maruyama, L. S. Oliveira, A. S. Britto, and R. Sabourin, “Intrapersonal Parameter Optimization for Offline Handwritten Signature Augmentation,” *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1335–1350, 2021, doi: 10.1109/TIFS.2020.3033442.

[75] J. Jiang, S. Lai, L. Jin, Y. Zhu, J. Zhang, and B. Chen, “Forgery-free signature verification with stroke-aware cycle-consistent generative adversarial network,” *Neurocomputing*, vol. 507, pp. 345–357, Oct. 2022, doi: 10.1016/j.neucom.2022.08.017.

[76] E. N. Zois, S. Said, D. Tsourounis, and A. Alexandridis, “Subscripto multiplex: A Riemannian symmetric positive definite strategy for offline signature verification,” *Pattern Recognition Letters*, vol. 167, pp. 67–74, Mar. 2023, doi: 10.1016/j.patrec.2023.02.002.

[77] P. Maergner *et al.*, “Combining graph edit distance and triplet networks for offline signature verification,” *Pattern Recognition Letters*, vol. 125, pp. 527–533, Jul. 2019, doi: 10.1016/j.patrec.2019.06.024.

[78] L. Liu, L. Huang, F. Yin, and Y. Chen, “Offline signature verification using a region based deep metric learning network,” *Pattern Recognition*, vol. 118, p. 108009, Oct. 2021, doi: 10.1016/j.patcog.2021.108009.

[79] Y. Zhu, S. Lai, Z. Li, and L. Jin, “Point-to-Set Similarity Based Deep Metric Learning for Offline Signature Verification,” in *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Sep. 2020, pp. 282–287. doi: 10.1109/ICFHR2020.2020.00059.

[80] A. Hamadene and Y. Chibani, “One-Class Writer-Independent Offline Signature Verification Using Feature Dissimilarity Thresholding,” *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 6, pp. 1226–1238, 2016, doi: 10.1109/TIFS.2016.2521611.

[81] T. Longjam, D. R. Kisku, and P. Gupta, “Writer independent handwritten signature verification on multi-scripted signatures using hybrid CNN-BiLSTM: A novel approach,” *Expert Systems with Applications*, vol. 214, p. 119111, Mar. 2023, doi: 10.1016/j.eswa.2022.119111.

[82] V. L. F. Souza, A. L. I. Oliveira, R. M. O. Cruz, and R. Sabourin, “A white-box analysis on the writer-independent dichotomy transformation applied to offline handwritten signature verification,” *Expert Systems with Applications*, vol. 154, p. 113397, Sep. 2020, doi: 10.1016/j.eswa.2020.113397.

[83] T. B. Viana, V. L. F. Souza, A. L. I. Oliveira, R. M. O. Cruz, and R. Sabourin, “Contrastive Learning of Handwritten Signature Representations for Writer-Independent Verification,” in *2022 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2022, pp. 01–09. doi: 10.1109/IJCNN55064.2022.9892428.

[84] T. B. Viana, V. L. F. Souza, A. L. I. Oliveira, R. M. O. Cruz, and R. Sabourin, “A multi-task approach for contrastive learning of handwritten signature feature representations,” *Expert Systems with Applications*, vol. 217, p. 119589, May 2023, doi: 10.1016/j.eswa.2023.119589.

[85] M. S. Hanif and M. Bilal, “A Metric Learning Approach for Offline Writer Independent Signature Verification,” *Pattern Recognit. Image Anal.*, vol. 30, no. 4, pp. 795–804, Oct. 2020, doi: 10.1134/S1054661820040173.

[86] H. Li, P. Wei, and P. Hu, “AVN: An Adversarial Variation Network Model for Handwritten Signature Verification,” *IEEE Transactions on Multimedia*, vol. 24, pp. 594–608, 2022, doi: 10.1109/TMM.2021.3056217.

[87] C. Li, F. Lin, Z. Wang, G. Yu, L. Yuan, and H. Wang, “DeepHSV: User-Independent Offline Signature Verification Using Two-Channel CNN,” in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sep. 2019, pp. 166–171. doi: 10.1109/ICDAR.2019.00035.

[88] P. Wei, H. Li, and P. Hu, “Inverse Discriminative Networks for Handwritten Signature Verification,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5764–5772.

[89] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust Face Recognition via Sparse Representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb. 2009, doi: 10.1109/TPAMI.2008.79.

[90] J. Yang, K. Yu, Y. Gong, and T. Huang, “Linear spatial pyramid matching using sparse coding for image classification,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 1794–1801. doi: 10.1109/CVPR.2009.5206757.

[91] E. N. Zois, D. Tsourounis, I. Theodorakopoulos, A. L. Kesidis, and G. Economou, “A Comprehensive Study of Sparse Representation Techniques for Offline Signature Verification,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 68–81, Jan. 2019, doi: 10.1109/TBIOM.2019.2897802.

[92] S. Leutenegger, M. Chli, and R. Y. Siegwart, “BRISK: Binary Robust invariant scalable keypoints,” in *2011 International Conference on Computer Vision*, Nov. 2011, pp. 2548–2555. doi: 10.1109/ICCV.2011.6126542.

[93] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, “Proximal methods for sparse hierarchical dictionary learning,” in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, in ICML’10. Madison, WI, USA: Omnipress, Jun. 2010, pp. 487–494.

[94] E. N. Zois, M. Papagiannopoulou, D. Tsourounis, and G. Economou, “Hierarchical Dictionary Learning and Sparse Coding for Static Signature Verification,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 432–442.

[95] D. Tsourounis, I. Theodorakopoulos, and E. N. Zois, “Handwritten Signature Verification via Deep Sparse Coding Architecture,” presented at the 2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), IEEE, 2018, pp. 1–5.

[96] A. Soleimani, K. Fouladi, and B. N. Araabi, “UTSig: A Persian offline signature dataset,” *IET Biometrics*, vol. 6, no. 1, pp. 1–8, 2016, doi: 10.1049/iet-bmt.2015.0058.

[97] D. L. Donoho, “For most large underdetermined systems of equations, the minimal ℓ1-norm near-solution approximates the sparsest near-solution,” *Communications on Pure and Applied Mathematics*, vol. 59, no. 7, pp. 907–934, 2006, doi: 10.1002/cpa.20131.

[98] D. Donoho, “Neighborly Polytopes And Sparse Solution Of Underdetermined Linear Equations,” 2005. Accessed: May 25, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Neighborly-Polytopes-And-Sparse-Solution-Of-Linear-Donoho/626703b4b5d8f2188ec53d82d8cb9e6868edc145

[99] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, “A Survey of Sparse Representation: Algorithms and Applications,” *IEEE Access*, vol. 3, pp. 490–530, 2015, doi: 10.1109/ACCESS.2015.2430359.

[100] S. Shariatmadari, S. Emadi, and Y. Akbari, “Patch-based offline signature verification using one-class hierarchical deep learning,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 22, no. 4, pp. 375–385, Dec. 2019, doi: 10.1007/s10032-019-00331-2.

[101] S. Masoudnia, O. Mersa, B. N. Araabi, A.-H. Vahabie, M. A. Sadeghi, and M. N. Ahmadabadi, “Multi-Representational Learning for Offline Signature Verification using Multi-Loss Snapshot Ensemble of CNNs,” *Expert Systems with Applications*, vol. 133, pp. 317–330, 2019.

[102] L. G. Hafemann, R. Sabourin, and L. S. Oliveira, “Learning features for offline handwritten signature verification using deep convolutional neural networks,” *Pattern Recognition*, vol. 70, pp. 163–176, 2017.

[103] L. G. Hafemann, L. S. Oliveira, and R. Sabourin, “Fixed-sized representation learning from offline handwritten signatures of different sizes,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 21, no. 3, pp. 219–232, 2018.

[104] E. N. Zois, I. Theodorakopoulos, D. Tsourounis, and G. Economou, “Parsimonious Coding and Verification of Offline Handwritten Signatures,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 134–143. Accessed: Sep. 18, 2020. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017_workshops/w6/html/Zois_Parsimonious_Coding_and_CVPR_2017_paper.html

[105] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded Up Robust Features,” in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 404–417. doi: 10.1007/11744023_32.

[106] D. Hutchison *et al.*, “BRIEF: Binary Robust Independent Elementary Features,” in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 778–792. doi: 10.1007/978-3-642-15561-1_56.

[107] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, Jun. 2005, pp. 886–893 vol. 1. doi: 10.1109/CVPR.2005.177.

[108] R. Arandjelovic and A. Zisserman, “All About VLAD,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 1578–1585. doi: 10.1109/CVPR.2013.207.

[109] J. Sivic and A. Zisserman, “Efficient Visual Search of Videos Cast as Text Retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 591–606, Apr. 2009, doi: 10.1109/TPAMI.2008.111.

[110] D. Kastaniotis, F. Fotopoulou, I. Theodorakopoulos, G. Economou, and S. Fotopoulos, “HEp-2 cell classification with Vector of Hierarchically Aggregated Residuals,” *Pattern Recognition*, vol. 65, pp. 47–57, May 2017, doi: 10.1016/j.patcog.2016.12.013.

[111] H. Jegou, M. Douze, C. Schmid, and P. Perez, “Aggregating local descriptors into a compact image representation,” IEEE, Jun. 2010, pp. 3304–3311. doi: 10.1109/CVPR.2010.5540039.

[112] H. Jegou, F. Perronnin, M. Douze, J. S&#x00E1;nchez, P. Perez, and C. Schmid, “Aggregating Local Image Descriptors into Compact Codes,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012, doi: 10.1109/TPAMI.2011.235.

[113] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, in NIPS’12. USA: Curran Associates Inc., 2012, pp. 1097–1105. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999134.2999257

[114] “Multi-scale Orderless Pooling of Deep Convolutional Activation Features | SpringerLink.” https://link.springer.com/chapter/10.1007/978-3-319-10584-0_26 (accessed Jan. 14, 2021).

[115] C. Liu, J. Yuen, and A. Torralba, “Nonparametric scene parsing: Label transfer via dense scene alignment,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 1972–1979. doi: 10.1109/CVPR.2009.5206536.

[116] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

[117] Y. LeCun *et al.*, “Comparison of Learning Algorithms for Handwritten Digit Recognition,” in *INTERNATIONAL CONFERENCE ON ARTIFICIAL NEURAL NETWORKS*, 1995, pp. 53–60.

[118] P. Foggia, G. Percannella, A. Saggese, and M. Vento, “Pattern recognition in stained HEp-2 cells: Where are we now?,” *Pattern Recognition*, vol. 47, 2014, doi: 10.1016/j.patcog.2014.01.010.

[119] S. Liu, M. Li, Z. Zhang, B. Xiao, and T. S. Durrani, “Multi-Evidence and Multi-Modal Fusion Network for Ground-Based Cloud Recognition,” *Remote Sensing*, vol. 12, no. 3, Art. no. 3, Jan. 2020, doi: 10.3390/rs12030464.

[120] S. Liu, M. Li, Z. Zhang, X. Cao, and T. S. Durrani, “Ground-Based Cloud Classification Using Task-Based Graph Convolutional Network,” *Geophysical Research Letters*, vol. 47, no. 5, p. e2020GL087338, 2020, doi: https://doi.org/10.1029/2020GL087338.

[121] J. S. Chung and A. Zisserman, “Lip reading in the wild,” in *Asian Conference on Computer Vision*, Springer, 2016, pp. 87–103.

[122] L. Zheng, Y. Yang, and Q. Tian, “SIFT Meets CNN: A Decade Survey of Instance Retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1224–1244, May 2018, doi: 10.1109/TPAMI.2017.2709749.

[123] H. Wang and S. Hou, “Facial Expression Recognition based on The Fusion of CNN and SIFT Features,” in *2020 IEEE 10th International Conference on Electronics Information and*

Emergency Communication (ICEIEC), Jul. 2020, pp. 190–194. doi: 10.1109/ICEIEC49280.2020.9152361.

[124] W. Lin, K. Hasenstab, G. Moura Cunha, and A. Schwartzman, “Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment,” *Sci Rep*, vol. 10, no. 1, Art. no. 1, Nov. 2020, doi: 10.1038/s41598-020-77264-y.

[125] A. Tripathi, T. V. A. Kumar, T. K. Dhansetty, and J. S. Kumar, “Real Time Object Detection using CNN,” *International Journal of Engineering & Technology*, vol. 7, no. 2.24, Art. no. 2.24, Apr. 2018, doi: 10.14419/ijet.v7i2.24.11994.

[126] A. Dudhal, H. Mathkar, A. Jain, O. Kadam, and M. Shirole, “Hybrid SIFT Feature Extraction Approach for Indian Sign Language Recognition System Based on CNN,” in *Proceedings of the International Conference on ISMAC in Computational Vision and Bio-Engineering 2018 (ISMAC-CVB)*, D. Pandian, X. Fernando, Z. Baig, and F. Shi, Eds., in Lecture Notes in Computational Vision and Biomechanics. Cham: Springer International Publishing, 2019, pp. 727–738. doi: 10.1007/978-3-030-00665-5_72.

[127] T. Connie, M. Al-Shabi, W. P. Cheah, and M. Goh, “Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator,” in *Multi-disciplinary Trends in Artificial Intelligence*, S. Phon-Amnuaisuk, S.-P. Ang, and S.-Y. Lee, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 139–149. doi: 10.1007/978-3-319-69456-6_12.

[128] A. Kumar, N. Jain, C. Singh, and S. Tripathi, “Exploiting SIFT Descriptor for Rotation Invariant Convolutional Neural Network,” in *2018 15th IEEE India Council International Conference (INDICON)*, Dec. 2018, pp. 1–5. doi: 10.1109/INDICON45594.2018.8987153.

[129] C. Weiyue, J. Geng, and K. Lin, “Facial Expression Recognition with Small Samples Under Convolutional Neural Network,” in *6GN for Future Wireless Networks*, S. Shi, R. Ma, and W. Lu, Eds., in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Cham: Springer International Publishing, 2022, pp. 383–396. doi: 10.1007/978-3-031-04245-4_34.

[130] M. K. Vidhyalakshmi, E. Poovammal, V. Bhaskar, and J. Sathyanarayanan, “Novel Similarity Metric Learning Using Deep Learning and Root SIFT for Person Re-identification,” *Wireless Pers Commun*, vol. 117, no. 3, pp. 1835–1851, Apr. 2021, doi: 10.1007/s11277-020-07948-1.

[131] Q. Zhao *et al.*, “A CNN-SIFT Hybrid Pedestrian Navigation Method Based on First-Person Vision,” *Remote Sensing*, vol. 10, no. 8, Art. no. 8, Aug. 2018, doi: 10.3390/rs10081229.

[132] S. K. Park, J. H. Chung, T. K. Kang, and M. T. Lim, “Binary dense sift flow based two stream CNN for human action recognition,” *Multimed Tools Appl*, vol. 80, no. 28, pp. 35697–35720, Nov. 2021, doi: 10.1007/s11042-021-10795-2.

[133] D. Varga, “No-Reference Quality Assessment of Authentically Distorted Images Based on Local and Global Features,” *Journal of Imaging*, vol. 8, no. 6, Art. no. 6, Jun. 2022, doi: 10.3390/jimaging8060173.

[134] P. K. R. Yelampalli, J. Nayak, and V. H. Gaidhane, “Daubechies wavelet-based local feature descriptor for multimodal medical image registration,” *IET Image Processing*, vol. 12, no. 10, pp. 1692–1702, Apr. 2018, doi: 10.1049/iet-ipr.2017.1305.

[135] S. Luan *et al.*, “Gabor Convolutional Networks,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2018, pp. 1254–1262. doi: 10.1109/WACV.2018.00142.

[136] E. Oyallon, E. Belilovsky, and S. Zagoruyko, “Scaling the Scattering Transform: Deep Hybrid Networks,” *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5619–5628, Oct. 2017, doi: 10.1109/ICCV.2017.599.

[137] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, “PCANet: A Simple Deep Learning Baseline for Image Classification?,” *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017–5032, Dec. 2015, doi: 10.1109/TIP.2015.2475625.

[138] R. Zeng, J. Wu, L. Senhadji, and H. Shu, “Tensor object classification via multilinear discriminant analysis network,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 1971–1975. doi: 10.1109/ICASSP.2015.7178315.

[139] Y. Gan, J. Liu, J. Dong, and G. Zhong, “A PCA-Based Convolutional Network.” arXiv, May 14, 2015. doi: 10.48550/arXiv.1505.03703.

[140] D. Wu, J. Wu, R. Zeng, L. Jiang, L. Senhadji, and H. Shu, “Kernel principal component analysis network for image classification.” arXiv, Dec. 20, 2015. doi: 10.48550/arXiv.1512.06337.

[141] Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, “Oriented Response Networks,” IEEE, Jul. 2017, pp. 4961–4970. doi: 10.1109/CVPR.2017.527.

[142] M. Jaderberg, K. Simonyan, A. Zisserman, and koray kavukcuoglu, “Spatial Transformer Networks,” in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., Curran Associates, Inc., 2015, pp. 2017–2025. [Online]. Available: http://papers.nips.cc/paper/5854-spatial-transformer-networks.pdf

[143] F. Guc and Y. Chen, “Sensor Fault Diagnostics Using Physics-Informed Transfer Learning Framework,” *Sensors*, vol. 22, no. 8, Art. no. 8, Jan. 2022, doi: 10.3390/s22082913.

[144] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, “Physics-informed machine learning,” *Nat Rev Phys*, vol. 3, no. 6, Art. no. 6, Jun. 2021, doi: 10.1038/s42254-021-00314-5.

[145] G.-S. Xie, X.-Y. Zhang, S. Yan, and C.-L. Liu, “Hybrid CNN and Dictionary-Based Models for Scene Recognition and Domain Adaptation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 6, pp. 1263–1274, Jun. 2017, doi: 10.1109/TCSVT.2015.2511543.

[146] F. Perronnin and D. Larlus, “Fisher vectors meet Neural Networks: A hybrid classification architecture,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 3743–3752. doi: 10.1109/CVPR.2015.7298998.

[147] M. Xi, L. Chen, D. Polajnar, and W. Tong, “Local binary pattern network: A deep learning approach for face recognition,” in *2016 IEEE International Conference on Image Processing (ICIP)*, Sep. 2016, pp. 3224–3228. doi: 10.1109/ICIP.2016.7532955.

[148] S. Chen, “LBPNet: Inserting Local Binary Patterns into Neural Networks to Enhance Manipulation Invariance of Fake Face Detection,” in *2021 International Conference on Digital Society and Intelligent Systems (DSInS)*, Dec. 2021, pp. 212–217. doi: 10.1109/DSInS54396.2021.9670608.

[149] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, Jun. 2006, pp. 2169–2178. doi: 10.1109/CVPR.2006.68.

[150] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, “HEp-2 cells classification via sparse representation of textural features fused into dissimilarity space,” *Pattern Recognition*, vol. 47, no. 7, pp. 2367–2378, 2014, doi: https://doi.org/10.1016/j.patcog.2013.09.026.

[151] S. Kornblith, J. Shlens, and Q. V. Le, “Do Better ImageNet Models Transfer Better?,” *arXiv:1805.08974 [cs, stat]*, May 2018, Accessed: Jun. 02, 2018. [Online]. Available: http://arxiv.org/abs/1805.08974

[152] I. Nigam, S. Agrawal, R. Singh, and M. Vatsa, “Revisiting HEp-2 Cell Image Classification,” *IEEE Access*, vol. 3, pp. 3102–3113, 2015, doi: 10.1109/ACCESS.2015.2504125.

[153] P. Agrawal, M. Vatsa, and R. Singh, “HEp-2 Cell Image Classification: A Comparative Analysis,” in *Machine Learning in Medical Imaging*, G. Wu, D. Zhang, D. Shen, P. Yan, K. Suzuki, and F. Wang, Eds., Springer International Publishing, 2013, pp. 195–202.

[154] S. Ensafi, S. Lu, A. A. Kassim, and C. L. Tan, “A Bag of Words Based Approach for Classification of HEp-2 Cell Images,” in *Proceedings of the 2014 1st Workshop on Pattern Recognition Techniques for Indirect Immunofluorescence Images*, in I3AWORKSHOP ’14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 29–32. doi: 10.1109/I3A.Workshop.2014.11.

[155] S. Ensafi, S. Lu, A. A. Kassim, and C. L. Tan, “Accurate HEp-2 cell classification based on sparse bag of words coding,” *Computerized Medical Imaging and Graphics*, vol. 57, pp. 40–49, 2017, doi: https://doi.org/10.1016/j.compmedimag.2016.08.002.

[156] G. Csurka, “Visual categorization with bags of keypoints,” *undefined*, 2004, Accessed: Jun. 24, 2022. [Online]. Available: https://www.semanticscholar.org/paper/Visual-categorization-with-bags-of-keypoints-Csurka/b91180d8853d00e8f2df7ee3532e07d3d0cce2af

[157] 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*, vol. 24, no. 7, pp. 971–987, Jul. 2002, doi: 10.1109/TPAMI.2002.1017623.

[158] Z. Guo, L. Zhang, and D. Zhang, “A Completed Modeling of Local Binary Pattern Operator for Texture Classification,” *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1657–1663, Jun. 2010, doi: 10.1109/TIP.2010.2044957.

[159] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *CoRR*, vol. abs/1409.1556, 2014.

[160] J. Zhang, P. Liu, F. Zhang, and Q. Song, “CloudNet: Ground-Based Cloud Classification With Deep Convolutional Neural Network,” *Geophysical Research Letters*, vol. 45, no. 16, pp. 8665–8672, 2018, doi: 10.1029/2018GL077787.

[161] D. Tsourounis, D. Kastaniotis, C. Theoharatos, A. Kazantzidis, and G. Economou, “SIFT-CNN: When Convolutional Neural Networks Meet Dense SIFT Descriptors for Image and Sequence Classification,” *Journal of Imaging*, vol. 8, no. 10, Art. no. 10, Oct. 2022, doi: 10.3390/jimaging8100256.

[162] M. Li, S. Liu, and Z. Zhang, “Dual Guided Loss for Ground-Based Cloud Classification in Weather Station Networks,” *IEEE Access*, vol. 7, pp. 63081–63088, 2019, doi: 10.1109/ACCESS.2019.2916905.

[163] S. Liu, L. Duan, Z. Zhang, and X. Cao, “Hierarchical Multimodal Fusion for Ground-Based Cloud Classification in Weather Station Networks,” *IEEE Access*, vol. 7, pp. 85688–85695, 2019, doi: 10.1109/ACCESS.2019.2926092.

[164] W. Zhu *et al.*, “Classification of Ground-Based Cloud Images by Improved Combined Convolutional Network,” *Applied Sciences*, vol. 12, no. 3, Art. no. 3, Jan. 2022, doi: 10.3390/app12031570.

[165] S. Agrawal, V. R. Omprakash, and Ranvijay, “Lip reading techniques: A survey,” in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, Jul. 2016, pp. 753–757. doi: 10.1109/ICATCCT.2016.7912100.

[166] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, “Temporal Convolutional Networks for Action Segmentation and Detection,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 156–165. Accessed: Sep. 14, 2020. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Lea_Temporal_Convolutional_Networks_CVPR_2017_paper.html

[167] Y. Jining, M. Lin, L. Wang, R. Rajiv, and A. Y. Zomaya, “Temporal Convolutional Networks for the Advance Prediction of ENSO,” *Scientific Reports (Nature Publisher Group)*, vol. 10, no. 1, 2020.

[168] B. Martinez, P. Ma, S. Petridis, and M. Pantic, “Lipreading Using Temporal Convolutional Networks,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6319–6323. doi: 10.1109/ICASSP40776.2020.9053841.

[169] D. Kastaniotis, D. Tsourounis, and S. Fotopoulos, “Lip Reading modeling with Temporal Convolutional Networks for medical support applications,” *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2020, doi: 10.1109/CISP-BMEI51763.2020.9263634.

[170] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip Reading Sentences in the Wild,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 3444–3453. doi: 10.1109/CVPR.2017.367.

[171] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, “End-to-End Audiovisual Speech Recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 6548–6552. doi: 10.1109/ICASSP.2018.8461326.

[172] T. Stafylakis and G. Tzimiropoulos, “Combining Residual Networks with LSTMs for Lipreading,” in *Interspeech 2017*, ISCA, Aug. 2017, pp. 3652–3656. doi: 10.21437/Interspeech.2017-85.

[173] S. Cheng *et al.*, “Towards Pose-Invariant Lip-Reading,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 4357–4361. doi: 10.1109/ICASSP40776.2020.9054384.

[174] C. Wang, “Multi-Grained Spatio-temporal Modeling for Lip-reading.,” in *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, 2019, p. 276. [Online]. Available: https://bmvc2019.org/wp-content/uploads/papers/1211-paper.pdf

[175] L. Courtney and R. Sreenivas, “Learning from Videos with Deep Convolutional LSTM Networks,” *arXiv:1904.04817 [cs]*, Apr. 2019, Accessed: Apr. 05, 2021. [Online]. Available: http://arxiv.org/abs/1904.04817

[176] M. Luo, S. Yang, S. Shan, and X. Chen, “Pseudo-Convolutional Policy Gradient for Sequence-to-Sequence Lip-Reading,” *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020, doi: 10.1109/FG47880.2020.00010.

[177] X. Weng and K. Kitani, “Learning Spatio-Temporal Features with Two-Stream Deep 3D CNNs for Lipreading.,” in *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, 2019, p. 269. [Online]. Available: https://bmvc2019.org/wp-content/uploads/papers/0016-paper.pdf

[178] J. Xiao, S. Yang, Y.-H. Zhang, S. Shan, and X. Chen, “Deformation Flow Based Two-Stream Network for Lip Reading,” *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020, doi: 10.1109/FG47880.2020.00132.

[179] X. Zhao, S. Yang, S. Shan, and X. Chen, “Mutual Information Maximization for Effective Lip Reading,” *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020, doi: 10.1109/FG47880.2020.00133.

[180] Y.-H. Zhang, S. Yang, J. Xiao, S. Shan, and X. Chen, “Can We Read Speech Beyond the Lips? Rethinking RoI Selection for Deep Visual Speech Recognition,” *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020, doi: 10.1109/FG47880.2020.00134.

[181] D. Feng, S. Yang, S. Shan, and X. Chen, “Learn an Effective Lip Reading Model without Pains.,” *CoRR*, vol. abs/2011.07557, 2020, [Online]. Available: https://arxiv.org/abs/2011.07557

[182] X. Pan, P. Chen, Y. Gong, H. Zhou, X. Wang, and Z. Lin, “Leveraging Unimodal Self-Supervised Learning for Multimodal Audio-Visual Speech Recognition,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 4491–4503. doi: 10.18653/v1/2022.acl-long.308.

[183] M. Kim, J. Hong, S. J. Park, and Y. M. Ro, “Multi-Modality Associative Bridging Through Memory: Speech Sound Recollected From Face Video,” presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 296–306. Accessed: Sep. 02, 2022.

[Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Kim_Multi-Modality_Associative_Bridging_Through_Memory_Speech_Sound_Recollected_From_Face_ICCV_2021_paper.html

[184] D. Tsourounis, D. Kastaniotis, and S. Fotopoulos, “Lip Reading by Alternating between Spatiotemporal and Spatial Convolutions,” *Journal of Imaging*, vol. 7, no. 5, Art. no. 5, May 2021, doi: 10.3390/jimaging7050091.

[185] M. Kim, J. H. Yeo, and Y. M. Ro, “Distinguishing Homophenes Using Multi-Head Visual-Audio Memory for Lip Reading.” arXiv, Apr. 04, 2022. doi: 10.48550/arXiv.2204.01725.

[186] A. Koumparoulis and G. Potamianos, “Accurate and Resource-Efficient Lipreading with Efficientnetv2 and Transformers,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 8467–8471. doi: 10.1109/ICASSP43922.2022.9747729.

[187] S. J. Raudys and A. K. Jain, “Small sample size effects in statistical pattern recognition: recommendations for practitioners,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252–264, Mar. 1991, doi: 10.1109/34.75512.

[188] M. Diaz, M. A. Ferrer, G. S. Eskander, and R. Sabourin, “Generation of Duplicated Off-Line Signature Images for Verification Systems,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 5, pp. 951–964, May 2017, doi: 10.1109/TPAMI.2016.2560810.

[189] M. A. Ferrer, J. F. Vargas, A. Morales, and A. Ordonez, “Robustness of Offline Signature Verification Based on Gray Level Features,” *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 966–977, Jun. 2012, doi: 10.1109/TIFS.2012.2190281.

[190] J. F. Vargas, M. A. Ferrer, C. M. Travieso, and J. B. Alonso, “Off-line Handwritten Signature GPDS-960 Corpus,” presented at the Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, Sep. 2007, pp. 764–768. doi: 10.1109/ICDAR.2007.4377018.

[191] M. Parmar, N. Puranik, D. Joshi, and S. Malpani, “Image Processing Based Signature Duplication and its Verification.” Rochester, NY, Apr. 30, 2020. doi: 10.2139/ssrn.3645426.

[192] A. Natarajan, B. S. Babu, and X.-Z. Gao, “Signature warping and greedy approach based offline signature verification,” *Int. j. inf. tecnol.*, vol. 13, no. 4, pp. 1279–1290, Aug. 2021, doi: 10.1007/s41870-021-00689-9.

[193] M. M. Yapıcı, A. Tekerek, and N. Topaloğlu, “Deep learning-based data augmentation method and signature verification system for offline handwritten signature,” *Pattern Anal Applic*, vol. 24, no. 1, pp. 165–179, Feb. 2021, doi: 10.1007/s10044-020-00912-6.

[194] D. C. Yonekura and E. B. Guedes, "Offline Handwritten Signature Authentication with Conditional Deep Convolutional Generative Adversarial Networks," in *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, SBC, Nov. 2021, pp. 482–491. doi: 10.5753/eniac.2021.18277.

[195] M. Diaz, M. A. Ferrer, D. Impedovo, M. I. Malik, G. Pirlo, and R. Plamondon, "A Perspective Analysis of Handwritten Signature Technology," *ACM Comput. Surv.*, vol. 51, no. 6, p. 117:1-117:39, Jan. 2019, doi: 10.1145/3274658.

[196] D. Bertolini, L. S. Oliveira, E. Justino, and R. Sabourin, "Reducing forgeries in writer-independent off-line signature verification through ensemble of classifiers," *Pattern Recognition*, vol. 43, no. 1, pp. 387–396, Jan. 2010, doi: 10.1016/j.patcog.2009.05.009.

[197] J.-P. Drouhard, R. Sabourin, and M. Godbout, "A neural network approach to off-line signature verification using directional PDF," *Pattern Recognition*, vol. 29, no. 3, pp. 415–424, Mar. 1996, doi: 10.1016/0031-3203(95)00092-5.

[198] J. Fierrez-Aguilar, N. Alonso-Hermira, G. Moreno-Marquez, and J. Ortega-Garcia, "An Off-line Signature Verification System Based on Fusion of Local and Global Information," in *Biometric Authentication*, D. Maltoni and A. K. Jain, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2004, pp. 295–306. doi: 10.1007/978-3-540-25976-3_27.

[199] R. Ghosh, "A Recurrent Neural Network based deep learning model for offline signature verification and recognition system," *Expert Systems with Applications*, p. 114249, Nov. 2020, doi: 10.1016/j.eswa.2020.114249.

[200] J. Ji, C. Chen, and X. Chen, "Off-Line Chinese Signature Verification: Using Weighting Factor on Similarity Computation," in *2010 2nd International Conference on E-business and Information System Security*, May 2010, pp. 1–4. doi: 10.1109/EBISS.2010.5473588.

[201] A. Nordgaard and B. Rasmusson, "The likelihood ratio as value of evidence—more than a question of numbers," *Law, Probability and Risk*, vol. 11, no. 4, pp. 303–315, Dec. 2012, doi: 10.1093/lpr/mgs019.

[202] D. Rivard, E. Granger, and R. Sabourin, "Multi-feature extraction and selection in writer-independent off-line signature verification," *IJDAR*, vol. 16, no. 1, pp. 83–103, Mar. 2013, doi: 10.1007/s10032-011-0180-6.

[203] B. Schafer and S. Viriri, "An off-line signature verification system," in *2009 IEEE International Conference on Signal and Image Processing Applications*, Nov. 2009, pp. 95–100. doi: 10.1109/ICSIPA.2009.5478727.

[204] T. Steinherz, D. Doermann, E. Rivlin, and N. Intrator, “Offline Loop Investigation for Handwriting Analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 193–209, Feb. 2009, doi: 10.1109/TPAMI.2008.68.

[205] P. S. Deng, H.-Y. M. Liao, C. W. Ho, and H.-R. Tyan, “Wavelet-Based Off-Line Handwritten Signature Verification,” *Computer Vision and Image Understanding*, vol. 76, no. 3, pp. 173–190, Dec. 1999, doi: 10.1006/cviu.1999.0799.

[206] A. Foroozandeh, Y. Akbari, M. J. Jalili, and J. Sadri, “Persian Signature Verification Based on Fractal Dimension Using Testing Hypothesis,” in *2012 International Conference on Frontiers in Handwriting Recognition*, Sep. 2012, pp. 313–318. doi: 10.1109/ICFHR.2012.254.

[207] V. Kiani, R. Pourreza, and H. R. Pourreza, “Offline signature verification using local radon transform and support vector machines,” *International Journal of Image Processing*, vol. 3, no. 5, pp. 184–194, 2009.

[208] A. Dutta, U. Pal, and J. Lladós, “Compact correlated features for writer independent signature verification,” in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec. 2016, pp. 3422–3427. doi: 10.1109/ICPR.2016.7900163.

[209] J. Hu and Y. Chen, “Offline Signature Verification Using Real Adaboost Classifier Combination of Pseudo-dynamic Features,” in *2013 12th International Conference on Document Analysis and Recognition*, Aug. 2013, pp. 1345–1349. doi: 10.1109/ICDAR.2013.272.

[210] M. I. Malik, M. Liwicki, A. Dengel, S. Uchida, and V. Frinken, “Automatic Signature Stability Analysis and Verification Using Local Features,” in *2014 14th International Conference on Frontiers in Handwriting Recognition*, Sep. 2014, pp. 621–626. doi: 10.1109/ICFHR.2014.109.

[211] M. I. Malik, S. Ahmed, M. Liwicki, and A. Dengel, “FREAK for Real Time Forensic Signature Verification,” in *2013 12th International Conference on Document Analysis and Recognition*, Aug. 2013, pp. 971–975. doi: 10.1109/ICDAR.2013.196.

[212] M. Okawa, “From BoVW to VLAD with KAZE features: Offline signature verification considering cognitive processes of forensic experts,” *Pattern Recognition Letters*, vol. 113, pp. 75–82, Oct. 2018, doi: 10.1016/j.patrec.2018.05.019.

[213] J. Ruiz-del-Solar, C. Devia, P. Loncomilla, and F. Concha, “Offline Signature Verification Using Local Interest Points and Descriptors,” in *Progress in Pattern Recognition, Image Analysis and Applications*, J. Ruiz-Shulcloper and W. G. Kropatsch, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008, pp. 22–29. doi: 10.1007/978-3-540-85920-8_3.

[214] Y. Serdouk, H. Nemmour, and Y. Chibani, "Topological and textural features for off-line signature verification based on artificial immune algorithm," presented at the Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of, Aug. 2014, pp. 118–122. doi: 10.1109/SOCPAR.2014.7007991.

[215] M. B. Yilmaz, B. Yanikoglu, C. Tirkaz, and A. Kholmatov, "Offline signature verification using classifier combination of HOG and LBP features," in *2011 International Joint Conference on Biometrics (IJCB)*, Oct. 2011, pp. 1–7. doi: 10.1109/IJCB.2011.6117473.

[216] E. N. Zois, I. Theodorakopoulos, D. Tsourounis, and G. Economou, "Parsimonious Coding and Verification of Offline Handwritten Signatures," presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Jul. 2017, pp. 636–645. doi: 10.1109/CVPRW.2017.92.

[217] E. N. Zois, I. Theodorakopoulos, and G. Economou, "Offline Handwritten Signature Modeling and Verification Based on Archetypal Analysis," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5514–5523. Accessed: Nov. 10, 2020. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Zois_Offline_Handwritten_Signature_ICCV_2017_paper.html

[218] D. Gumusbas and T. Yildirim, "Offline Signature Identification and Verification Using Capsule Network," in *2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)*, Jul. 2019, pp. 1–5. doi: 10.1109/INISTA.2019.8778228.

[219] M. B. Yılmaz and K. Öztürk, "Recurrent Binary Patterns and CNNs for Offline Signature Verification," in *Proceedings of the Future Technologies Conference (FTC) 2019*, K. Arai, R. Bhatia, and S. Kapoor, Eds., in Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2020, pp. 417–434. doi: 10.1007/978-3-030-32523-7_29.

[220] B. Ribeiro, I. Gonçalves, S. Santos, and A. Kovacec, "Deep Learning Networks for Off-Line Handwritten Signature Recognition," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, C. San Martin and S.-W. Kim, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2011, pp. 523–532. doi: 10.1007/978-3-642-25085-9_62.

[221] Khalajzadeh Hurieh, M. Mansouri, and M. Teshnehlab, "Persian Signature Verification using Convolutional Neural Networks," *International Journal of Engineering Research and Technology (IJERT)*, vol. 1, no. 2, pp. 7–12, 2012.

[222] Z. Zhang, X. Liu, and Y. Cui, "Multi-phase Offline Signature Verification System Using Deep Convolutional Generative Adversarial Networks," in *2016 9th International Symposium on Computational Intelligence and Design (ISCID)*, Dec. 2016, pp. 103–107. doi: 10.1109/ISCID.2016.2033.

[223] S. Dey, A. Dutta, J. I. Toledo, S. K. Ghosh, J. Lladós, and U. Pal, "Signet: Convolutional siamese network for writer independent offline signature verification," *arXiv preprint arXiv:1707.02131*, 2017.

[224] L. G. Hafemann, R. Sabourin, and L. S. Oliveira, "Writer-independent feature learning for offline signature verification using deep convolutional neural networks," presented at the 2016 International Joint Conference on Neural Networks (IJCNN), IEEE, 2016, pp. 2576–2583.

[225] L. G. Hafemann, R. Sabourin, and L. S. Oliveira, "Characterizing and evaluating adversarial examples for Offline Handwritten Signature Verification," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 8, pp. 2153–2166, 2019.

[226] L. G. Hafemann, R. Sabourin, and L. Oliveira, "Meta-Learning for Fast Classifier Adaptation to New Users of Signature Verification Systems," *IEEE Transactions on Information Forensics and Security*, 2020, doi: 10.1109/TIFS.2019.2949425.

[227] T. M. Maruyama, L. S. Oliveira, A. S. Britto Jr, and R. Sabourin, "Intrapersonal Parameter Optimization for Offline Handwritten Signature Augmentation," *arXiv:2010.06663 [cs]*, Oct. 2020, Accessed: Nov. 02, 2020. [Online]. Available: http://arxiv.org/abs/2010.06663

[228] M. B. Yilmaz and K. Öztürk, "Hybrid User-Independent and User-Dependent Offline Signature Verification with a Two-Channel CNN," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2018, pp. 639–6398. doi: 10.1109/CVPRW.2018.00094.

[229] O. Mersa, F. Etaati, S. Masoudnia, and B. Araabi, "Learning Representations from Persian Handwriting for Offline Signature Verification, a Deep Transfer Learning Approach," *2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, 2019, doi: 10.1109/PRIA.2019.8785979.

[230] T. Younesian, S. Masoudnia, R. Hosseini, and B. N. Araabi, "Active Transfer Learning for Persian Offline Signature Verification," *arXiv preprint arXiv:1903.06255*, 2019.

[231] A. Bellet, A. Habrard, and M. Sebban, "A Survey on Metric Learning for Feature Vectors and Structured Data," *arXiv:1306.6709 [cs, stat]*, Feb. 2014, Accessed: Dec. 17, 2020. [Online]. Available: http://arxiv.org/abs/1306.6709

[232] H. Rantzsch, H. Yang, and C. Meinel, “Signature embedding: Writer independent offline signature verification with deep metric learning,” presented at the International symposium on visual computing, Springer, 2016, pp. 616–625.

[233] A. Soleimani, B. N. Araabi, and K. Fouladi, “Deep multitask metric learning for offline signature verification,” *Pattern Recognition Letters*, vol. 80, pp. 84–90, 2016.

[234] J. Chapran, “Biometric writer identification: feature analysis and classification,” *Int. J. Patt. Recogn. Artif. Intell.*, vol. 20, no. 04, pp. 483–503, Jun. 2006, doi: 10.1142/S0218001406004831.

[235] F. Kleber, S. Fiel, M. Diem, and R. Sablatnig, “CVL-DataBase: An Off-Line Database for Writer Retrieval, Writer Identification and Word Spotting,” in *2013 12th International Conference on Document Analysis and Recognition*, Aug. 2013, pp. 560–564. doi: 10.1109/ICDAR.2013.117.

[236] M. Sharif, M. A. Khan, M. Faisal, M. Yasmin, and S. L. Fernandes, “A Framework for Offline Signature Verification System: Best Features Selection Approach,” *Pattern Recognition Letters*, 2018, doi: 10.1016/j.patrec.2018.01.021.

[237] M. R. Pourshahabi, M. H. Sigari, and H. R. Pourreza, “Offline Handwritten Signature Identification and Verification Using Contourlet Transform,” in *2009 International Conference of Soft Computing and Pattern Recognition*, Dec. 2009, pp. 670–673. doi: 10.1109/SoCPaR.2009.132.

[238] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, in ICML’10. Madison, WI, USA: Omnipress, Jun. 2010, pp. 807–814.

[239] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” *Proceedings of the International Conference on Machine Learning*, vol. 1, 2020, Accessed: Oct. 08, 2020. [Online]. Available: https://proceedings.icml.cc/paper/2020/hash/36452e720502e4da486d2f9f6b48a7bb

[240] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum Contrast for Unsupervised Visual Representation Learning,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738. Accessed: Oct. 08, 2020. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/He_Momentum_Contrast_for_Unsupervised_Visual_Representation_Learning_CVPR_2020_paper.html

[241] I. Misra and L. van der Maaten, “Self-Supervised Learning of Pretext-Invariant Representations,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6707–6717. Accessed: Oct. 08, 2020. [Online]. Available:

https://openaccess.thecvf.com/content_CVPR_2020/html/Misra_Self-Supervised_Learning_of_Pretext-Invariant_Representations_CVPR_2020_paper.html

[242] J. Wang *et al.*, “Learning Fine-Grained Image Similarity with Deep Ranking,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1386–1393. doi: 10.1109/CVPR.2014.180.

[243] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv:1412.6980 [cs]*, Jan. 2017, Accessed: Oct. 09, 2020. [Online]. Available: http://arxiv.org/abs/1412.6980

[244] R. Sabourin, R. Plamondon, and G. Lorette, “Off-line Identification With Handwritten Signature Images: Survey and Perspectives,” in *Structured Document Image Analysis*, H. S. Baird, H. Bunke, and K. Yamamoto, Eds., Berlin, Heidelberg: Springer, 1992, pp. 219–234. doi: 10.1007/978-3-642-77281-8_10.

[245] J. Galbally, M. Gomez-Barrero, and A. Ross, “Accuracy evaluation of handwritten signature verification: Rethinking the random-skilled forgeries dichotomy,” in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, Oct. 2017, pp. 302–310. doi: 10.1109/BTAS.2017.8272711.

[246] M. Blumenstein, M. A. Ferrer, and J. F. Vargas, “The 4NSigComp2010 Off-line Signature Verification Competition: Scenario 2,” in *2010 12th International Conference on Frontiers in Handwriting Recognition*, Nov. 2010, pp. 721–726. doi: 10.1109/ICFHR.2010.117.

[247] K. He, X. Zhang, S. Ren, and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1026–1034. doi: 10.1109/ICCV.2015.123.

[248] K. Stapor, P. Ksieniewicz, S. García, and M. Woźniak, “How to design the fair experimental classifier evaluation,” *Applied Soft Computing*, vol. 104, p. 107219, Jun. 2021, doi: 10.1016/j.asoc.2021.107219.

[249] R. V. Hogg and J. Ledolter, *Engineering statistics*. Macmillan Publishing Company, 1987.

[250] M. M. Kumar and N. B. Puhan, “Off-line signature verification: upper and lower envelope shape analysis using chord moments,” *IET Biometrics*, vol. 3, no. 4, pp. 347–354, 2014, doi: 10.1049/iet-bmt.2014.0024.

[251] A. K. Bhunia, A. Alaei, and P. P. Roy, “Signature verification approach using fusion of hybrid texture features,” *Neural Comput & Applic*, vol. 31, no. 12, pp. 8737–8748, Dec. 2019, doi: 10.1007/s00521-019-04220-x.

[252] J. Wen, B. Fang, Y. Y. Tang, and T. Zhang, “Model-based signature verification with rotation invariant features,” *Pattern Recognition*, vol. 42, no. 7, pp. 1458–1466, Jul. 2009, doi: 10.1016/j.patcog.2008.10.006.

[253] Y. Serdouk, H. Nemmour, and Y. Chibani, “Handwritten signature verification using the quad-tree histogram of templates and a Support Vector-based artificial immune classification,” *Image and Vision Computing*, vol. 66, pp. 26–35, Oct. 2017, doi: 10.1016/j.imavis.2017.08.004.

[254] M. Okawa, “Synergy of foreground–background images for feature extraction: Offline signature verification using Fisher vector with fused KAZE features,” *Pattern Recognition*, vol. 79, pp. 480–489, Jul. 2018, doi: 10.1016/j.patcog.2018.02.027.

[255] M. A. Ferrer, J. B. Alonso, and C. M. Travieso, “Offline geometric parameters for automatic signature verification using fixed-point arithmetic,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 993–997, Jun. 2005, doi: 10.1109/TPAMI.2005.125.

[256] V. Nguyen, M. Blumenstein, and G. Leedham, “Global Features for the Off-Line Signature Verification Problem,” in *2009 10th International Conference on Document Analysis and Recognition*, Jul. 2009, pp. 1300–1304. doi: 10.1109/ICDAR.2009.123.

[257] M. B. Yılmaz and B. Yanıkoğlu, “Score level fusion of classifiers in off-line signature verification,” *Information Fusion*, vol. 32, pp. 109–119, Nov. 2016, doi: 10.1016/j.inffus.2016.02.003.

[258] M. Parodi, J. C. Gomez, and A. Belaïd, “A Circular Grid-Based Rotation Invariant Feature Extraction Approach for Off-line Signature Verification,” in *2011 International Conference on Document Analysis and Recognition*, Sep. 2011, pp. 1289–1293. doi: 10.1109/ICDAR.2011.259.

[259] G. Pirlo and D. Impedovo, “Cosine similarity for analysis and verification of static signatures,” *IET Biometrics*, vol. 2, no. 4, pp. 151–158, 2013, doi: 10.1049/iet-bmt.2013.0012.

[260] G. Pirlo and D. Impedovo, “Verification of Static Signatures by Optical Flow Analysis,” *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 5, pp. 499–505, 2013, doi: 10.1109/THMS.2013.2279008.

[261] Y. Serdouk, H. Nemmour, and Y. Chibani, “A New Handwritten Signature Verification System Based on the Histogram of Templates Feature and the Joint Use of the Artificial Immune System with SVM,” in *Computational Intelligence and Its Applications*, A. Amine, M. Mouhoub, O. Ait Mohamed, and B. Djebbar, Eds., in IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing, 2018, pp. 119–127. doi: 10.1007/978-3-319-89743-1_11.

[262] F. Vargas, M. Ferrer, C. Travieso, and J. Alonso, “Off-line Handwritten Signature GPDS-960 Corpus,” in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, Sep. 2007, pp. 764–768. doi: 10.1109/ICDAR.2007.4377018.

[263] O. Mersa, F. Etaati, S. Masoudnia, and B. N. Araabi, “Learning Representations from Persian Handwriting for Offline Signature Verification, a Deep Transfer Learning Approach,” *2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, pp. 268–273, Mar. 2019, doi: 10.1109/PRIA.2019.8785979.

[264] D. Tsourounis, I. Theodorakopoulos, E. N. Zois, and G. Economou, “From text to signatures: Knowledge transfer for efficient deep feature learning in offline signature verification,” *Expert Systems with Applications*, vol. 189, p. 116136, Mar. 2022, doi: 10.1016/j.eswa.2021.116136.

[265] N. Arab, H. Nemmour, and Y. Chibani, “A new synthetic feature generation scheme based on artificial immune systems for robust offline signature verification,” *Expert Systems with Applications*, vol. 213, p. 119306, Mar. 2023, doi: 10.1016/j.eswa.2022.119306.

[266] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow Twins: Self-Supervised Learning via Redundancy Reduction,” in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 12310–12320. Accessed: Mar. 22, 2022. [Online]. Available: https://proceedings.mlr.press/v139/zbontar21a.html

[267] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, “Signature verification using a" siamese" time delay neural network,” presented at the Advances in neural information processing systems, 1994, pp. 737–744.

[268] D. Avola, M. J. Bigdello, L. Cinque, A. Fagioli, and M. R. Marini, “R-SigNet: Reduced space writer-independent feature learning for offline writer-dependent signature verification,” *Pattern Recognition Letters*, vol. 150, pp. 189–196, Oct. 2021, doi: 10.1016/j.patrec.2021.06.033.

[269] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 815–823. doi: 10.1109/CVPR.2015.7298682.

[270] W. Chen, X. Chen, J. Zhang, and K. Huang, “Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-Identification,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 403–412. Accessed: Apr. 27, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Chen_Beyond_Triplet_Loss_CVPR_2017_paper.html

[271] Q. Wan and Q. Zou, “Learning Metric Features for Writer-Independent Signature Verification using Dual Triplet Loss,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, Jan. 2021, pp. 3853–3859. doi: 10.1109/ICPR48806.2021.9413091.

[272] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[273] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7132–7141. doi: 10.1109/CVPR.2018.00745.

[274] L. Liu, L. Huang, F. Yin, and Y. Chen, “Offline signature verification using a region based deep metric learning network,” *Pattern Recognition*, vol. 118, p. 108009, Oct. 2021, doi: 10.1016/j.patcog.2021.108009.

[275] L. Liu, L. Huang, F. Yin, and Y. Chen, “Off-Line Signature Verification Using a Region Based Metric Learning Network,” in *Pattern Recognition and Computer Vision*, J.-H. Lai, C.-L. Liu, X. Chen, J. Zhou, T. Tan, N. Zheng, and H. Zha, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 74–86. doi: 10.1007/978-3-030-03398-9_7.

[276] S. Lai and L. Jin, “Learning Discriminative Feature Hierarchies for Off-Line Signature Verification,” in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Aug. 2018, pp. 175–180. doi: 10.1109/ICFHR-2018.2018.00039.

[277] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[278] V. Ruiz, I. Linares, A. Sanchez, and J. F. Velez, “Off-line handwritten signature verification using compositional synthetic generation of signatures and Siamese Neural Networks,” *Neurocomputing*, vol. 374, pp. 30–41, Jan. 2020, doi: 10.1016/j.neucom.2019.09.041.

[279] E. Parcham, M. Ilbeygi, and M. Amini, “CBCapsNet: A novel writer-independent offline signature verification model using a CNN-based architecture and capsule neural networks,” *Expert Systems with Applications*, vol. 185, p. 115649, Dec. 2021, doi: 10.1016/j.eswa.2021.115649.

[280] F. Chollet, “Xception: Deep Learning With Depthwise Separable Convolutions,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258. Accessed: Dec. 12, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Chollet_Xception_Deep_Learning_CVPR_2017_paper.html

[281] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, “Aggregated Residual Transformations for Deep Neural Networks,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500. Accessed: Dec. 12, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Xie_Aggregated_Residual_Transformations_CVPR_2017_paper.html

[282] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-ResNet and the impact of residual connections on learning,” in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, in AAAI’17. San Francisco, California, USA: AAAI Press, Feb. 2017, pp. 4278–4284.

[283] A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” *ArXiv*, vol. abs/1704.04861, 2017.

[284] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, “Learning Transferable Architectures for Scalable Image Recognition,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 8697–8710. doi: 10.1109/CVPR.2018.00907.

[285] X. Lu, L. Huang, and F. Yin, “Cut and Compare: End-to-end Offline Signature Verification Network,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, Jan. 2021, pp. 3589–3596. doi: 10.1109/ICPR48806.2021.9412377.

[286] Y.-J. Xiong and S.-Y. Cheng, “Attention Based Multiple Siamese Network for Offline Signature Verification,” in *Document Analysis and Recognition – ICDAR 2021*, J. Lladós, D. Lopresti, and S. Uchida, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 337–349. doi: 10.1007/978-3-030-86334-0_22.

[287] D. Engin, A. Kantarci, S. Arslan, and H. Kemel Ekenel, “Offline Signature Verification on Real-World Documents,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 808–809.

[288] T. Younesian, S. Masoudnia, R. Hosseini, and B. N. Araabi, “Active Transfer Learning for Persian Offline Signature Verification,” in *2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, Mar. 2019, pp. 234–239. doi: 10.1109/PRIA.2019.8786013.

[289] A. Jain, S. K. Singh, and K. Pratap Singh, “Multi-task learning using GNet features and SVM classifier for signature identification,” *IET Biometrics*, vol. 10, no. 2, pp. 117–126, 2021, doi: 10.1049/bme2.12007.

[290] N. Çalik, O. C. Kurban, A. R. Yilmaz, T. Yildirim, and L. Durak Ata, “Large-scale offline signature recognition via deep neural networks and feature embedding,” *Neurocomputing*, vol. 359, pp. 1–14, Sep. 2019, doi: 10.1016/j.neucom.2019.03.027.

[291] R. Ghosh, “A Recurrent Neural Network based deep learning model for offline signature verification and recognition system,” *Expert Systems with Applications*, p. 114249, Nov. 2020, doi: 10.1016/j.eswa.2020.114249.

[292] M. B. Yılmaz and K. Öztürk, “Recurrent Binary Patterns and CNNs for Offline Signature Verification,” in *Proceedings of the Future Technologies Conference (FTC) 2019*, K. Arai, R. Bhatia, and S. Kapoor, Eds., in Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2020, pp. 417–434. doi: 10.1007/978-3-030-32523-7_29.

[293] H. Li, P. Wei, and P. Hu, “Static-Dynamic Interaction Networks for Offline Signature Verification,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, Art. no. 3, May 2021, doi: 10.1609/aaai.v35i3.16284.

[294] P. R. Prajapati, S. Poudel, M. Baduwal, S. Burlakoti, and S. P. Panday, “Signature Verification using Convolutional Neural Network and Autoencoder,” *Journal of the Institute of Engineering*, vol. 16, no. 1, Art. no. 1, Apr. 2021, doi: 10.3126/jie.v16i1.36533.

[295] H. Li, P. Wei, and P. Hu, “AVN: An Adversarial Variation Network Model for Handwritten Signature Verification,” *IEEE Transactions on Multimedia*, pp. 1–1, 2021, doi: 10.1109/TMM.2021.3056217.

[296] H. Li, H. Li, H. Zhang, and W. Yuan, “Black-box attack against handwritten signature verification with region-restricted adversarial perturbations,” *Pattern Recognition*, vol. 111, p. 107689, Mar. 2021, doi: 10.1016/j.patcog.2020.107689.

[297] S. Roy, D. Sarkar, S. Malakar, and R. Sarkar, “Offline signature verification system: a graph neural network based approach,” *J Ambient Intell Human Comput*, Nov. 2021, doi: 10.1007/s12652-021-03592-0.

[298] J.-X. Ren, Y.-J. Xiong, H. Zhan, and B. Huang, “2C2S: A two-channel and two-stream transformer based framework for offline signature verification,” *Engineering Applications of Artificial Intelligence*, vol. 118, p. 105639, Feb. 2023, doi: 10.1016/j.engappai.2022.105639.

[299] S. Manna, S. Chattopadhyay, S. Bhattacharya, and U. Pal, “SWIS: Self-Supervised Representation Learning For Writer Independent Offline Signature Verification,” *arXiv:2202.13078 [cs, eess]*, Feb. 2022, Accessed: Mar. 30, 2022. [Online]. Available: http://arxiv.org/abs/2202.13078

[300] S. Chattopadhyay, S. Manna, S. Bhattacharya, and U. Pal, “SURDS: Self-Supervised Attention-guided Reconstruction and Dual Triplet Loss for Writer Independent Offline Signature Verification.” arXiv, Jun. 26, 2022. doi: 10.48550/arXiv.2201.10138.

[301] M. Hussain, J. J. Bird, and D. R. Faria, “A Study on CNN Transfer Learning for Image Classification,” in *Advances in Computational Intelligence Systems*, A. Lotfi, H. Bouchachia, A. Gegov, C. Langensiepen, and M. McGinnity, Eds., Springer International Publishing, 2019, pp. 191–202.

[302] G. Hinton, O. Vinyals, and J. Dean, “Distilling the Knowledge in a Neural Network.” arXiv, Mar. 09, 2015. doi: 10.48550/arXiv.1503.02531.

[303] J. Ba and R. Caruana, “Do Deep Nets Really Need to be Deep?,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2014. Accessed: Jul. 06, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2014/hash/ea8fcd92d59581717e06eb187f10666d-Abstract.html

[304] Y. Bengio, A. Courville, and P. Vincent, “Representation Learning: A Review and New Perspectives.” arXiv, Apr. 23, 2014. doi: 10.48550/arXiv.1206.5538.

[305] I. Theodorakopoulos, G. Economou, S. Fotopoulos, and C. Theoharatos, “Local manifold distance based on neighborhood graph reordering,” *Pattern Recognition*, vol. 53, pp. 195–211, May 2016, doi: 10.1016/j.patcog.2015.12.006.

[306] I. Theodorakopoulos, F. Fotopoulou, and G. Economou, “Geometric Regularization of Local Activations for Knowledge Transfer in Convolutional Neural Networks,” *Information*, vol. 12, no. 8, Art. no. 8, Aug. 2021, doi: 10.3390/info12080333.

[307] I. Theodorakopoulos, F. Fotopoulou, and G. Economou, “Local Manifold Regularization for Knowledge Transfer in Convolutional Neural Networks,” in *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA*, Jul. 2020, pp. 1–8. doi: 10.1109/IISA50023.2020.9284400.

[308] J. B. Kruskal, “On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem,” *Proceedings of the American Mathematical Society*, vol. 7, no. 1, pp. 48–50, 1956, doi: 10.2307/2033241.

[309] R. Zemel and M. Carreira-Perpiñán, “Proximity Graphs for Clustering and Manifold Learning,” in *Advances in Neural Information Processing Systems*, MIT Press, 2004. Accessed: Apr. 10, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2004/hash/dcda54e29207294d8e7e1b537338b1c0-Abstract.html

[310] I. Theodorakopoulos and D. Tsourounis, “A Geometric Perspective on Feature-Based Distillation,” in *Advancements in Knowledge Distillation: Towards New Horizons of Intelligent*

Systems, 1st ed.in Studies in Computational Intelligence, no. 1100. Springer Nature, 2023, pp. 33–63. [Online]. Available: https://doi.org/10.1007/978-3-031-32095-8_2

[311] M. A. Ferrer, J. F. Vargas, A. Morales, and A. Ordonez, “Robustness of Offline Signature Verification Based on Gray Level Features,” *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 966–977, Jun. 2012, doi: 10.1109/TIFS.2012.2190281.

[312] F. Kleber, S. Fiel, M. Diem, and R. Sablatnig, “CVL-DataBase: An Off-Line Database for Writer Retrieval, Writer Identification and Word Spotting,” in *2013 12th International Conference on Document Analysis and Recognition*, Aug. 2013, pp. 560–564. doi: 10.1109/ICDAR.2013.117.

[313] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, JMLR Workshop and Conference Proceedings, Mar. 2010, pp. 249–256. Accessed: Jul. 20, 2022. [Online]. Available: https://proceedings.mlr.press/v9/glorot10a.html