Design of a deep neural network architecture dedicated to handwriting synthesis using kinematic sensors from a digital pen.

Florent Imbert

COLLEGE DOCTORAL BRETAGNE
MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

THÈSE DE DOCTORAT DE

L'INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES

ÉCOLE DOCTORALE N° 601
Mathématiques, Télécommunications, Informatique, Signal, Systèmes, Électronique
Spécialité : *Informatique*

Par
IMBERT Florent

Conception d'une architecture de réseaux de neurones profonds dédiée à la synthèse d'écriture manuscrite à partir de capteurs cinématiques d'un stylo numérique.

Thèse présentée et soutenue à Rennes, le 21 novembre 2024
Unité de recherche : IRISA
Thèse N° : 24ISAR 20 / D24-20

Rapporteurs avant soutenance :

Andreas FISCHER	Full professor, University of Applied Sciences and Arts Western Switzerland (HES-SO)
Clément Chatelain	Maître de conférences HDR, INSA Rouen Normandie

Composition du Jury :

Président :	Elisa Fromont	Professeur des universités, Université de Rennes
Examinateurs :	Andreas Fischer	Full professor, University of Applied Sciences and Arts Western Switzerland (HES-SO)
	Clément Chatelain	Maître de conférences HDR, INSA Rouen Normandie
	Elisa Fromont	Professeur des universités, Université de Rennes
	Nicolas Ragot	Professeur des universités, Polytech Tours
Dir. de thèse :	Eric Anquetil	Professeur des universités, INSA Rennes
Co-dir. de thèse :	Romain Tavenard	Professeur des universités, Université Rennes 2

Invité(s) :

Co-encadr. de thèse :	Yann Soullard	Maître de conférences, Université Rennes 2

Remerciements

Je voudrais d'abord exprimer ma reconnaissance à l'ensemble des membres du jury pour l'intérêt qu'ils ont porté à cette thèse. Je remercie Andreas Fisher et Clément Chatelain d'avoir bien voulu rapporter ce manuscrit. Merci aussi à Elisa Fromont et Nicolas Ragot d'avoir accepté d'être les examinateurs de ma soutenance.

Une immense gratitude va naturellement à mes encadrants, Eric Anquetil, Romain Tavenard et Yann Soullard, pour leurs conseils avisés. Je leur suis également reconnaissant pour le temps précieux qu'ils m'ont consacré, tant pour partager leurs connaissances et expériences que pour relire mes articles, et notamment ce manuscrit.

Je remercie chaleureusement tous les membres de l'équipe Shadoc (ex IntuiDoc), avec une pensée particulière pour mon camarade Omar Krichen, pour ses partages de sagesse et de conseils, tout comme nos moments de loisirs, que ce soit aux échecs, au badminton ou au tennis.

Je ne peux terminer ces remerciements sans avoir une pensée pour ma compagne qui m'a accompagné et soutenu pendant toutes ces années. Je tiens également à remercier mes parents pour leur soutien sans faille à mes projets depuis de nombreuses années.

RÉSUMÉ EN FRANÇAIS

Contexte

L'utilisation des technologies numériques dans l'éducation s'est intensifiée. Au cours de la dernière décennie, de nombreuses études se sont concentrées sur la mise en œuvre de la technologie numérique dans les écoles, dans le but de créer des logiciels éducatifs qui soutiennent efficacement l'apprentissage des élèves et des enseignants. Avant qu'une application éducative puisse être considérée comme pratique, il est crucial de développer des solutions informatiques à la fois fiables et robustes.

Le projet ANR Franco-Allemand KIHT (Fig. 1), dont ma thèse fait partie, se concentre sur l'apprentissage de l'écriture.

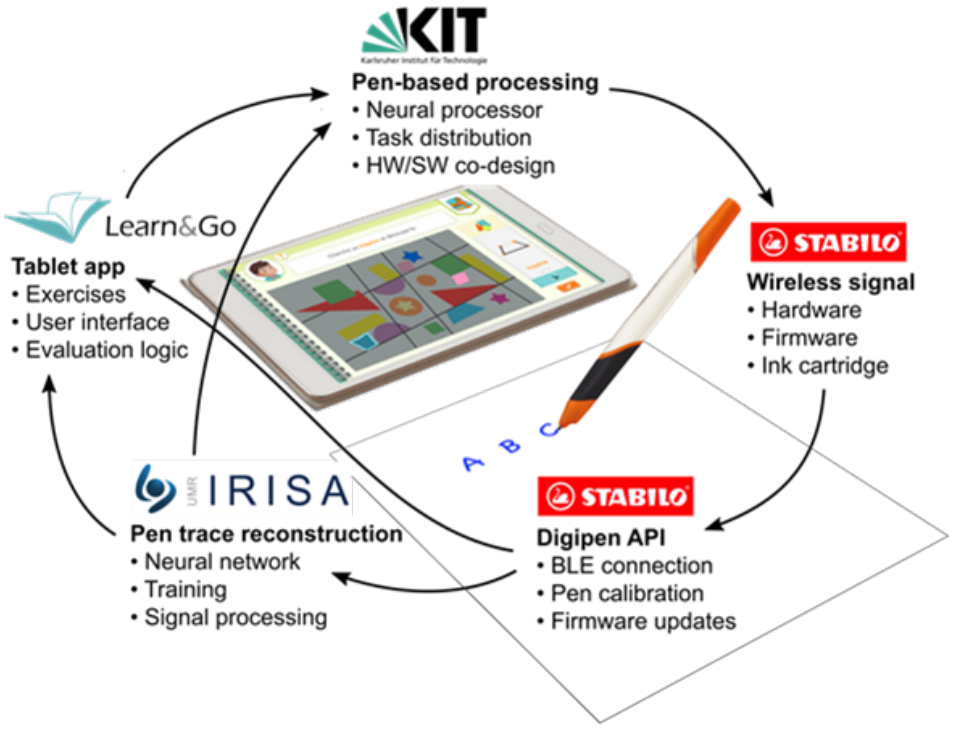


Figure 1 – Vue d'ensemble du projet KIHT.

L'objectif de ce projet est de développer un dispositif d'apprentissage intelligent pour l'écriture automatique, composé d'éléments existants, qui peut être mis à la disposition du plus grand nombre d'étudiants possible. La base est l'application « Kaligo » de la société française Learn&Go, qui jusqu'à présent dépendait d'ordinateurs tablettes coûteux avec un dispositif d'écriture dédié et nécessitait d'écrire sur l'écran tactile capacitif. L'association du stylo électronique « DigiPen » de la société STABILO et de l'application Kaligo vise précisément à rendre cela possible : une aide à l'apprentissage de l'écriture manuscrite dont le plus grand nombre d'enfants peuvent bénéficier. En outre, le Digipen permet non seulement d'utiliser n'importe quelles tablettes disponibles dans le commerce, mais aussi il peut être utilisé sur n'importe quel support (papier), ce qui est un grand avantage pour le développement de l'écriture des enfants.

Notre objectif au sein de l'institut de recherche français IRISA, et plus particulièrement de l'équipe Intuidoc/ShaDoc, est de mener des recherches sur la conception d'un système s'appuyant sur des algorithmes d'apprentissage automatique pour générer automatiquement des traces manuscrites en ligne à partir de signaux produits par des capteurs du stylo numérique, tandis que l'institut allemand ITIV, qui fait partie de l'institut de technologie de Karlsruhe KIT, étudie divers concepts visant à intégrer les algorithmes d'apprentissage automatique que nous développons dans le stylo numérique. De cette manière, la complexité globale du système est répartie entre le logiciel et le matériel, ce qui permet une exécution rapide et efficace des algorithmes. L'objectif est de permettre la reconstruction en ligne du tracé du stylo à partir des données du capteur inertiel à faible coût Digipen.

Pour relever les défis inhérents à la reconstruction des traces de stylo à partir des données inertielles, il est essentiel d'exploiter des techniques avancées apprentissage profond. Les méthodes traditionnelles, telles que la double intégration, introduisent une dérive importante dans les données, ce qui rend les résultats presque inutilisables. En revanche, l'application de réseaux neuronaux formés à la fois sur les données des capteurs et sur les traces de stylo réelles permettent d'éviter ou limiter une telle dérive, ce qui permet d'obtenir des reconstructions d'une qualité nettement supérieure.

Prétraitement

Le prétraitement est une étape cruciale dans la préparation des données des capteurs et des tablettes pour l'analyse, en particulier lorsqu'il s'agit de méthodes d'apprentissage profond. L'un des principaux défis de ce processus est la variation des fréquences d'échantillonnage

entre les données du capteur et celles de la tablette dont la trace est utilisée pour l'apprentissage d'un modèle. Cet écart peut entraîner des points de données mal alignés, ce qui complique la comparaison et l'analyse précises des informations. En outre, le signal de la tablette est susceptible d'être perdu lorsque le stylo est levé trop haut, ce qui entraîne des lacunes dans les données qui doivent être corrigées. Pour garantir l'efficacité des modèles d'apprentissage profond, il est essentiel de prétraiter les données de manière à ce que la vérité terrain et les données des capteurs aient des tailles identiques. Cela implique un pipeline de prétraitement détaillé dans la figure 2.

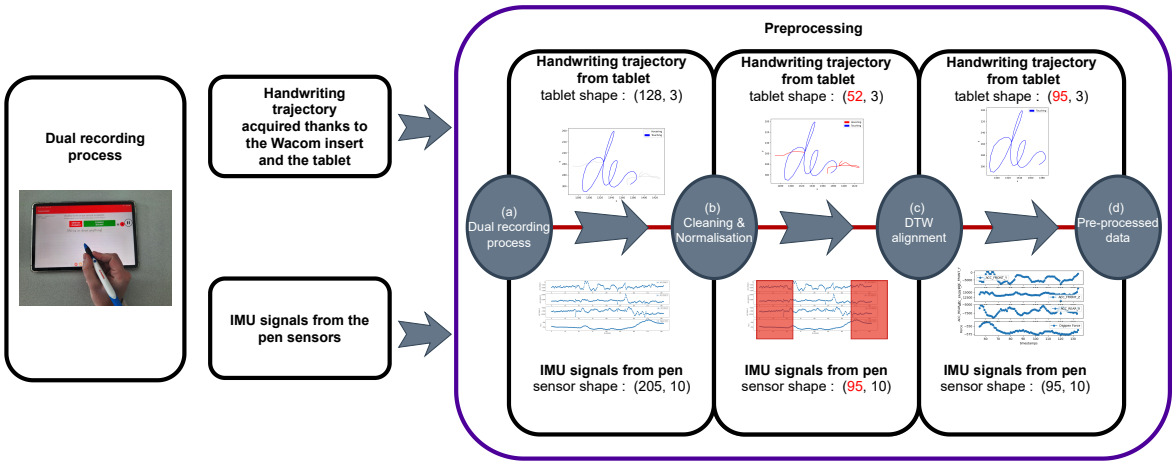


Figure 2 – Notre pipeline de prétraitement. (a) grâce à la double acquisition, nous récupérons les signaux du Digipen et la vérité terrain, (b) nous supprimons le début et la fin du survol, qui ne sont pas des données liées à l'écriture manuscrite, (c) nous alignons la vérité terrain et les signaux du capteur à l'aide de l'algorithme DTW, (d) nous obtenons les données prétraitées utilisées pour l'entraînement.

Afin de respecter autant que possible la dynamique de l'écriture qui est présente dans les données d'entrée, qui sont des accélérations, nous proposons une approche d'alignement basée sur l'algorithme Dynamic Time Warping (DTW) (Figure 2(C)). Comme le taux d'échantillonnage des données du capteur est plus élevé que celui des données de la tablette, nous avons choisi d'augmenter l'échantillonnage des données de la tablette pour qu'il corresponde à la longueur des données du capteur. En effet, nous voulons conserver autant de données de capteurs que possible pour aider à la reconstruction, et nous ne voulons pas limiter notre impact sur la dynamique des données de capteurs.

Un modèle expert de la trace écrite

Afin de prendre en compte les états passés et futurs lors de la mise en correspondance d'une séquence d'entrée avec un signal de sortie, nous proposons d'utiliser une architecture de réseau neuronal convolutionnel temporel non causal (Temporal Convolutional Network - TCN). Nous nommons cet expert de la trace écrite TEM pour "Touching Expert Model" car il est entraîné sur les coups de toucher uniquement.

Une architecture convolutive (Convolutional neural network - CNN) peut capturer des caractéristiques spatiales qui se réfèrent à la disposition des points de données d'une séquence et à la relation entre eux au sein de la séquence. L'avantage du TCN sur le CNN est sa capacité à capturer un contexte plus éloigné avec moins de profondeur, grâce à des convolutions dilatées. Cela permet au TCN de regarder plus loin dans le passé et le futur tout en conservant une architecture de réseau moins profonde, réduisant ainsi le risque de rencontrer des disparition du gradient qui survient généralement avec des réseaux plus profonds.

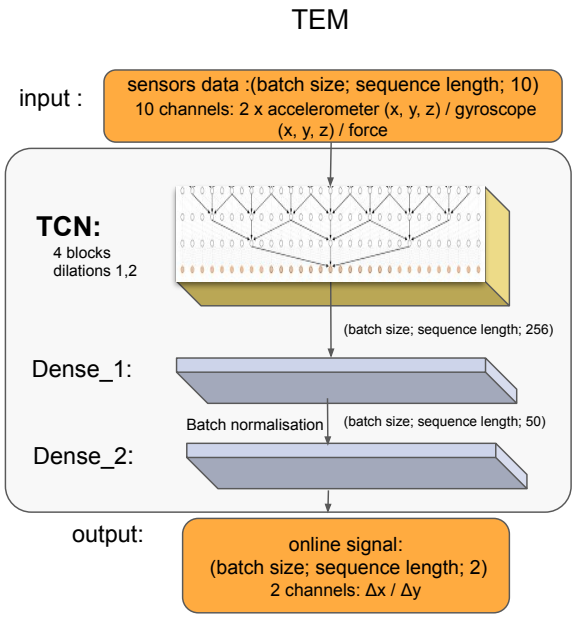


Figure 3 – Notre modèle TEM pour la reconstruction de trajectoires manuscrites.

Une approche de mélange d'experts (Mixture-Of-Experts - MOE) pour une meilleure collaboration et une meilleure spécification des tâches

Nous pouvons formaliser notre problème comme un apprentissage multitâche, en ce sens que nous avons deux tâches liées, la première étant la prédiction de l'écriture elle-même, la seconde le mouvement du stylo entre ces différentes parties. Ces deux tâches diffèrent par leur dynamique et leur nature (signaux bidimensionnels et tridimensionnels). En pratique, pour prendre en compte ces différentes natures de signaux, nous proposons de combiner deux réseaux neuronaux, chacun étant dédié à l'une de ces natures de signaux.

Inspirés par nos travaux précédents où une architecture basée sur un réseau de neurones est entraînée sur des touchers de crayon, produisant des reconstructions dégradées sur les parties plume haute, nous suggérons d'utiliser cette architecture de réseau comme modèle expert pour les touchés de crayon. Nous restons convaincus que le réseau TCN est la bonne architecture de réseau pour traiter ces données IMU pour les raisons mentionnées précédemment.

Nous nous sommes donc intéressés à la manière d'entraîner notre modèle, en particulier on veux traiter les parties plume haute pour correctement repositionner la trace après une trajectoire de survol. Compte tenu de la spécificité des parties plume haute, nous suggérons d'utiliser le modèle sur des séquences complètes, car nous pensons que le fait de donner autant de contexte que possible peut être bénéfique pour la prédiction des trajectoires. La raison pour laquelle nous entraînons notre réseau sur des séquences entières, plutôt que d'isoler les traits de crayon, vient de la dynamique complexe des interactions entre le stylo et la tablette. En disposant de la séquence complète, le modèle obtient des informations sur les modèles de transition entre les touchers du stylo et les parties plume haute.

C'est pourquoi nous proposons une nouvelle approche d'apprentissage (Fig 4) : avec deux réseaux neuronaux en parallèle, le premier sur les touchers et le second sur les séquences complètes.

Méthodes d'adaptation de domaine pour le traitement des données enfants

Un stylo numérique équipé de capteurs cinématiques permet aux utilisateurs d'écrire sur n'importe quelle surface tout en préservant simultanément une trajectoire numérique de l'écriture. L'un des principaux problèmes réside dans la différence entre les signaux capturés par les adultes et les enfants. Pour une trace d'écriture similaire, on observe de grandes différences dans les signaux des capteurs en raison des différences de vitesse et de confiance dans le geste d'écriture des enfants. Pour y remédier, nous étudions une approche d'adaptation au domaine pour construire une représentation intermédiaire unifiée des caractéristiques visant à faciliter la reconstruction de la trajectoire. Nous démontrons l'intérêt des méthodes d'adaptation au domaine pour tirer parti des connaissances existantes afin de les appliquer dans différents contextes. Plus précisément, nous comparons notre approche d'adaptation au domaine avec deux autres méthodes : l'entraînement du modèle à partir de zéro et l'ajustement d'un modèle préentraîné sur les données spécifiques considérées.

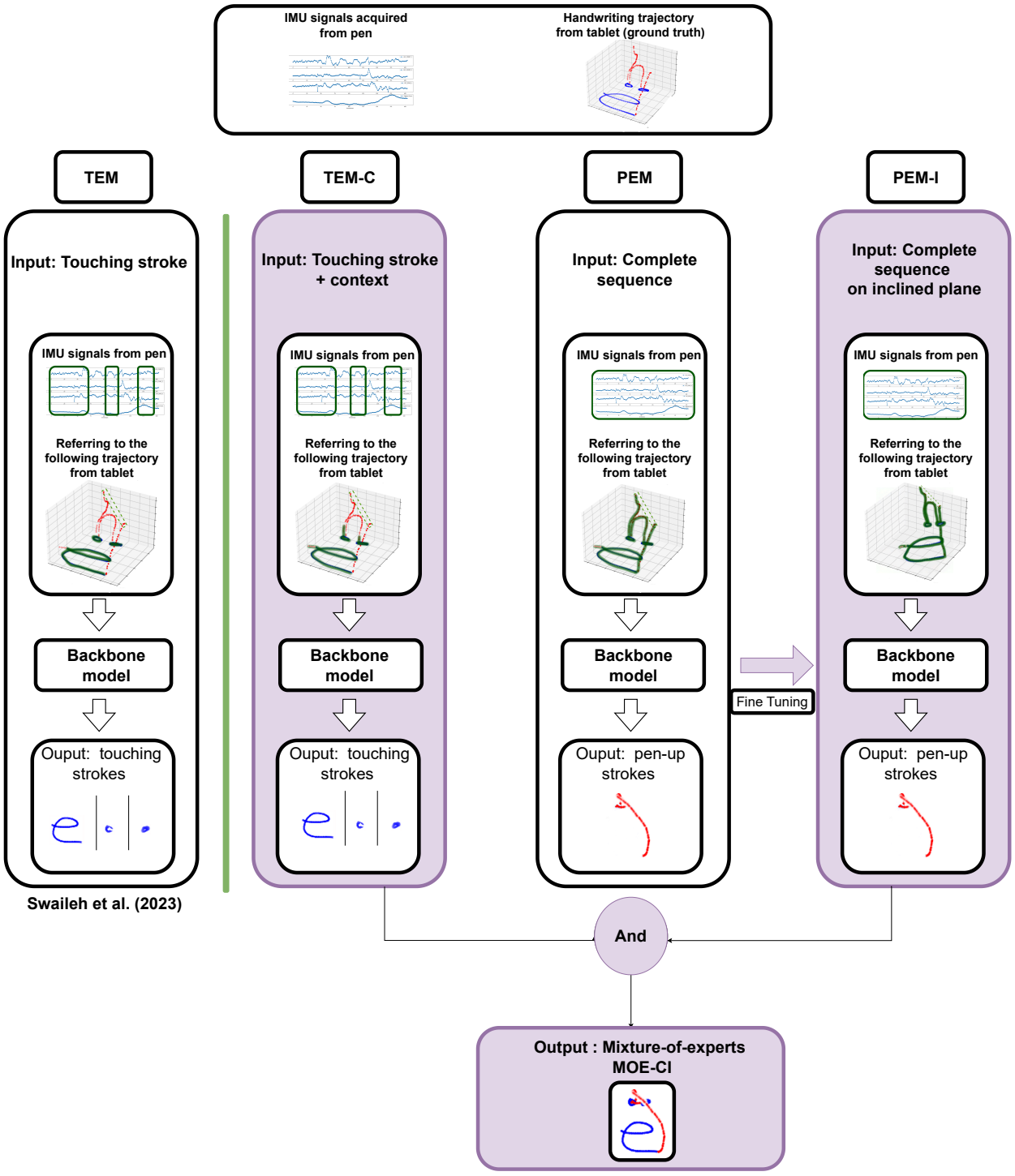


Figure 4 – Notre approche MOE-CI fusionne les avancées pour les deux modèles experts : TEM-C, qui améliore le modèle expert du toucher (TEM) en intégrant le contexte du survol (en vert), qui donne plus de contexte et facilite la transition du survol au toucher, HEM-I, qui améliore le modèle expert du survol (HEM) par un ajustement sur les données 3D pour une meilleure compréhension de cette troisième dimension.

TABLE OF CONTENTS

List of acronyms

— **DNN** : Deep Neural Network

— **RNN** : Recurrent Neural Network

— **LSTM** : Long Short-Term Memory

— **BLSTM** : Bidirectional Long Short-Term Memory

— **CNN** : Convolutional Neural Network

— **FCN** : Fully Convolutional Network

— **TCN** : Temporal Convolutional Networks

— **ReLU** : *Rectified Linear Unit*

— **SGD** : Stochastic Gradient Descent

— **seq2seq** : Sequence-to-Sequence model

— **MSE** : Mean Squared Error

— **DTW** : Dynamic Time Warping

— **SOTA** : State-Of-The-Art

— **HMM** : Hidden Markov Model

— **PCA** : Principal Component Analysis

— **DA** : Domaine Adaptation

— **DANN** : Domain-Adversarial Training of Neural Networks

INTRODUCTION

General thesis context

Handwriting involves creating visual symbols on a surface to express thoughts and ideas. Like speech, it serves as a method to communicate, with these symbols corresponding to particular languages, enabling others to understand the message. Over the course of a person's life, their ability to write progresses and changes uniquely, giving rise to a handwriting style that is distinct and individualized [Alaei et al., 2022].

Handwriting remains a crucial skill to maintain. Research indicates that using pen and paper enhances cognitive development and memory retention by activating various neural networks in the brain, thus boosting learning capacity [James, 2017]. Additionally, handwriting can foster creativity and improve writing skills, as it encourages focus and can lead to the generation of more innovative ideas during the drafting process.

While the faithful reproduction of letter shapes is widely accepted to be essential for legible writing, the importance of motor skills is much less appreciated. Routine handwriting is characterized by a smooth course of velocity over time with just one maximum of velocity within a stroke [Mai et al., 1994]. If words are written repeatedly, experienced writers retain the characteristic execution of the movement over all runs. Writing movements of this kind are understood by Mai and Marquardt to be automated [N. Mai and C. Marquardt, 2002]. Non-automated movements, on the other hand, are characterized by several maxima in the course of velocity, so they are executed with several motion impulses per stroke.

Learned movements are carried out automatically, that is, completely planned in advance and then no longer consciously controlled or corrected in detail. They are controlled by the cerebellum and the motor cortex and require no visual control [Marquardt et al., 1996]. In less experienced writers, however, writing is associated with conscious hand-eye coordination. This leads to less automated movements. The cerebrum, which is responsible for controlling cognitive movements, then takes control of this non-automated motion execution. This conscious movement control not only leads to significantly slower writing,

but also heavily burdens the brain's working memory and makes it difficult to focus on spelling or content at the same time.

Handwriting plays a crucial role in academic success, as it enhances the learning process. It significantly improves the retention and understanding of learned material [Mueller et al., 2014]. According to another study [Oviatt et al., 2012], handwritten papers contain 38% more relevant concepts compared to typed papers, indicating an improved creativity when using a pen rather than a keyboard. Askvik et al. show that handwriting is vital to facilitate and optimize learning [Askvik et al., 2020]. The French National Ministry of Education has noted that students' spelling skills have consistently declined over the past few decades [1]. The same problem has also been observed in Germany, where a large proportion of students have severe handwriting problems. According to a 2019 survey [Schreibmotorik Institut in Kooperation mit dem Verband Bildung und Erziehung (VBE Bund) und den 16 VBE Landesverbänden, 2019], this concerns 53% of boys and 33% of girls in secondary education in Germany.

Concurrently, the use of digital technologies in education has intensified. Over the last decade numerous studies have focused on the implementation of digital technology in schools, aiming to create educational software that effectively supports students learning and teachers. If educational applications such as Kaligo offer instant visual feedback so that the child can understand his or her successes and errors by analyzing the tracings. Its main limitation is that it restricts writing to tablets. Hence the aim of this project is to remove this constraint.

My thesis is a part of the Kaligo-based Intelligent Handwriting Teacher (KIHT) project. The KIHT project (Fig. 5), is a French-German bilateral ANR project. It involves four partners: Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA, France), Karlsruhe Institute of Technology (KIT, Germany), Learn&Go (France) and STABILO (Germany). The objective of this project is to enhance children's writing skills by merging traditional and digital methods. It leverages the tactile benefits of writing on any surface while incorporating digital technology for greater adaptability and data management. Although current learning methods using tablets facilitate writing, the KIHT project aims to introduce a specialized pen, the Digipen, which can be used with both tablets and regular paper. This dual functionality significantly benefits the development of children's writing skills.

1. https://www.education.gouv.fr/les-performances-en-orthographe-des-eleves

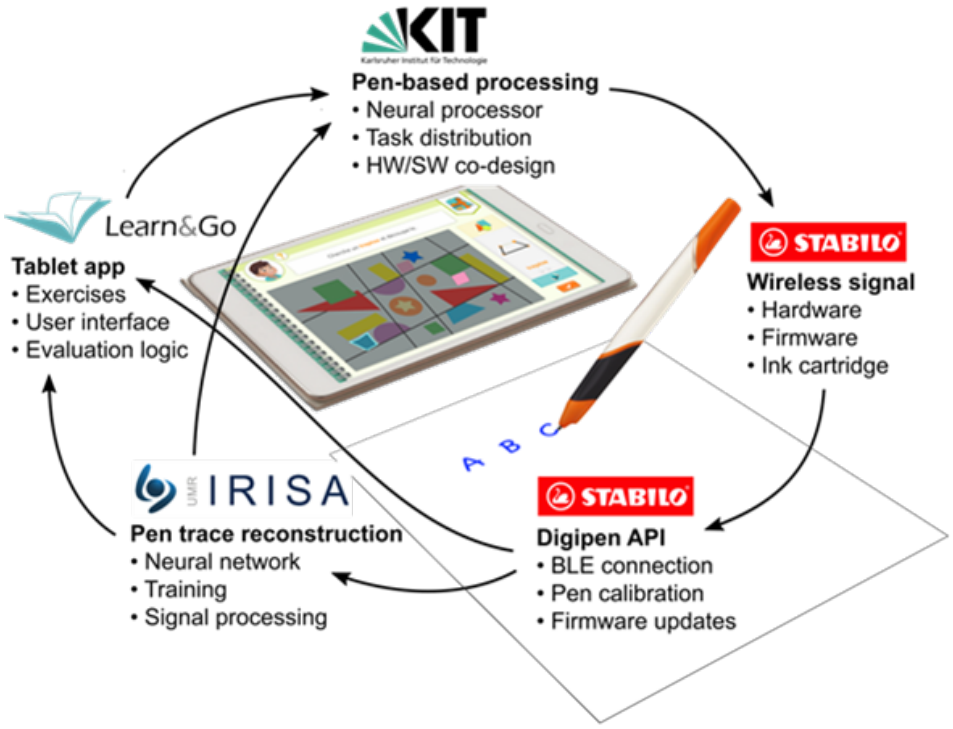


Figure 5 – Overview of the KIHT project.

The project consortium comprises two universities and two companies based in Germany and France. STABILO is in charge of creating the pen. The Karlsruhe Institute of Technology is developing concepts for integrating AI algorithms tailored to embedded hardware. Learn&Go is enhancing their Kaligo app to help students learn writing and spelling using a tablet and the electronic pen. IRISA is tasked with designing and developing a deep learning-based AI solution that reconstructs online handwriting trajectories from the pen.

Our objective is to conduct research on a system to automatically generate online handwritten traces from signals produced by digital pen sensors.

To address the inherent challenges of reconstructing pen traces from inertial data, it is essential to leverage advanced deep learning techniques. Traditional methods, such as double integration, introduce significant drift into the data, rendering the outcomes nearly unusable. In contrast, the application of neural networks trained on both sensor data and actual pen traces can be less subject to drift, yielding reconstructions of substantially higher quality. The initial encouraging results from the collaborative research conducted

by KIT and STABILO underscore the crucial role of deep learning algorithms in enhancing the accuracy and usability of these reconstructions (Fig. 6).

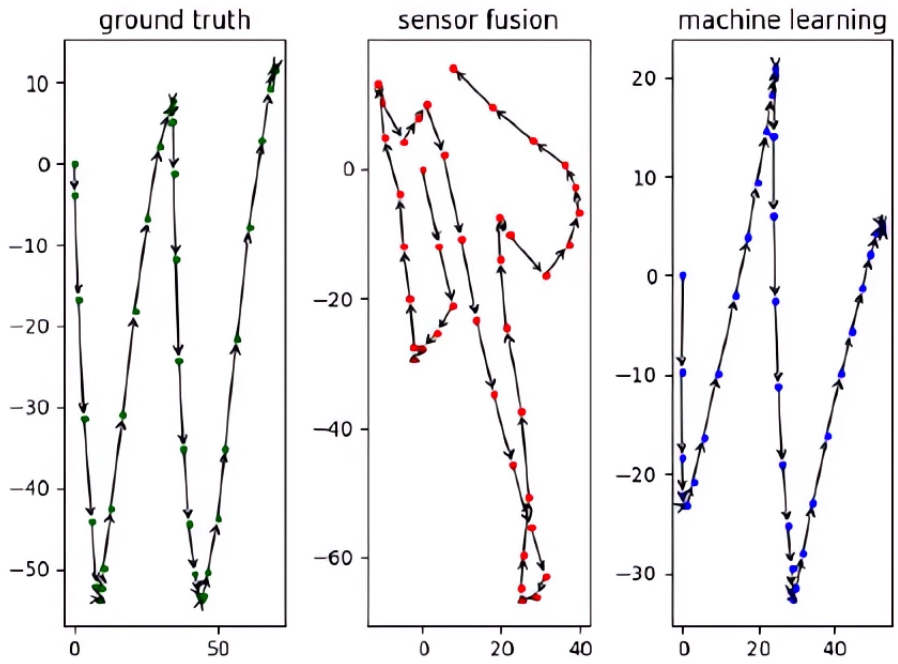


Figure 6 – Pen trace reconstruction. Left: Original trace. Center: Result from a mathematical reconstruction (sensor signals, which are acceleration, are converted into a trajectory by double integration). Right: First result after applying a trained neural network. Illustration from the project proposal.

Chapter 1

Presentation of the Inertial Measurement Units stylus and the reconstruction task

In this section, we present the specific features of the inertial measurement unit (IMU) sensor, and the challenges resulting from this technology.

1.1 Introduction to Inertial Measurement Units

An Inertial Measurement Unit (IMU) is an electronic device that captures the dynamics of a moving body, utilizing accelerometers and gyroscopes to measure linear acceleration and angular velocity, respectively. These sensors track the translational and rotational movements within a local reference frame oriented by gravity. Often, a magnetometer is also included to gauge the magnetic field's strength and direction, aiding in defining movements relative to an absolute coordinate system.

IMU have become omnipresent across various fields in recent years, including robotics, quality control, medical rehabilitation, and activity recognition. Their popularity is attributed to their compact size, lightweight nature, minimal power requirements, and affordability. Additionally, IMU operate independently as they are self-contained and do not rely on external references.

One application of IMU is to digitalize handwriting without the need for physical writing surfaces, aligning well with advancements in technology. However, this use case is not without challenges. A significant limitation of IMU is error accumulation over time, (named sensor drift) which leads to increasing inaccuracies in position data derived from integration of the sensors outputs. IMU are also susceptible to noise (temperature variations, vibrations, interference, etc.). These errors can be particularly problematic in handwriting reconstruction, where precision is crucial [Pan et al., 2018].

1.2 Digipen: an Inertial Measurement Units stylus for digital handwriting trace reconstruction

Digital devices play a crucial role in enhancing the learning experience for both students and teachers by facilitating active learning techniques and offering immediate feedback [Simonnet et al., 2019]. The literature on e-learning [Atilola et al., 2014] highlights the accuracy and reliability of computer-based analysis for generating relevant feedback for correction or guidance. Based on this, pen-based tablet applications have been developed to provide personalized feedback to children [Krichen et al., 2022].

Despite the increasing reliance on digital platforms, there remains a need for children to learn writing on paper, as it remains the most widely used surface. To address this, digital pens, such as the Digipen stylus developed by STABILO, have been equipped with kinematic sensors to track pen movements (Fig. 1.1). Such a pen allows to capture handwriting gestures on any surface such as paper, to visualize and analyze the handwriting reconstruction on a digital medium (e.g. tablet, computer).

More specifically, version 6.0 of the Digipen is equipped with:

— a front accelerometer;

— a rear accelerometer;

— a front gyroscope;

— a magnetometer;

— a force sensor;

— a microcontroller which is used for reading the sensor data, processing it for wireless transfer and running the Bluetooth stack.

Each sensor has its own reference system (Fig. 1.2).

From this pen, our goal is to reconstruct its digital trace. To achieve this, we are focusing on deep learning methods, which have recently seen significant advances, particularly in the field of remote sensing and tracking systems [Marvasti-Zadeh et al., 2022]. The integration of IMU sensors offers a cost-effective solution, enabling us to create an affordable pen. However, the use of IMU sensors comes with certain limitations, which will be discussed in detail in the following chapter.

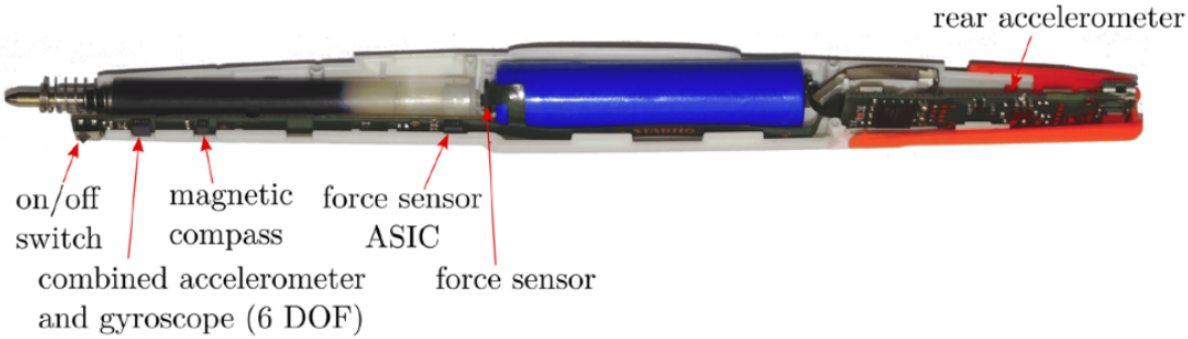


Figure 1.1 – © STABILO International Digipen's sensor location

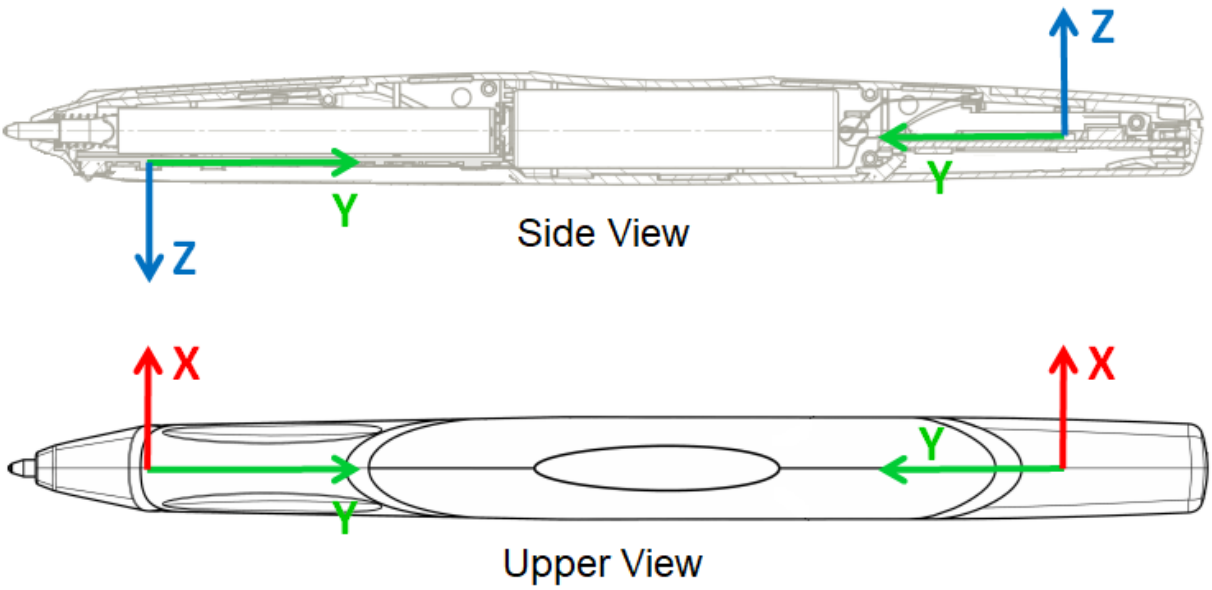


Figure 1.2 – © STABILO International Digipen's sensor reference

Chapter 2

Handwiting trajectory reconstruction challenges

Reconstructing handwritten text from kinematic signals stands at the crossroads of multiple disciplines, including biomechanics, signal processing, machine learning, and computational linguistics. This intricate process involves translating the complex motion data generated during writing, which is often influenced by noise and variability in kinematic measurements, shaped by an individual's unique writing style and the conditions of writing, into the dynamic nature of handwriting itself. This data is captured through accelerometers, gyroscopes, and other sensors, providing a detailed representation of the writing process. It is a challenge marked by several obstacles which are discussed in the following.

2.1 Pen related challenges

Among the most difficult tasks in reconstructing handwritten text from kinematic signals is dealing with the significant noise and variability present in the data. This variability can originate from the sensors themselves. In fact, while IMU sensors have the advantage of being inexpensive, the downside is that they generate noisy signals. The main limitation of IMU is the accumulation of errors over time (called sensor drift), which leads to increasing inaccuracies in the data. These errors can be particularly problematic in handwriting reconstruction, where precision is crucial especially for e-learning in order to give appropriate feedback. As the aim is to be able to write on any support (i.e. paper or tablet), this also means greater variability in signals and noise. Indeed, data acquired on paper are noisier than that acquired on a tablet due to the greater friction on paper.

Working with the Stabilo Digipen involves a number of challenges induced by the hardware. One challenge we face is the need for the tablet's trajectory to be used as the ground truth for training, which differs from the trajectory of the pen. This difference

is particularly noticeable during pen-up movement above the tablet surface, a movement that is not always reliably detected by the tablet. Synchronisation of timestamps between devices and accuracy of sensor data are critical in applications involving digital pen and tablet interfaces. In addition, grouping Bluetooth points, particularly for transmitting data between a pen and a tablet, involves collecting and buffering 12 data points before forming a packet for transmission. This method enhances efficiency by reducing transmission overhead but this process introduces complexities due to varying transmission speeds and the loss of the original dynamic.

Another important factor is that the ground truth is inherently dependent on the characteristics of the tablet, such as its model, screen size and sampling rate. These variables can dramatically affect the interpretation of the pen's position and movements, leading to discrepancies in the captured data from the tablet.

2.2 Writer challenges

Another source of variability stems from the inherent differences in an individual's writing technique. The act of writing is a complex motor task that requires coordinated movements of the fingers, wrists, and arms, with each person having a distinct handwriting style that may include a preference for cursive or block letters. In addition, handwriting is composed of two distinct parts (Fig 2.1): the act of writing itself and the process of pen-up movements, which involves repositioning the trajectory of the writing instrument. The writing phase encompasses the direct contact of the pen or pencil with the surface (in 2D), creating visible marks. In contrast, pen-up movements involves subtle, above-the-surface movements that adjust the position and angle of the writing tool (in 3D), setting up for the next stroke or letter. The dual nature of these different actions makes the study of handwriting more complicated.

Inter-scripter variability, particularly in the context of reconstructing handwriting from inertial measurement units, presents significant challenges that stem from the inherent differences between the handwriting gestures of children and adults. Children's handwriting gestures tend to be slower and more hesitant compared to those of adults. This discrepancy can largely be attributed to the developmental aspects of motor skills in children, who are still in the process of refining their coordination and muscle control. Such variability affects the handwriting reconstruction, as the sensors capture a wide range of motion dynamics.

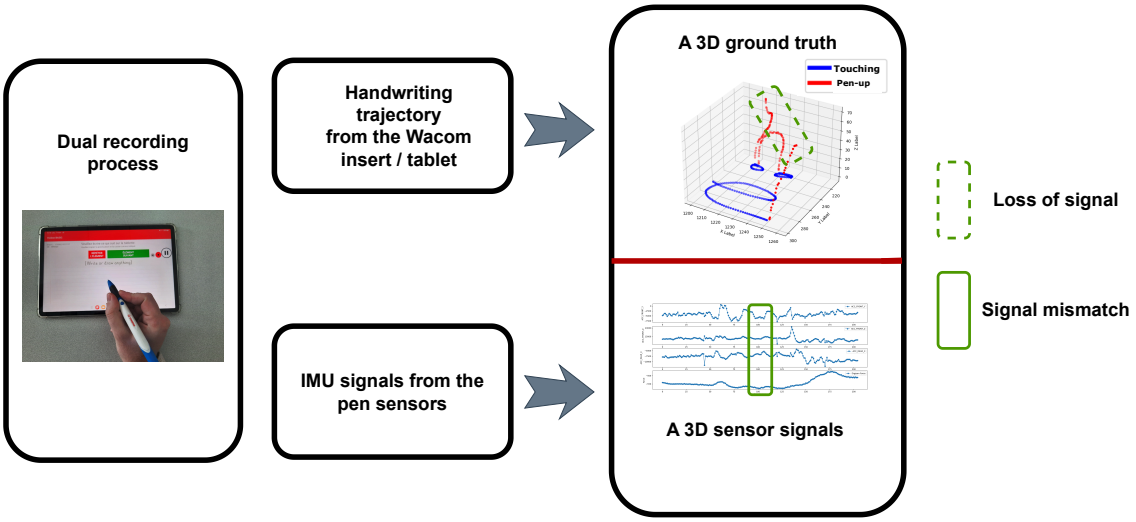


Figure 2.1 – Dual recording process, with a Wacom insert, to enable the acquisition of ground truth

In addition, the way the pen is held by the user adds another layer of complexity. The sensor data, including pressure, angle and speed, can vary greatly depending on the user's grip and style of movement. This variability requires sophisticated algorithms that can account for these differences in order to accurately interpret the pen's intended trajectory and translate it into digital input.

2.3 Deep learning training challenges

To train deep learning models, it is necessary to have an input data / ground truth pair, which implies having a Digipen tablet acquisition process. Furthermore, the alignment of signal data with a ground truth is crucial. This alignment ensures that the input features are correctly mapped to the target outputs, facilitating accurate predictions and model performance.

However, the most critical factor is data availability. Without sufficient data, the model's performance and generalization will be limited, necessitating the design of an adapted architecture that can work with the constrained dataset.

2.4 Conclusion

Reconstructing handwritten text from kinematic signals is an intricate endeavor that intertwines multiple fields including biomechanics, signal processing, and machine learning. The process is fraught with challenges arising from both the technology used and the inherent variability in human writing.

Pen related challenges involve dealing with noisy and variable sensor data, particularly from IMU, which suffer from drift and accuracy issues. Synchronization of data and the handling of transmission delays add additional layers of complexity.

Writer related challenges stem from the diversity in individual writing styles and techniques. The distinction between the act of writing and the pen-up movements process requires consideration. Variability between different writers, including age related differences in handwriting dynamics and the effect of the user's style, presents additional obstacles that must be addressed.

Deep learning training challenges highlight the need for a robust dataset that aligns input data with ground truth. The availability of sufficient and well aligned data is critical for training accurate models. With limited data, specialized architectures and techniques must be employed to ensure the model's performance and generalization capabilities.

Addressing these challenges requires a multidisciplinary approach, incorporating signal processing techniques, and machine learning algorithms.

Thesis contributions

The objective of this thesis is to design an efficient machine learning method based on neural networks for reconstructing handwriting using data from kinematic sensors. Our approach follows a constructive methodology, beginning with data collection. Initially, we introduce a processing pipeline to handle touch strokes. This is followed by the implementation of a Mixture-of-Experts model to process both touching and pen-up inputs. Finally, we explore domain adaptation techniques to address variations in data from children and different surfaces (tablet and paper).

First contribution: Data collection

In reconstructing handwriting trajectories, achieving precision at every step is crucial, underscoring the need to start with the cleanest possible data before training a neural

network. Initially, we faced the challenge of not having access to the necessary data, prompting us to prioritize data collection. Consequently we made 2 datasets publicly accessible [1], to facilitate research and development in this area. In order to collect this dataset, a data acquisition protocol was established to enable the simultaneous acquisition of Digipen sensor data and Wacom trajectory data, which serves as ground truth.

Second contribution: A first preprocessing chain

Reconstructing handwriting trajectories demands high precision throughout the process. Hence, our initial task involved establishing a comprehensive preprocessing pipeline, tailored specifically to address the complexities of IMU data. The key preprocessing step for training a deep neural network consists in aligning ground truth and sensor data to obtain sequences of identical size.

Third contribution: A model dedicated to touching part

In our endeavor to reconstruct handwriting trajectories, we are initially focusing on the touching strokes, as they represent the simplest component of the signal to process. This strategic decision allows us to refine our methods on a less complex aspect of handwriting, setting a solid foundation for more comprehensive analysis in later stages. To effectively process these touching strokes while preserving the crucial temporal context, we propose utilizing a Temporal Convolutional Network (TCN) architecture. This approach is designed to maintain the sequential integrity of the handwriting signal, facilitating a more accurate reconstruction of the handwriting.

Fourth contribution: A new mixture of expert models approach.

Our latest approach in handwriting trajectory reconstruction focus on the reconstruction of touching parts, where we have access to reliable ground truth data during training. Notably, while touching trajectories offer 2D data that can be accurately captured, pen-up trajectories present a more complex 3D challenge, with ground truth signals only discernible within a 7mm height. Our work underscores the critical role of dynamics in accurately reconstructing trajectories from IMU data. In this new phase, we aim to broaden our approach to include both touching and pen-up aspects of handwriting. By introducing two expert networks, we intend to tackle the distinct characteristics of IMU signals,

1. `https://www-shadoc.irisa.fr/irisa-kiht-s-and-kiht-public-datasets/`

leveraging the successful outcomes from our prior focus on pencil touching trajectories to inform a more holistic strategy in capturing the nuanced dynamics of handwriting.

Fifth contribution: Domain adaptation method to process children data

The transition from reconstructing adult handwriting captured on tablets to child handwriting necessitates a methodological approach. This expansion is crucial due to the significant variances in motor skills and stylus or pen pressure between adults and children. These variations introduce complexities in signal processing and pattern recognition, highlighting the need for domain adaptation algorithms that can accurately interpret and reconstruct the diverse range of handwriting styles.

Manuscript organization

The thesis is organized as follows. The first part presents the state of the art in handwriting reconstruction (Part: I), including an overview of the different neural networks that can be used to process time series. This section sets the foundation by exploring existing literature, methodologies, and the challenges associated with handwriting reconstruction. It begins with a comprehensive review of deep learning and sequence-to-sequence models (Chapter 3), emphasizing their relevance to the task at hand. This is followed by an evaluation of various handwriting reconstruction methods (Chapter 4), highlighting the criteria and metrics used to assess their effectiveness. We then focus on the method dedicated to handwriting reconstruction (Chapter 5), with a particular focus on methods using the Digipen, which will serve as a baseline for our work. The section concludes with a discussion on domain adaptation for handwriting data from IMU sensors (Chapter 6), addressing the complexities of transitioning data from one domain to another.

The second part is dedicated to our contributions (Part: II). It starts with detailing the data choice and acquisition protocol (Chapter 7), explaining the selection criteria and the steps taken to ensure data quality. The next chapter introduces the first complete preprocessing chain developed for handwriting data (Chapter 8), showcasing the steps involved and the impact on reconstruction quality. Following this, we present TCN based model dedicated to handling touching strokes, known as the Touching Expert Model (TEM) (Chapter 9). Subsequently, we describe our mixture-of-experts approach (Chapter

10), designed to achieve the best reconstructions for each part of the writing process through collaborative strategies and task specification. The next part of the thesis focuses on our domain adaptation methods (Chapter 11), focusing on transitioning from adult data to child handwriting. Experimental results will be presented as they come in, to provide a clearer understanding of the problems and the solutions we have proposed.

Finally, the manuscript concludes with a general conclusion (Chapter 11.5) that highlights the key contributions of the research, discusses the broader implications of the findings, and suggests potential directions for future work.

Part I

State of the art

Few studies have focused on reconstructing handwritten text from Inertial mesurement unit (IMU) data. Consequently, in this review of the state-of-the-art, we will first introduce deep learning techniques specifically designed for processing time series data. This will be followed by a presentation of the evaluation metrics commonly used to assess and compare online handwriting reconstruction methods. Subsequently, we will provide an overview of various approaches that have been proposed for handwriting reconstruction. Finally, the last part of this state-of-the-art review is dedicated to domain adaptation methods for handwriting reconstruction.

Chapter 3

Deep learning and seqtoseq

Advances in the field of deep learning, particularly in processing sequential data, have been significantly driven by the development of various neural network architectures. Each architecture possesses unique features suited to different types of data and tasks. In this chapter, we explore three main categories: Recurrent Neural Networks (RNN and its variants), Convolutional Neural Networks (CNN and TCN), and Transformers, focusing on their fundamental principles, advantages, and disadvantages in relation to the challenges of this thesis.

3.1 Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNN) are specifically designed to handle sequential data. They use recurring connections to propagate information over time (Fig. 3.1), allowing the network to maintain a state or memory of previous inputs. This capability is crucial for processing sequences where context and the order of observations are important. Enhancements such as Long Short-Term Memory (LSTM) [Hochreiter et al., 1997] and Gated Recurrent Units (GRU) [Chung et al., 2014] have been developed to address the problem of vanishing gradients, which makes learning long-term dependencies difficult for vanilla RNNs. LSTM consists of memory cells that maintain a cell state, along with

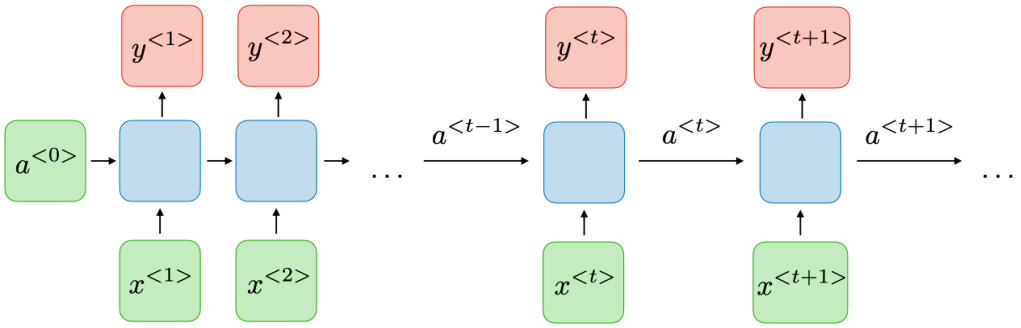


Figure 3.1 – Representation of RNN, visual from stanford.edu

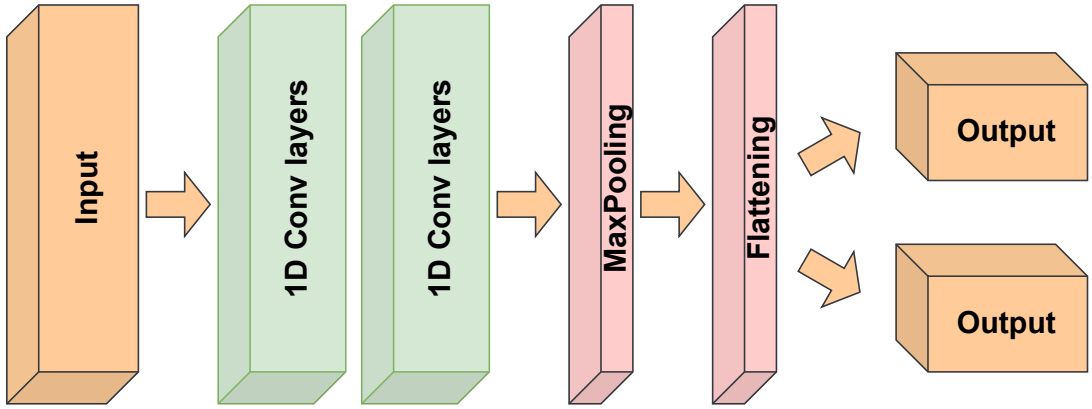


Figure 3.2 – 1D CNN Network architecture from [Pathour et al., Apr. 2023]

gates (input, forget, and output gates) that regulate the flow of information, allowing the network to selectively remember or forget information over long sequences. GRUs simplify the LSTM architecture by combining the forget and input gates into a single update gate and using a reset gate to control the flow of information. These advanced architectures are capable of capturing long-distance information in the data, making them ideal for tasks like language modeling and speech recognition. However, they face challenges, LSTMs and GRUs have more parameters than vanilla recurrent neurons due to their architectures involving multiple gates. This increased number of parameters leads to higher computational complexity, requiring more memory and processing power. Consequently, training LSTMs and GRUs often necessitates larger datasets to effectively learn patterns and dependencies. They have also difficulty in processing very long sequences, and speed limitations due to sequential calculations.

3.2 Convolutional Neural Networks (CNN)

On the other hand, Convolutional Neural Networks (CNN) utilize convolutional filters to process data. Primarily known for their application in image processing, where they excel due to their ability to extract significant patterns and features from spatial data, CNN are also suited for other types of structured data. Pooling layers are frequently utilized to decrease the spatial dimensions of feature maps, effectively compressing the data while preserving essential features (Fig. 3.2).

The advantages of CNN include their ability to reduce the number of parameters through the use of shared weights and their translational invariance, making them robust to variations in data positioning. The effectiveness of these designs in capturing long-term dependencies is potentially limited by their focus on local processing. This limitation arises from their specified filter size and receptive field, and their lack of internal memory makes them less suitable for processing temporal sequences.

Causal convolutions are a type of convolution designed for temporal data, ensuring that the model respects the chronological order of the data. Specifically, this means that the prediction $h_t = f(x_0, \ldots, x_t))$ made by the model at timestep t is only based on the current and past timesteps $x_1, \ldots, x_t$ and does not depend on any future timesteps $x_{t+1}, \ldots, x_T$.

3.3 Temporal Convolutional Networks (TCN)

Temporal Convolutional Networks (TCN) are a type of neural network designed for processing sequential data and can be either causal or non-causal. Causal TCN ensure that the output at any given time step depends only on the current and past inputs, making them suitable for real time applications where future data is not available. Non-causal TCN, on the other hand, allow the output to depend on both past and future inputs, making them ideal for offline applications where the entire sequence is known in advance. Both variants use dilated convolutions to efficiently increase the receptive field without a proportional increase in computational complexity. A notable example of a TCN is WaveNet [Oord et al., 2016], which uses dilated causal convolutions for effective modeling of audio waveforms (Fig. 3.3).

TCN offer several advantages, such as the ability to process variable-length sequences and perform parallel computations, which speeds up training. They are particularly adapted to capture long-term dependencies [Ehteram et al., 2024]. However, unlike RNN, they do not maintain an internal state for sequences, which can be a drawback for some applications requiring memory of past events.

3.4 Transformers

Finally, the architecture of Transformers [Vaswani et al., 2017] represents a significant advancement in sequence processing through its attention mechanism.

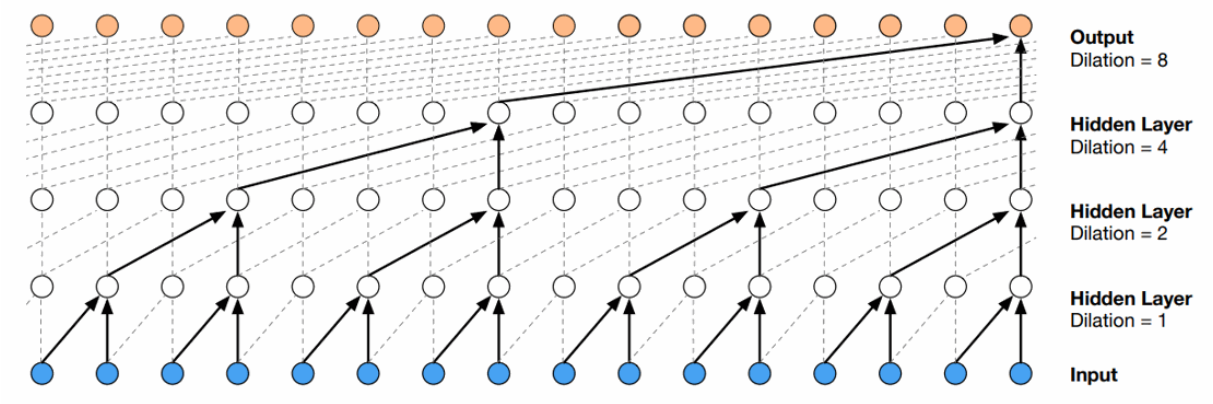


Figure 3.3 – Visualization of stack dilated causal convolution from [Oord et al., 2016]

It relies on attention mechanisms, where the model evaluates the relevance of each elements in a sequences with respect to every other element, enabling it to capture complex dependencies and contextual meanings. Transformers use multi-head attention to process multiple pieces of information in parallel, enhancing the model's ability to handle large-scale tasks efficiently. This allows models to differentially weigh parts of a sequence, thereby facilitating the capture of long-range relationships between elements. Transformers offer advantages such as parallelism efficiency, which significantly speeds up training. In addition, they mitigate the effects of vanishing gradients during backpropagation, which only occur over the network's depth rather than over time as seen with RNNs, as the samples are not processed sequentially, but rather in parallel. Due to their many parameters and the attention mechanism, they are prone to overfitting and therefore require a large amount of training data.

3.5 Conclusion on deep learning methods

In conclusion, our framework for reconstructing writing from IMU sensors must address challenges such as handling long sequences, time-series data, and noise. Recurrent Neural Networks (RNN) often struggle with vanishing gradients in long sequences, and Transformers may be less effective when dealing with smaller datasets. Convolutional Neural Networks (CNN) also tend to fall short in capturing the essential temporal dependencies of time-series data. In contrast, Temporal Convolutional Networks (TCN) perform well in these areas by effectively modeling long-range dependencies and leverag-

ing the strengths of convolutional methods. In the context of the thesis, given the limited amount of data, transformers models do not seem to be the best solution, and RNNs do not seem to be the best option due to the problem of vanishing gradient, which is why convolutional networks will be studied in this thesis.

Chapter 4

Evaluation of Handwriting Reconstruction

Having relevant metrics, whether for evaluation or as a loss function, is a key element in the reconstruction and evaluation of handwriting, which is why we now present the main metrics in the field.

Evaluation methods vary across studies, with some relying on qualitative assessments of reconstructed trajectories [Nguyen et al., 2021], while others use recognition performance as a proxy for evaluation [Huang et al., 2022]. Metrics like Root Mean Squared Error (RMSE), Dynamic Time Warping (DTW) and Fréchet distance are also employed in certain cases to assess reconstruction accuracy [Chen et al., 2022]. The diverse approaches underscore the complexity of evaluating handwriting trajectory reconstruction from digital stylus inputs.

Note that when evaluating the reconstruction of a trajectory, several criteria must be taken into account: shape, direction and dynamics. Here, we focus on shape and direction, with a view to adding the evaluation of dynamics. In fact, we want a faithful reconstruction of the shape to produce feedback on the quality of the layout. We will take a more in depth look at these three metrics.

The MSE between two multivariate time series $A \in \mathbb{R}^{T_A \times z}$, and $B \in \mathbb{R}^{T_B \times z}$ of equal feature dimensionality z and have an equal lengths $T_A = T_B$ is:

$$MSE(A, B) = \frac{1}{T_A} \sum_{i=1}^{n} \|a_i - b_i\|^2 \tag{4.1}$$

The DTW algorithm ([Sakoe et al., 1978]) is based on dynamic programming to assess the similarity between time series. For two multivariate time series $A \in \mathbb{R}^{T_A \times z}$, and $B \in \mathbb{R}^{T_B \times z}$ of equal feature dimensionality z and respective lengths T_A and T_B, the DTW is written as follows:

$$DTW(A, B) = \min_{\delta \in E(A,B)} \sum_{(i,j) \in \delta} d(a_i, b_j) \tag{4.2}$$

where $E(A, B)$ represents the set of all admissible alignments between A and B and d is a distance metric in $\mathbb{R}^z$. Commonly, the squared Euclidean distance $d(a_i, b_j) = \|a_i - b_j\|^2$ is used.

Several variants of this metric exist. [Mohamed Moussa et al., 2023] proposed DTWseg, an adaptation of the Dynamic Time Warping (DTW) algorithm, designed to improve the matching of online handwriting signals. Instead of the traditional point-to-point Euclidean distance, DTWseg employs a point-to-segment distance metric, where each segment corresponds to a stroke. This adjustment greatly reduces sensitivity to variations in signal sampling rates, which are often caused by differences in acquisition frequencies or writing speeds. Consequently, it eliminates the need for resampling, which can overlook important dynamic characteristics of handwriting.

The Fréchet distance, for two multivariate time series $A \in \mathbb{R}^{T_A \times z}$, and $B \in \mathbb{R}^{T_B \times z}$ of equal feature dimensionality z and respective lengths T_A and T_B is defined by the following equation:

$$F(A, B) = \min_{\delta \in E(A,B)} \max_{(i,j) \in \delta} d(a_i, b_j) \tag{4.3}$$

More concretely (Fig. 4.1), MSE is a straightforward metric that computes the average squared difference between corresponding elements in two sequences. It does not account for any time shifts or sequence order variations, and is sensitive to outliers. The DTW distance is the minimum cumulative distance between the sequences when mapped onto each other. DTW consider time shifts by finding an optimal alignment. The Fréchet distance between two curves measures the minimum distance required for two points, each moving along their respective curves, to simultaneously traverse their paths from start to finish while staying as close as possible to each other. All these metrics are sensitive to outliers. DTW and Fréchet are computationally intensive algorithms.

It's also important to remember, depending on what you want to measure, that resampling, normalization, and recentering are preprocessing steps that significantly impact theses three metrics (DTW, MSE, Fréchet Distance). Resampling standardizes the temporal scale of data sequences, making them more comparable, especially beneficial for DTW. Normalization scales data into a uniform range, enhancing the comparability and meaningfulness of all three metrics by focusing on patterns and shapes rather than magnitude. Recentering impact is more specific to the application context but generally recentering can help to align sequences. Together, these preprocessing steps can improve the accuracy and relevance of measurements for a given objective.

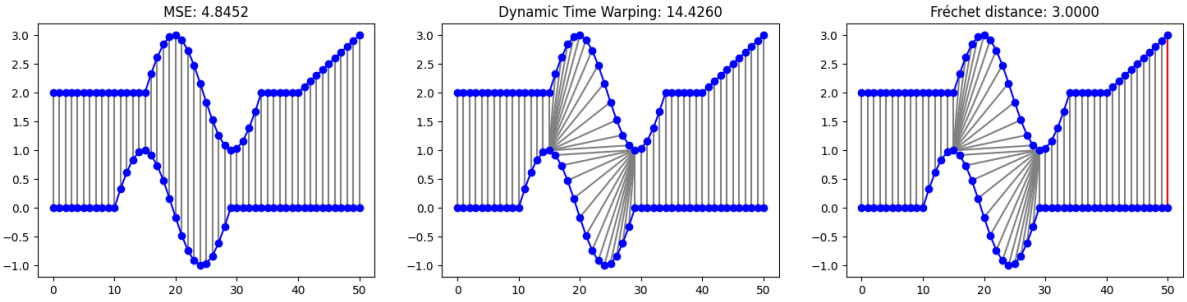


Figure 4.1 – Comparison between Euclidean distance, DTW and Fréchet distance. Note that, for visualization, time series are shifted vertically, but one should imagine that feature value ranges (y-axis values) match.

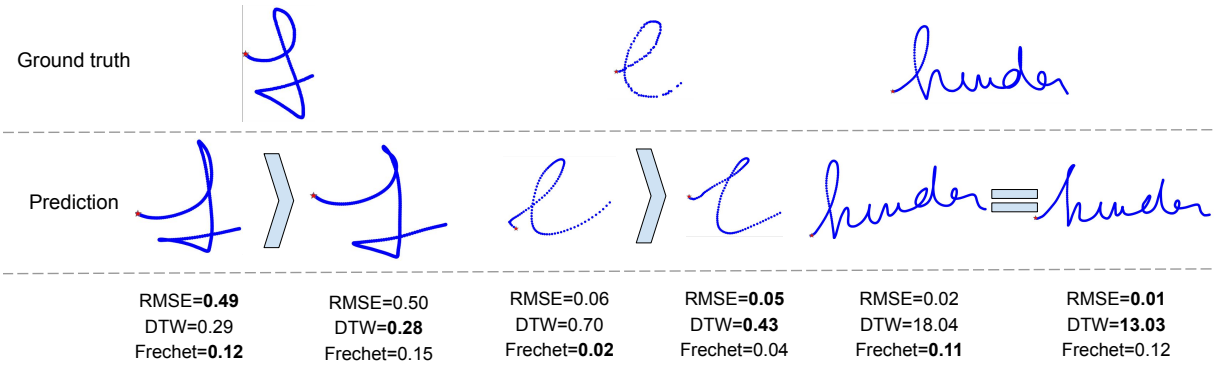


Figure 4.2 – Metrics comparison. On the top, the ground truth and on the bottom two predictions coming from different models.

Appropriate assessment metrics are needed to evaluate trajectory reconstruction to provide the end-user with useful feedback. Fréchet distance doesn't disagree with visual evaluation like DTW and RMSE. Figure 4.2 indicates that the left upper loop of the f is better reconstructed. To confirm this, we conducted several experiments with qualitative verification. The conclusion was that the Fréchet distance appeared more robust and closer to human perception.

Chapter 5

Handwriting reconstruction

The field of digital devices for note-taking, drawing, and handwriting learning is rapidly expanding. However, most systems use display-based interfaces for digital handwriting acquisition, with only a few incorporating styluses equipped with motion tracking systems to reconstruct handwriting trajectories. These conventional approaches are often limited by the physical constraints of a screen or tablet, which can hinder user flexibility.

In contrast, research has been exploring alternative methods for handwriting trajectory reconstruction using Inertial Measurement Unit (IMU) sensors. IMU are small devices that track changes in their orientation and motion using a combination of accelerometers and gyroscopes. This technology has the potential to enable surface-free digital handwriting, where users can write with any device or object, without being confined to a specific display-based interface.

In this context, this thesis aims to push the boundaries of conventional digital note-taking systems by developing a surface-free handwriting reconstruction pen using IMU sensors (the Digipen).

5.1 Handwriting Trajectory Reconstruction

The field of handwriting trajectory reconstruction can be broadly categorized into two main approaches:

- active surface (tablet), to collect a digital ink;
- passive surface, another medium is needed for the handwriting reconstruction:
  - reconstruction from pen-tip optical tracking systems;
  - reconstruction from offline handwriting images;
  - reconstruction from IMU signals.

We present these approaches in the following sections.

5.1.1 Pen-based digital tablets

In the realm of digitizing handwritten text, various methodologies have emerged, each leveraging distinct technologies to accurately capture the intricacies and personal nuances of human handwriting. Among the most prevalent methods are the use of pen-based digital tablets. These devices, including those compatible with the Samsung S Pen, Apple Pencil, Microsoft Surface Pen, and Wacom styluses, are recognized for their precision and responsiveness. The core of these devices is Electro-Magnetic Resonance (EMR) technology for S pen and Wacom, whereas Apple and Microsoft use capacitive sensing coupled with sensors in their pens. However, while these devices offer high precision and ease of use, they often come with a significant limitation: they tie users to specific hardware ecosystems.

5.1.2 Handwriting trajectory reconstruction using information from a camera-based system

By utilizing the visual information captured by cameras, it is possible to reconstruct the dynamic path of a pen or stylus during handwriting. This section delves into the methodologies and challenges associated with this process. The Anoto pen [Liao et al., Jan. 2008], enables writing on paper while simultaneously capturing digital ink. This system utilizes a specialized pen and paper duo, where the paper is embedded with a complex, invisible dot pattern. This pattern allows the accompanying digital pen to accurately determine its position on the sheet. The pen, a conventional ink-based writing tool to the eye, houses a tiny camera and processor under its hood. These components work in tandem to capture hundreds of images per second, deciphering the pen's movements and translating them into digital text or illustrations stored within the pen. However, while these devices can synthesize handwriting, they have two major limitations: the need to use special paper and an often expensive pen.

[Ott et al., 2022] combine IMU with a camera to reconstruct the writing trajectory. They propose a novel approach to classifying and reconstructing trajectories in online handwriting recognition using a multi-task learning framework. The authors propose a neural network architecture that integrates the learning of both classification tasks, such as character recognition, and trajectory regression tasks, using IMU and camera-based data inputs. Their proposed approach utilizes multi-task learning (MTL) (Fig. 5.1), which exploits the differences and commonalities across the two tasks (classification and

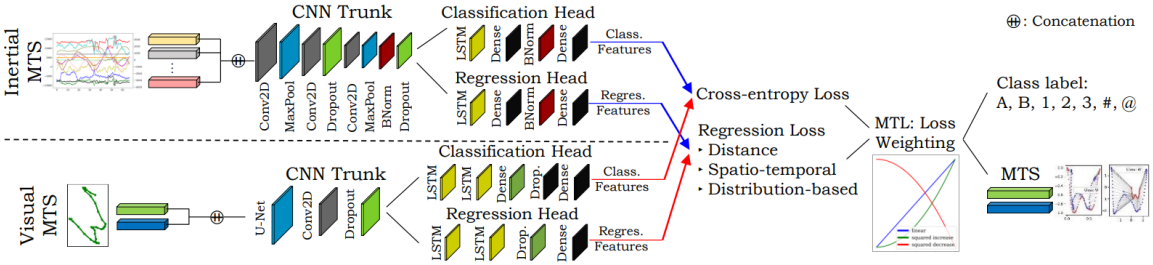


Figure 5.1 – [Ott et al., 2022] multi-task learning (MTL) approach using convolutional neural networks (CNN) for processing inertial measurement unit (IMU) and visual datasets. The architecture features a shared CNN trunk with separate heads for classification and trajectory prediction.

trajectory regression) to improve performance on each. The architecture is a 2-branch CNN, where the 2 inputs are video and IMU signals. They combine cross-entropy loss for classification with distance and similarity losses for trajectory regression. By doing so, it achieves notable improvements in handwriting recognition tasks, reducing errors and variance in trajectory prediction while also enhancing character classification accuracy. In the regression section, the study compares several loss functions (MSE, Andrew's Sine, Huber, Pearson Correlation, Cosine Similarity, MSE + Pearson Correlation, MSE + Cosine Similarity, and MSE + Wasserstein). Among these, only MSE, Andrew's Sine, Huber, and MSE + Pearson Correlation yield satisfactory and comparable results. The regression branch comprises three convolutional layers followed by an LSTM layer, demonstrating that a relatively small architecture can achieve promising outcomes.

5.1.3 Handwriting trajectory reconstruction from offline handwriting images

Some work has been done on the transition from offline to online trajectories to reintroduce documents into an online flow. This work is particularly interesting in terms of the metrics used to evaluate online reconstruction. Let's explore a few methods.

[Chen et al., 2022] propose an approach for generating online handwriting trajectories from offline images. Traditional evaluation metrics only considered writing order, neglecting glyph fidelity. To address this, [Chen et al., 2022] introduced two new metrics: Adaptive Intersection on Union (AIoU) for assessing glyph fidelity by eliminating stroke width influences, and Length-Independent Dynamic Time Warping (LDTW) for align-

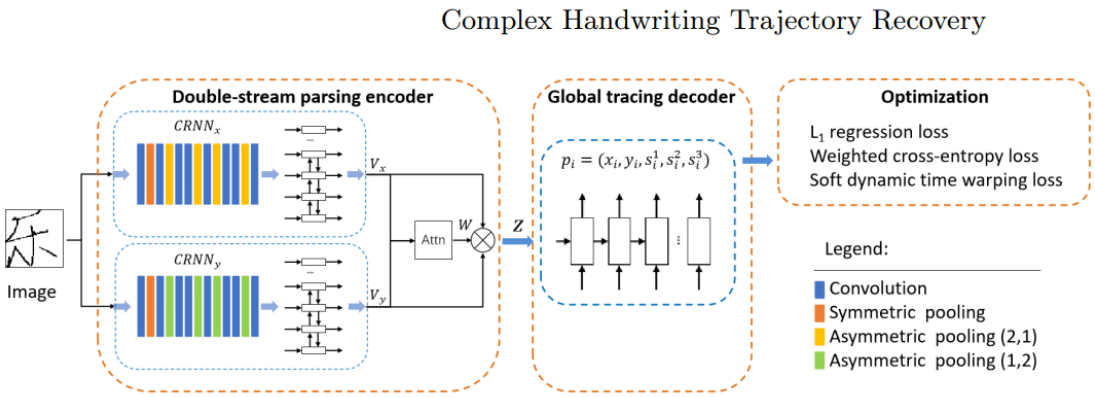


Figure 5.2 – [Chen et al., 2022] method for a given static handwriting image, the double-stream parsing encoder examines the stroke context and interprets the glyph structure, extracting features that the global tracing decoder will use to predict trajectory points.

ing handwriting trajectories of varying lengths. They propose a double-stream parsing encoder for glyph structure analysis and a global tracing decoder to predict trajectory points (Fig. 5.2). Using DTW or its variations as a loss function introduces a major disadvantage which is the calculation cost. The computational cost of DTW is significant because it requires calculating distances between all possible pairs of points between the two sequences, leading to a quadratic time complexity. This can become particularly burdensome with longer sequences, making the training of models excessively slow and computationally expensive.

To retrieve online handwriting from offline, [Mohamed Moussa et al., 2023] introduces DTWseg which was previously described (in chapter 4). The paper establishes a new benchmark for evaluating state-of-the-art methods in offline to online handwriting conversion using this innovative metric, underscoring DTWseg's potential in providing more accurate and meaningful comparisons for online handwriting signals.

[Zhu et al., 2024], takes Chinese handwritten images as its input. Initially, a feature map is generated through a spatial encoder (a ResNet). This feature map is then processed by the decoding network, which utilizes both the temporal information from a GRU and the spatial features from the encoder, along with stroke features. The decoding network outputs a heatmap. This heatmap serves two purposes: predicting the state at the current

moment and forecasting the handwriting point for the subsequent moment. The loss used is the MSE.

An important lesson from this section is the frequent use of Dynamic Time Warping (DTW) in handwriting reconstruction problems. This makes DTW and its variants a natural choice, especially for matching sensor signals to ground truth data.

5.1.4 Handwriting trajectory reconstruction from Inertial Measurement Unit (IMU) sensors

Specialized hardware, such as Inertial Measurement Unit (IMU) sensors, allows for the digitization of handwriting on conventional paper by recording writing movements and producing a digital trace. Such sensors are installed in the Digipen pen we are working with in our project. In this section, we will explore various methods for reconstructing trajectories using IMUs. We will begin with traditional methods and then move on to those that incorporate deep learning techniques. Then, we will focus on methods utilizing the Stabilo Digipen. Finally, we will discuss approaches that deal with recognition rather than reconstruction. Although recognition is a different and generally easier task, these approaches and models can still provide valuable insights for our work.

However, there is limited research focusing on IMU-based trajectory reconstruction. For instance, in the work by [Pan et al., 2016], traditional approaches are employed (Fig. 5.3). This method involves signal preprocessing for trajectory reconstruction, utilizing the integration of linear discriminant analysis (LDA) for movement detection. The trajectory is reconstructed from the last known point and speed.

[Miyagawa et al., 2000] utilized a pen equipped with an accelerometer and a gyroscope to capture triaxial linear acceleration and angular velocity. These measurements are initially processed through low-pass filters to eliminate unwanted frequency components. Subsequently, the linear acceleration and angular velocity are adjusted using a coordinate transformation matrix to account for the pen's orientation during writing. The pen's position is derived by doubly integrating the filtered and transformed accelerometer data (Fig. 5.4). This process calculates the pen's displacement, which is essential for reconstructing the written characters. The reconstructed characters are then visually represented in various combinations of the three-dimensional axes provided by the sensors, such as the x-y, x-z, and y-z combinations.

The IMUPEN, a device integrating an accelerometer and two gyroscopes, employs a methodology similar to previous techniques for digitizing handwriting [Wang et al., 2009].

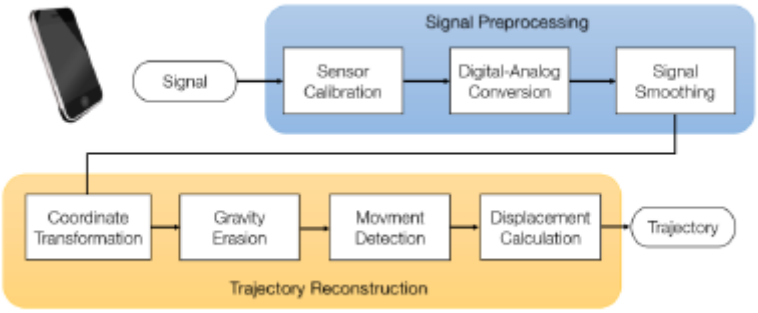


Figure 5.3 – [Pan et al., 2016] pipeline for trajectory reconstruction, where the trajectory is mathematically reconstructed from the last known point and speed after some preprocessing steps.

Initially, the recorded signals are calibrated to adjust for any sensor offsets. Subsequently, a moving filter is applied to mitigate high-frequency noise from the data, ensuring that only relevant signal components are preserved. To accurately represent the orientation of the pen during use, quaternions are utilized. These quaternions are then transformed to a reference frame using a coordinate transformation matrix, aligning the data with a standardized orientation. Position estimation is subsequently computed through integral methods, which calculate the cumulative path of the pen based on the processed sensor data. This refined data is then used for the task of recognizing handwritten digits (Fig. 5.5).

[Bu et al., 2021] attached an inertial measurement unit (IMU) to the pen's tail, the system captures handwriting movements (Fig. 5.6). They extract the motion segments from IMU readings based on the short time energy (STE). Principal component analysis (PCA) is used to detect the writing plane, distinguishing on-plane writing from off-plane movements to capture effective handwriting during continuous writing processes. The IMU displacement is then calculated by double integration of the global linear acceleration.

[Liu et al., 2020] explored the use of magnetic signals for handwriting recognition. They began by segmenting the continuous 3D magnetometer readings into smaller sections, with each section corresponding to an individual handwritten letter. After segmentation, each segment of magnetometer readings is processed independently, converting them into their equivalent handwriting trajectories. This conversion is achieved through projection and coordinate transformation.

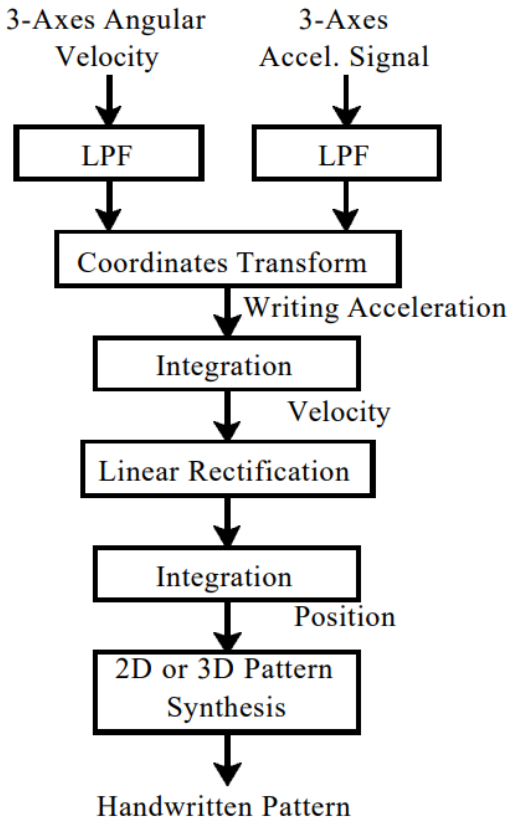


Figure 5.4 – [Miyagawa et al., 2000] pipeline for trajectory reconstruction. Based on a low-pass filter to denoise the signal. Followed by double integration to pass from acceleration to trajectory.

The Digipen [Harbaum et al., 2024] utilize IMU sensors, enhancing versatility by allowing users to write on various surfaces, including tablets, paper, or boards. Nonetheless, IMU based systems primarily track relative pen displacements, which may introduce inaccuracies due to the inherent noise in sensor data. During this thesis it will be our object of study. Let's focus now on [Wehbi et al., 2022] Digipen approach (Fig. 5.7), which we will then use as a benchmark for our approach.

This approach involves mapping movements captured by an IMU enhanced digital pen into relative displacement data, trained using a convolutional neural network (CNN) (5.8). It starts with a 1D convolution layer. Followed by a batch normalization layer. Next, a dropout layer is applied. This sequence of a 1D convolution, batch normalization, and dropout is repeated three times. The network finish with a TimeDistributed fully

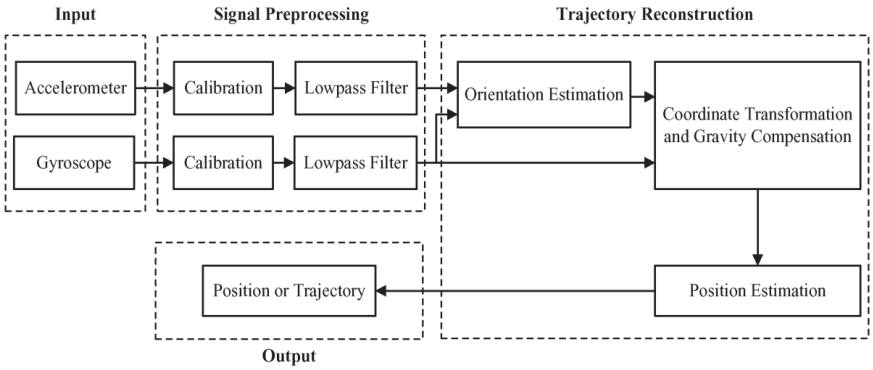


Figure 5.5 – [Wang et al., 2009] pipeline for trajectory reconstruction. The main steps are: First, apply a low-pass filter to denoise the signal. Then, double integration is performed to convert the acceleration data into trajectory information.

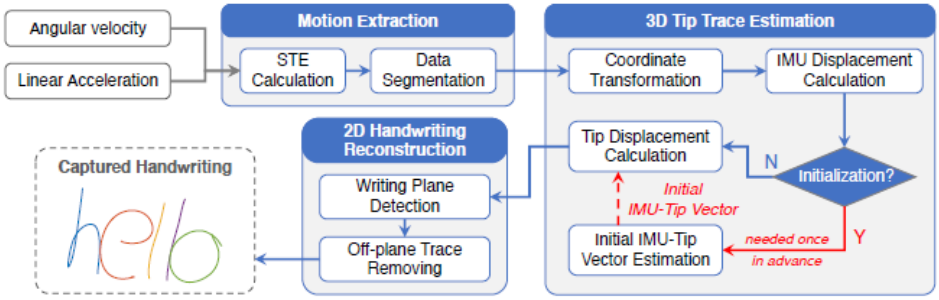


Figure 5.6 – [Bu et al., 2021] pipeline for trajectory reconstruction. IMU is read and STE used to extract motion segments. PCA distinguishes on-plane writing from off-plane movements. IMU displacement is calculated by integrating global linear acceleration twice.

connected layer, which allows the network to maintain the temporal sequence information while applying the same fully connected operation to every time step.

A special aspect of their work is to minimize preprocessing, segmentation, or post-processing alignment during inference. To ensure that the sensor signals and ground truth data have identical sizes, linear interpolation is applied. This involves adding points to the ground truth to match the number of sensor data points. Once the number of points is equal, all the points are distributed linearly along the trajectory. This manipulation has the effect of breaking the trajectory dynamics. Indeed, an linear interpolation breaks the dynamics of the trajectory by artificially smoothing the data, eliminating fluctuations and temporal relationships, which can lead to significant discrepancies and potential errors in

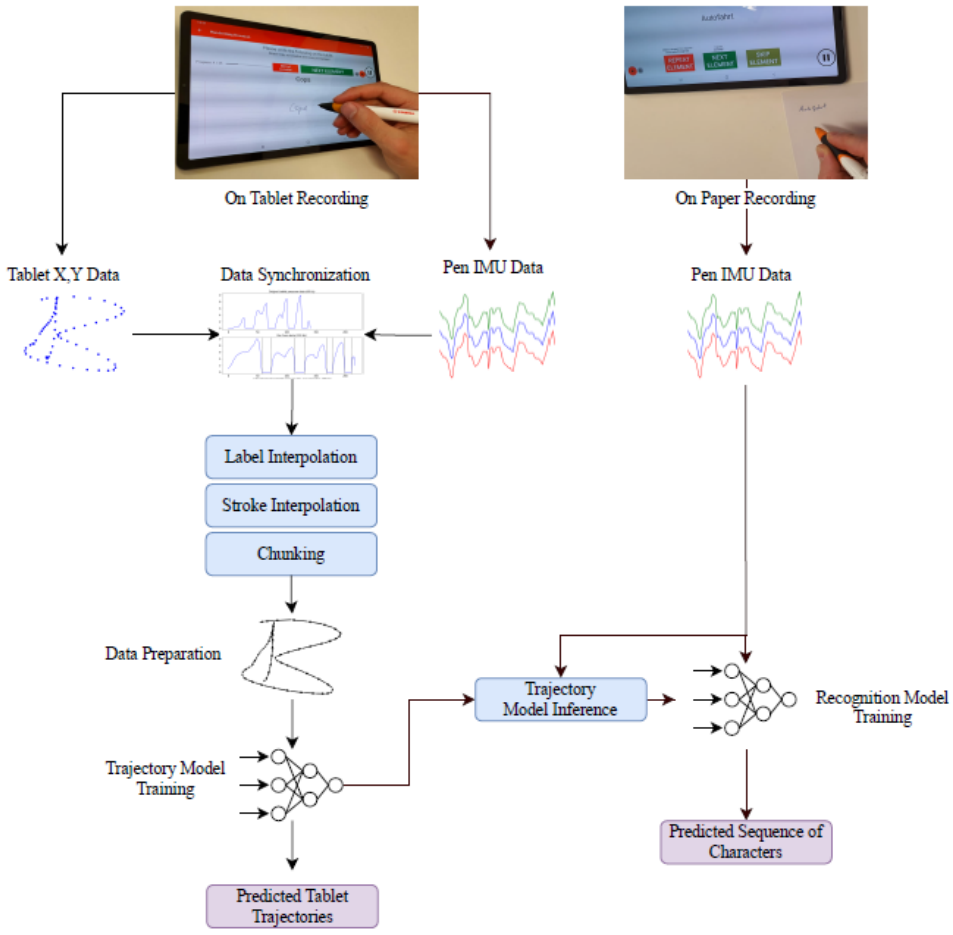


Figure 5.7 – [Wehbi et al., 2022] pipeline for trajectory reconstruction. The preprocessing steps are label interpolation, stroke interpolation, and chunking to prepare the data for training a trajectory model (CNN) that predicts handwriting trajectories.

the representation of real motion. For evaluation, the output trajectories were normalized using min–max scaling to evaluate the predictions on a unit scale. Root mean squared error (RMSE) was calculated to evaluate the similarity of the reconstructed trajectories in comparison to the ground truth from the original ones.

While few works focus on handwriting reconstruction from IMU sensors, online handwriting recognition from IMU sensor data has been explored using pen-tip trajectory signals and IMU sensor signals. It should be noted that, although the two objectives may seem close, they are different in nature. The state-of-the-art shows that recognizing

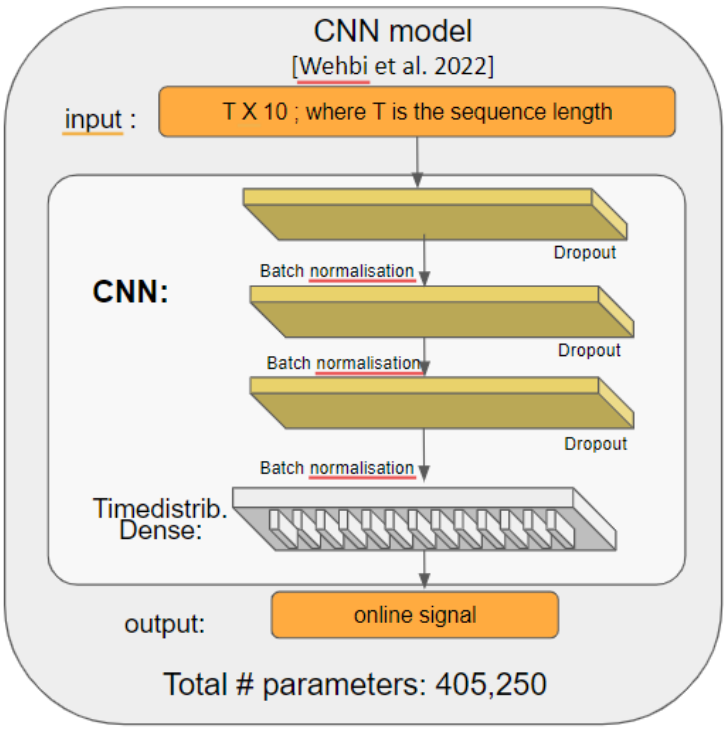


Figure 5.8 – CNN Architecture proposed by [Wehbi et al., 2022].

from an IMU signal is a less complex task than reconstructing a precise trajectory. Handwriting reconstruction focuses on the replication of the pen's exact path during writing, necessitating precise trajectory data. On the other hand, handwriting recognition aims to identify the characters or symbols penned, relying on interpreting representative global trace shapes rather than mimicking the motion.

A benchmark study by [Ott et al., 2022] compared various neural network architectures (CNN, LSTM, BILSTM, and transformers) for recognizing characters, symbols, words, and equations from IMU sensor data. Their work identifies the combination of Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BLSTM) as yielding the best results. However, the performance is not significantly different from that achieved by the other approach. The IAM-OnDB dataset is a widely used handwritten text dataset that consists of forms of handwritten English text, commonly employed for training and evaluating handwriting recognition systems. In terms of performance these methods have a character error rate on IAM-OnDB is about 10 % on average, which means that the problem is generally well understood. The conclusion is that the main

difficulty is not the choice of architecture, but understanding and processing the input signals.

Explorations into simultaneous online handwriting trajectory reconstruction and character recognition are also prevalent. In the context of the UbiComp 2021 Challenge, [Wegmeth et al., 2021] utilized a CNN/LSTM framework for recognizing mathematical expressions written with the Stabilo's Digipen, emphasizing the precision of label boundary identification. Their network is composed of 4 convolution layers followed by 5 LSTM layers.

Further studies have sought to optimize the data transmission efficiency between the Digipen and remote devices, as seen in [Kreß et al., 2022], or to adapt the model to different domains, as discussed in works by [Klaß et al., 2022]. The pursuit of explainability in the models, as undertaken by [Azimi et al., 2022], underscores the ongoing efforts to enhance interpretability within this domain.

A recent advancement [Wang et al., 2024] involves leveraging low-cost IMU sensors from smartphones for the recognition of 62 characters. This approach uniquely combines dynamic inertial signals with trajectory morphology, derived from the handwriting image, subsequently employing a CNN for classification. This innovation represents a stride towards accessible and efficient handwriting recognition, bridging inertial dynamics with visual traits to enhance model performance.

5.2 Discussions and conclusions

The state of the art in handwriting reconstruction has made significant strides, particularly in utilizing diverse methodologies for capturing and reconstructing handwritten text. Notable advancements include pen-based digital tablets, camera-based systems, offline handwriting, and IMU sensors. Each system offers distinct strengths and weaknesses that influence their practicality and accuracy.

System with active surface (tablet), provide high precision and a natural writing experience, but their reliance on specific hardware ecosystems limits their flexibility and increases costs. Camera-based systems, such as those using specialized pen-paper duos, offer accurate digital captures but are constrained by the need for expensive, proprietary paper and hardware.

Handwriting reconstruction from offline image has introduced innovative metrics like DTWseg and LDTW to improve glyph fidelity and trajectory alignment can be a source

of inspiration, especially for preprocessing. However, these approaches suffer from high computational costs due to their reliance on DTW, making them less practical for real-time applications.

IMU sensor-based methods have shown promise by allowing writing on various surfaces and leveraging deep learning techniques for trajectory reconstruction. Traditional approaches, which do not rely on deep learning, often struggle with sensor noise and error accumulation. These methods typically involve a series of preprocessing steps, such as applying low-pass filters to remove high-frequency noise and using coordinate transformation matrices to adjust for sensor orientation. However, the cumulative effect of noise and small errors during these preprocessing steps can lead to significant inaccuracies in the reconstructed trajectories over time.

In contrast, deep learning methods offer a more sophisticated approach to handling IMU data. These methods can learn and adapt to noise patterns in the data, potentially providing more accurate trajectory reconstructions. Despite their advantages, deep learning methods require careful alignment between the ground truth and sensor data during the training process. To address this, techniques such as linear interpolation are used to match the sizes of sensor signals and ground truth data, although this disrupt the dynamics of the original trajectory. Moreover, normalization methods like min-max scaling are applied to the output trajectories to ensure consistent evaluation metrics, facilitating more accurate comparisons between predicted and ground truth trajectory.

From this state of the art, no single neural network architecture stands out as superior to the others for handwriting trajectory reconstruction. Various architectures, including CNN, LSTM, and hybrid models, have been explored, each demonstrating strengths and limitations in different aspects of the task. As far as the objectives of the thesis are concerned, a convolutional network would seem to be an appropriate choice, given the small amount of data involved. In addition, it seems appropriate to use a matching-based metric such as DTW or Fréchet to align sensor signals and ground truth. In the next chapter, we will delve into the details of domain adaptation for handwriting data from IMU sensors.

Chapter 6

Domain adaptation for handwriting data from IMU sensors.

Handwriting analysis using IMU data holds significant potential for applications in education. However, a major challenge in leveraging IMU data for handwriting reconstruction lies in the domain discrepancy problem. IMU data can vary significantly across different devices, individuals, and environments due to variations in sensor placement, orientation, and personal writing styles. These variances can lead to significant performance degradation of machine learning models when applied to data from a different domain than they were trained on.

Domain adaptation (DA) techniques offer a solution to this problem by enabling models to generalize across different domains. DA methods aim to transfer knowledge learned from a source domain (e.g., IMU data from a specific device or user) to a target domain (e.g., IMU data from a different device or user) with a limited amount of labeled data from the target domain. This capability is crucial in handwriting reconstruction tasks where collecting and labeling large amounts of data for every new device or user can be impractical. In particular, obtaining handwriting data from specific populations, such as children, poses significant challenges. Children's handwriting evolves rapidly as they develop their motor skills, and capturing this dynamic process requires frequent data collection over extended periods. Additionally, ethical and privacy considerations make it difficult to gather large datasets from children, as parental consent and strict adherence to data protection regulations are required.

This introduction explores the landscape of domain adaptation techniques applied to handwriting data from IMU sensors. We will discuss the unique challenges posed by IMU based handwriting data, review existing domain adaptation methodologies, and highlight recent advancements in this field. The goal is to provide a comprehensive understanding of how domain adaptation can enhance the performance and robustness of handwriting reconstruction systems in the face of domain variability, ultimately paving the way for

more accurate and widely applicable handwriting analysis solutions using wearable IMU sensors.

6.1 Presentation

Depending on the data available, the domain adaptation can be:

— **Supervised**: You have labeled data from the target domain.

— **Semi-Supervised**: You have both labeled as well as unlabeled data from the target domain.

— **Unsupervised**: You have unlabeled data from the target domain.

The three prominent way to perform domain adaptation are: Divergence-based Domain Adaptation, Adversarial-based Domain Adaptation and Reconstruction-based Domain Adaptation.

6.1.1 Divergence-based Domain Adaptation

This approach focuses on minimizing the divergence or discrepancy between the distributions of the source and target domains. Techniques such as Maximum Mean Discrepancy (MMD), Kullback-Leibler (KL) divergence, and Wasserstein distance are commonly employed to quantify and reduce this statistical distance. The core intuition behind these methods is that the error in the target domain is bounded by the error in the source domain plus the distance between the source and target distributions. This means that, in order to effectively apply a model trained on the source domain to the target domain, we need to find a representation space where we remain good at classifying the source data while also ensuring that the source and target distributions are as close as possible.

6.1.1.1 Wasserstein Distance Guided Representation Learning (WDGRL)

This technique [Shen et al., 2018] aims to address the challenge of domain adaptation by learning feature representations that are invariant across different but related domains, which typically have varied data distributions. WDGRL employs a domain critic network to estimate the empirical Wasserstein distance between the source and target domains and then optimizes the feature extractor network to minimize this distance in an adversarial fashion (Fig. 6.1).

Empirical studies show this approach it outperforms other state-of-the-art methods in domain adaptation tasks across common sentiment and image classification datasets. Additionally, WDGRL integrates well with existing domain adaptation frameworks by replacing their representation learning components, and can be further customized for various domain-specific applications.

From a supervisory point of view, this method is used in the same way as the previous one.

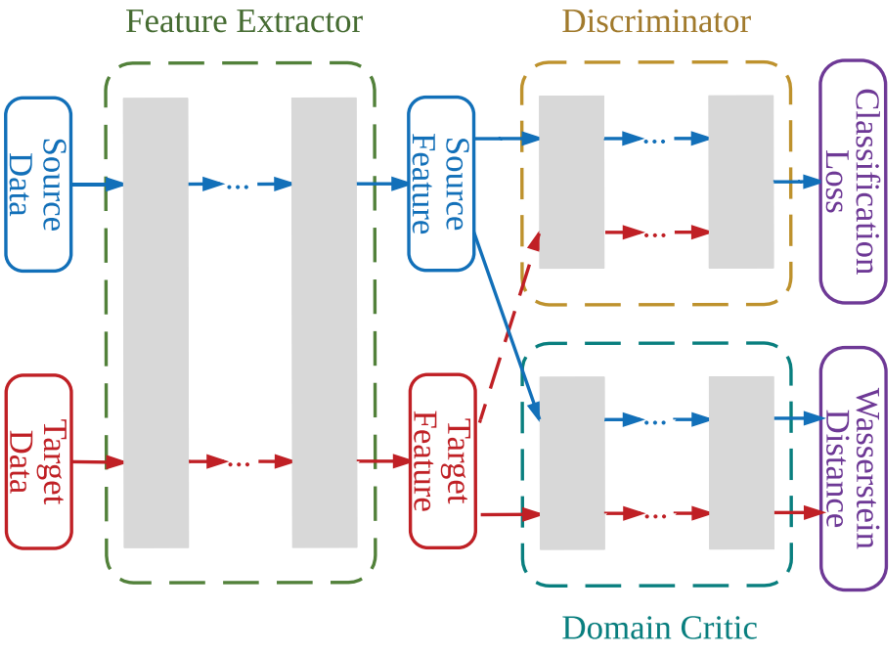


Figure 6.1 – [Shen et al., 2018] method, with Wasserstein Distance used as objective for the adaptation.

6.1.2 Adversarial-based Domain Adaptation

Inspired by Generative Adversarial Networks (GAN), this method employs a game-theoretic approach. It involves a discriminator and a feature extractor where the discriminator tries to distinguish between the source and target domain features, and the feature extractor learns to generate features that are domain-indistinguishable. This adversarial training helps in creating a feature space where the distinction between the domains is minimized, thereby aiding in better generalization of the model to the target domain.

6.1.2.1 Domain-Adversarial Training of Neural Networks (DANN)

Domain-Adversarial Training of Neural Networks [Ganin et al., 2016], are specialized neural network architectures designed to address the challenge of domain shift, where a model trained on one domain (source) is expected to perform well on a different but related domain (target). These networks operate by learning features that are domain-invariant, meaning they are useful and generalizable across both domains. This is typically achieved through a shared feature extractor on which additional components are built: a domain classifier and a task-specific classifier. The domain classifier tries to determine the domain of the input data, whereas the task-specific classifier focuses on predicting the actual labels.

Training involves a twist (Fig. 6.3): the domain classifier's gradients are reversed during backpropagation, which encourages the feature extractor to generate features that are indistinguishable between domains, thus fooling the domain classifier. This technique, known as adversarial training, helps the network to minimize the representation gap between the source and target domains, leading to better performance on the target domain without requiring extensive labeled data from it.

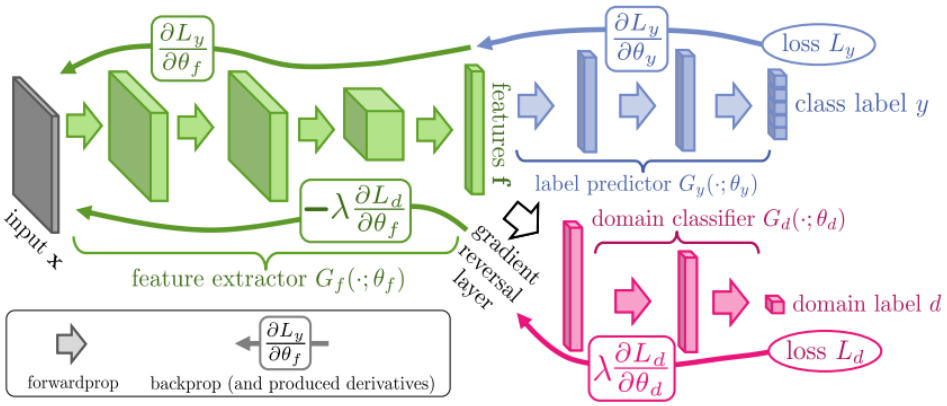


Figure 6.2 – DANN from [Ganin et al., 2016], consists of a feature extractor (green) and a label predictor (blue), forming a standard feed-forward structure. To achieve unsupervised domain adaptation, a domain classifier (red) is incorporated. This classifier is linked to the feature extractor via a gradient reversal layer, which multiplies the gradient by a negative constant during backpropagation. This training approach minimizes both the label prediction loss (for source domain examples) and the domain classification loss (for all samples).

6.1.3 Reconstruction-based Domain Adaptation

This technique utilizes the concept of reconstructing inputs in either the source or target domain from a common feature representation. By encouraging the model to maintain reconstruction ability across domains, the underlying feature representation becomes more robust to domain shifts. Autoencoders are a popular choice for implementing this strategy, where the encoder learns a domain-agnostic representation and the decoder reconstructs the domain-specific data.

6.1.3.1 Deep Reconstruction Classification Networks (DRCN)

[Ghifary et al., 2016] introduces Deep Reconstruction Classification Networks (DRCN), a convolutional network designed to simultaneously tackle two tasks: supervised prediction of source labels and unsupervised reconstruction of target data. The objective is to ensure that the label prediction function effectively classifies images in the target domain, with the reconstruction task serving as a supportive auxiliary function for enhancing label prediction adaptability. Unlike conventional pretraining-fine tuning methods, the DRCN employs a unique learning strategy that alternates between unsupervised and supervised training phases.

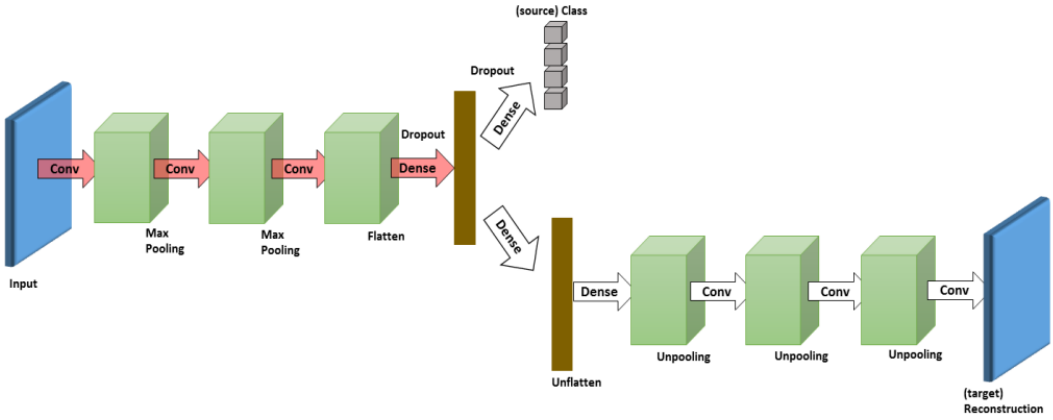


Figure 6.3 – DRCN [Ghifary et al., 2016], is structured around two main branches: the label prediction pipeline and the data reconstruction pipeline.

6.2 Conclusion

This chapter explores domain adaptation techniques applied to handwriting analysis using IMU sensors, addressing the challenge of domain discrepancy caused by variations in devices and individuals. The chapter introduces specific models like Domain-Adversarial Training of Neural Networks (DANN) and Wasserstein Distance Guided Representation Learning (WDGRL), which aim to create domain-invariant features to enhance the performance of handwriting reconstruction across different domains, with a focus on applications in education and data from children.

DANN is particularly well-suited for handwriting analysis using IMU sensors due to its focus on learning domain-invariant features through adversarial training. Its flexibility to operate in both supervised and unsupervised settings allows it to adapt to various scenarios. Moreover, DANN has been widely validated in various domain adaptation tasks, demonstrating superior performance in scenarios where there is a significant domain shift.

Part II

Contributions

Chapter 7

Data acquisition protocol

At the start of the project, we had no data. In fact, the FAU-EINNS proposed by[Wehbi et al., 2022] does exist; however, it is limited to words and includes only six writers, making it unsuitable for our needs. Additionally, the old version of the pen operates at 100 Hz, which may lack precision and is particularly far from the frequency of the tablet. So we set up a protocol and data collection with the project partners. In this section, we describe the choices we made. This will be followed by a description of the datasets that have been collected during the project. Some of these data has been made public to help future work in the field.

7.1 Data acquisition

Throughout this thesis, several versions of the digital pen were employed, reflecting advancements in pen technology and impacting the data acquisition process. At the beginning of the research project, Digipen Version 5.0 was utilized, operating at a 100 Hz sampling rate. During the first year, the pen was upgraded to Digipen Version 6.0, which offered sampling rates of 100 Hz, 200 Hz, and 400 Hz. Later, Version 6.3 was introduced, which incorporated sensors from different manufacturers. In the final year, Version 7.5 was developed with the front accelerometer removed, featuring dual acquisition capabilities for both paper and ground truth.

These technological advancements highlight the challenge of obtaining consistent data, as the pen's design was still evolving at the start of the research project. The changes in sensor technology and acquisition methods by Stabilo significantly impacted the development of the acquisition protocol and the analysis of the data throughout the research project.

7.1.1 Data acquisition challenges

To create datasets the acquisition process is quite challenging due to several difficulties which we will briefly remind you of. First, the stylus and the tablet have **different sampling rates**, so the selection of both of them is an important step. Another challenge concerns the recording of the **pen-up movements**. When the pen gets too far from the tablet (over 7mm high), the tablet stops recording a trace. This results to parts of the data coming from the stylus that do **not match any ground truth from the tablet**. In addition, a **drift** in the accelerometer measurements is possible, and the kinematic signals are transmitted from the pen to the tablet in sets of six points through Bluetooth. Variable transmission time delay results in asynchronous timestamps for kinematic and tablet data.

7.1.2 Acquisition tools description

To create a datasets of the handwriting trajectory reconstruction task, we use three equipments and tools:

— the Digipen is equipped with Wacom insert to be used with EMR acquisition technology;

— a Samsumg Galaxy S7 FE tablet, working with EMR technology;

— an Android application developped by Stabilo, that guides the collection in which the data from the tablet are associated to the ones of the pen.

For ground-truth acquisition, the Digipen pen is equipped with a Wacom insert, providing a ground-truth online trajectory trace of the handwriting that corresponds to the pen's IMU signals (Fig. 2.1).

As a reminder, the Digipen embeds the following IMU sensors (Fig.1.1): a gyroscope, a front and a rear accelerometer, a magnetometer, and a force sensor. Each of these sensors provides temporal reading values that describe the relative pen movement in 3-axes channels (x, y and z), except the force sensor which has only one channel. The readings of the sensors are buffered before being sent to the tablet via Bluetooth connection. This produces multi-variate time series made of 13 modalities (10 for Version 7.5) for every dataset sample.

The tablet choice is important, as the model, the screen size, and the sampling rate must be considered to get consistent online handwriting signals. Remember, tablet data serves as the ground truth for training our neural networks. Indeed, each tablet has

its own sampling rate so, only one model of tablet (Samsumg Galaxy S7 FE) has been used for the whole data collection campaign to obtain homogeneous ground truth. The Samsumg Galaxy S7 FE has a resolution of 2560 x 1600. The EMR technology captures the pencil signal up to 7mm high. The tablet dynamically adjusts its sampling rate based on the stylus's speed, increasing it for fast movements to capture detailed input accurately (up to 370Hz), and decreasing it for slow movements or pauses to conserve power and processing resources (down to 60Hz).

Stabilo has developed a dedicated mobile application to record (i) the handwriting trajectories using the Digipen integrating Wacom tip, (ii) the corresponding sensor signals from the Digipen. The application allows to calibrate the Digipen sensors' signals and setup the sampling rate of the Digipen sensors. The app provides instructions, allowing the user to simply follow the prompts and complete the required writing tasks. Data acquisition is pseudonymized with an individual ID assigned to each person. This ID is not stored in a database or linked to personal identities, ensuring privacy and security.

Before collecting a large amount of data, we chose an appropriate sampling frequency among 100Hz, 200Hz, and 400Hz. We opted for 400Hz as it closely aligns with the tablet's maximum sampling rate of 370Hz. This choice minimizes discrepancies in sequence size, which is crucial for reducing the impact of preprocessing alignment.

7.1.3 Data acquisition protocol

The recording process (Fig. 2.1) begins by selecting a set of predefined scripts to be written on the tablet surface using the Digipen. These two data sets are made up of the following two recording types:

— BASIC, consists of 34 samples to be written one by one during a single recording session. It is composed of five types: 15 characters, 10 words, 5 equations, 2 shapes and 2 word groups.

— EXTENDED, consists of 57 samples to be written one by one during a single recording session. It is composed of five types: 30 characters, 10 words, 5 equations, 4 shapes and 8 word groups.

These two distributions are chosen to facilitate the study of handwriting reconstruction. The datasets contains more characters than other types, focusing on essential strokes and forms. This approach helps models first learn the fundamental elements of handwriting before progressing to more complex forms. Words and groups of words have also

been collected, the length of which, linked to potential drifting and pen-up movements, increases the complexity. Equations and shapes, though less numerous, are included specifically to extend the model's capabilities in reconstructing non-textual handwriting elements by the presence of many pen-up phases. This gradual and intentional increase in complexity ensures that the model develops robust reconstruction skills, effectively transitioning from simple to more complex handwriting inputs, thus making it highly adaptable for both research and practical implementations in handwriting reconstruction.

While recording, a user holds the pen's on/off switch up, which is a natural way to take the Digipen due to grips designed on the pen to naturally position the fingers properly. Since the pen is held consistently, the sensors are oriented in the same direction. Consequently, for similar trajectories, the signals resemble each other closely.

Regarding data collection, we initially started the project with version 5.0 data from a select group of users. The first data collection campaign commenced with the release of the 6.0 pencil at the end of the first year, conducted by the four project partners for adults. Data from children was collected exclusively by Learn&Go in schools, resulting in a smaller dataset due to the greater complexity of collection. Six months later, a second campaign began, featuring an extended data collection strategy to address the issues we encountered.

7.2 Data cleaning

As part of the research project, hardware evolution over time has introduced challenges, particularly regarding consistency between Wacom acquisition and Digipen force measurements. Small samples often face incorrect pressure conditions, pencil lift are sometimes undetected and sometimes over-detected, resulting in a mismatch between sensor signals and tablet ground truth. To remedy this, we suggest the following method (Algorithme 1): We assess whether the sensor data and ground truth are aligned. In cases of minor discrepancies, we adjust the samples accordingly. However, if the mismatch is too significant, we exclude the sample from the training set.

Algorithm 1 Pseudocode for cleaning incorrect pressure recordings

```
function CLEANING(tablet_samples, sensors_samples)

    filtered_tablet_samples ← []    ▷ Initialize a list to store filtered tablet samples.
    filtered_sensors_samples ← [] ▷ Initialize a list to store filtered sensor samples.

    for (tab_sample, sens_sample) in (tablet_samples, sensors_sample) do
                                        ▷ Loop over pairs of tablet and sensor samples.

        tablet_partitions ← CUT_INTO_STROKE(tab_sample['pres'])
                            ▷ Find touching and pen-up partitions for the tablet sample.
        sensor_partitions ← CUT_INTO_STROKE(sens_sample['force'])
                          ▷ Find touching and pen-up partitions for the sensors sample.

        if len(tablet_partitions) ≠ len(sensor_partitions) then
              ▷ Check if the number of partitions differs between 'tablet' and 'sensor'.

            if abs(len(tablet_partitions) - len(sensor_partitions)) < 5 then
                    ▷ Check if the difference in partition count is small (less than 5).

                tab_sample, sens_sample ← MERGE(tab_sample, sens_sample, 20)
                     ▷ Stroke merging for those with a difference of less than 20 points

                Append tab_sample to filtered_tablet_samples
                Append sens_sample to filtered_sensors_samples
            end if
        else
            Append tab_sample to filtered_tablet_samples
            Append sens_sample to filtered_sensors_samples
        end if
    end for
    return filtered_tablet_samples, filtered_sensors_samples
end function
```

This approach aims to mitigate inconsistencies and enhance the accuracy and reliability of our data by ensuring that the pressure measurements from both Wacom and Digipen are aligned and consistent, while effectively handling small discrepancies through merging. Even if this cleaning is necessary to have coherent data between the sensor and the pen, it only affects about 3% of adults' recordings.

7.3 Datasets description

From this adult data collection, we have decided to make two datasets public to serve as benchmarks for future studies: the IRISA-KIHT-S dataset and the KIHT-Public dataset [1]. These datasets are composed of 30 recordings for the IRISA-KIHT-S dataset and 149 recordings for the KIHT-Public datasets. For each recording session, 4 different files are provided:

— The sensor signals file has 14 columns: milliseconds, accelerometer front (x, y, z), accelerometer rear (x, y, z), gyroscope (x, y, z), magnetometer (x, y, z), and force signals;

— Tablet signal files contain milliseconds, position coordinates (x, y, z), and force signals;

— The transcription (labels) file contains labels (the text to be written or the task to be carried out for a form) and the start and stop time-stamps for every sample;

— Additional files concerning the sensor calibration and recording meta data are provided.

We also worked on two private datasets during this thesis: IRISA-KIHT and KIHT-Private dataset. Each dataset includes BASIC and EXTENDED recordings. Note that the test recordings are identical across four datasets to ensure fair comparisons. The KIHT-Private dataset is an extension of the KIHT-Public dataset with an additional 300 recordings. The IRISA-KIHT-S dataset is a subset of the IRISA-KIHT dataset. Each datset is composed of five groups: characters, words, sentences, equations and form (Tab. 7.2). Note that the writers in the test set are not included in the train set.

Note that the line “Inclined data” in the table 7.1 refers to data acquired on inclined planes, which will be used in the method presented in section 10.2. Note that the IRISA-KIHT dataset 57 recording of German words are included, hence the higher proportion of words.

Note that child databases collected by Learn&Go project partner, will be detailed in the chapter 11 when they are used.

1. https://www-shadoc.irisa.fr/irisa-kiht-s-and-kiht-public-datasets/

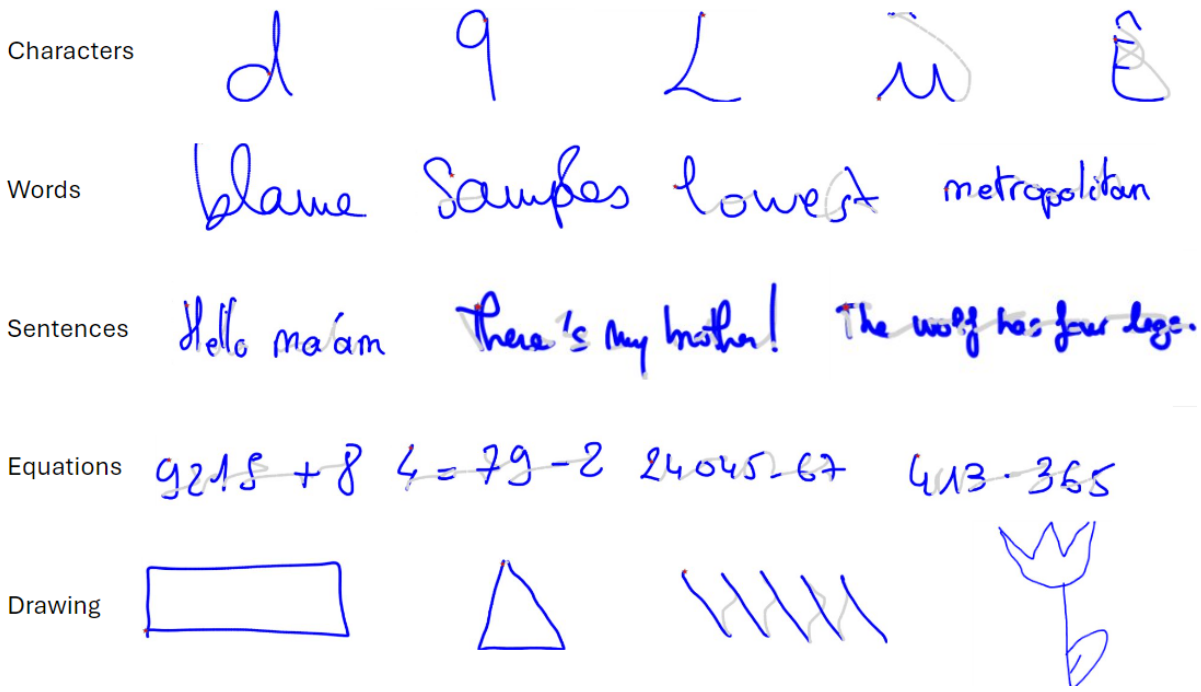


Figure 7.1 – Some examples of KIHT-Public data. Pen-up movements are in gray and pen-down (touching) strokes are in blue.

Table 7.1 – Adult datasets and their corresponding sets.

Sets	Datasets	Inclined data	# Writers	# Recordings	# Samples
Training	IRISA-KIHT		25	113	3770
	KIHT-Public		36	90	2761
		x	7	30	1368
	KIHT-Private		66	371	11811
		x	12	49	2234
	IRISA-KIHT-S		20	30	680
Test	Common test set		9	9	266
	IRISA-KIHT-S		10	10	340

7.4 Conclusion

In this chapter, we have detailed data choice and acquisition necessary for the handwriting trajectory reconstruction task. Beginning this work without data, we establish a protocol for data collection. The data acquisition protocol was designed to include a diverse range of handwriting samples, from basic characters to complex equations and form, to build a comprehensive dataset that supports various aspects of handwriting reconstruction.

Table 7.2 – Detailed distribution of each entry category in datasets.

Sets	Datasets	Inclined data	Characters	Words	Sentences	Equations	Drawing
Training	IRISA-KIHT		915	2306	102	305	122
	KIHT-Public		1217	812	163	406	163
		x	603	402	81	201	81
	KIHT-Private		5209	3474	697	1734	697
		x	985	657	132	328	132
	IRISA-KIHT-S		300	200	40	100	40
Test	Common test set		135	81	18	45	18
	IRISA-KIHT-S		150	100	20	50	20

The following chapters will use these databases to support the experiments. The first data collections consist primarily of adult data, which is why our focus will be on reconstructing handwriting trajectories for adults. This strategy allows for a clearer segmentation of challenges, preventing the accumulation of difficulties and offering a more focused understanding of the handwriting reconstruction problem.

Chapter 8

A first preprocessing chain

Preprocessing is a crucial step in the preparation of sensor and tablet data, especially when it comes to deep learning methods that require sequence pairs of the same size (input / ground truth). A primary challenges is the varying sampling frequencies between the sensor and tablet data. This discrepancy can lead to misaligned data points, making it difficult to accurately analyze the information. Additionally, the tablet signal is prone to being lost when the pen is raised too high above the tablet, resulting in gaps in the data that need to be addressed. To ensure the effectiveness of deep learning models, it is essential to preprocess the data such that the ground truth and the sensor data have identical sizes. This involves a preprocessing pipeline detailed in figure 8.1 and presented in the next sections. We will present the preprocessing steps involved in the training phase. It is important to note that the preprocessing for the test phase is identical, except for the Dynamic Time Warping (DTW) alignment and the exclusion of the ground truth component.

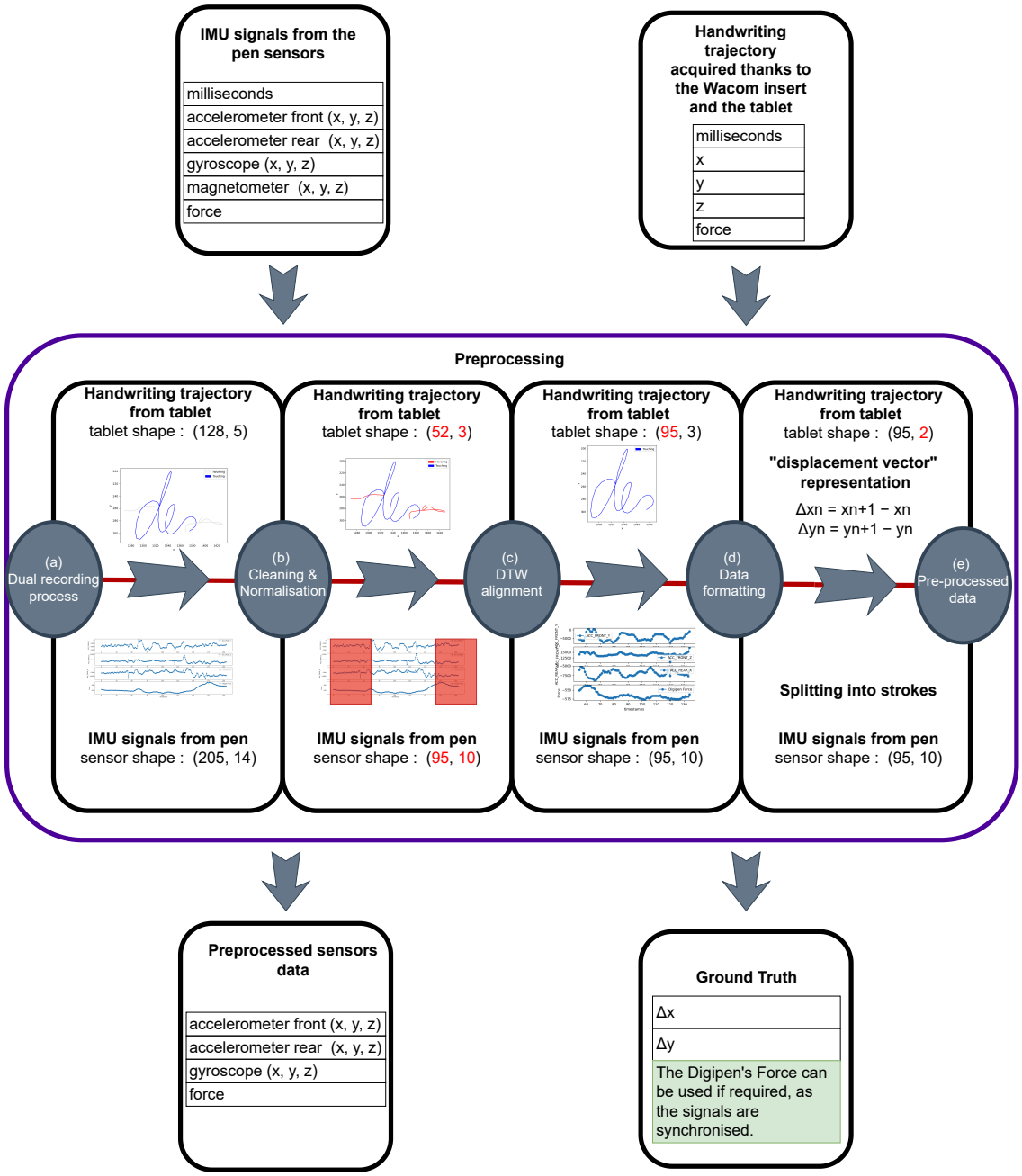


Figure 8.1 – Our proposed preprocessing pipeline. (a) Thanks to dual acquisition, we recover Digipen signals (14 channels: 2 x 3 (x, y, z) for accelerometer, 3 for gyroscope, 3 for magnetometer and 1 for the force) and the ground truth (3 channels: x, y and pressure), (b) we remove start and end pen-up movements, which are not related to handwriting, (c) we align ground truth and sensor signals using DTW, (d) data formatting, (e) we obtain the preprocessed data used for training.

8.1 Cleaning and normalization

8.1.1 Dimension reduction

The first preprocessing step is to remove the 3 magnetometer channels (x, y, z) from the sensor channels. As the data is collected on a tablet, the electromagnetic field detected is that of the tablet and not that of the earth. Therefore, in order not to bias the network by mapping it to the electromagnetic field specific to a tablet, we have removed the magnetometer data.

We have also decided to remove the z-coordinate from the ground truths, as our primary focus is on the writing within the (x, y) plane. The timestamps of the Digipen and the tablet are "kept" during preprocessing.

8.1.2 Signal splitting and normalization

First, signals are divided into spans that correspond to the written samples. The samples are segmented based on the timestamps provided in the label.csv file, which specifies the start and end times for each sample. These timestamps were recorded at the time of data collection.

Irrelevant parts of the input signals are removed. Those parts are the start and end pen-up movements that does not refer to pen-up and pen-down actions to write the given script (Fig. 8.1(b)). Cutting is performed using Digipen's force sensor by detecting the presence or absence of a zero-pressure reading. In order to maintain the interoperability of the system between the different versions of Digipen, accelerometers, gyroscope, magnetometer and force signals are normalized by their maximum values (as reported by the manufacturer).

8.2 DTW alignment

Due to the different sampling rates between the stylus and the tablet, an alignment process is crucial to get a mapping between each input points from the stylus and each output points from the tablet. In addition, the sensor data are acquired in packets of 6 data points and the recorded timestamps are not equally spaced due to the Bluetooth transmission time delay (Fig. 8.2).

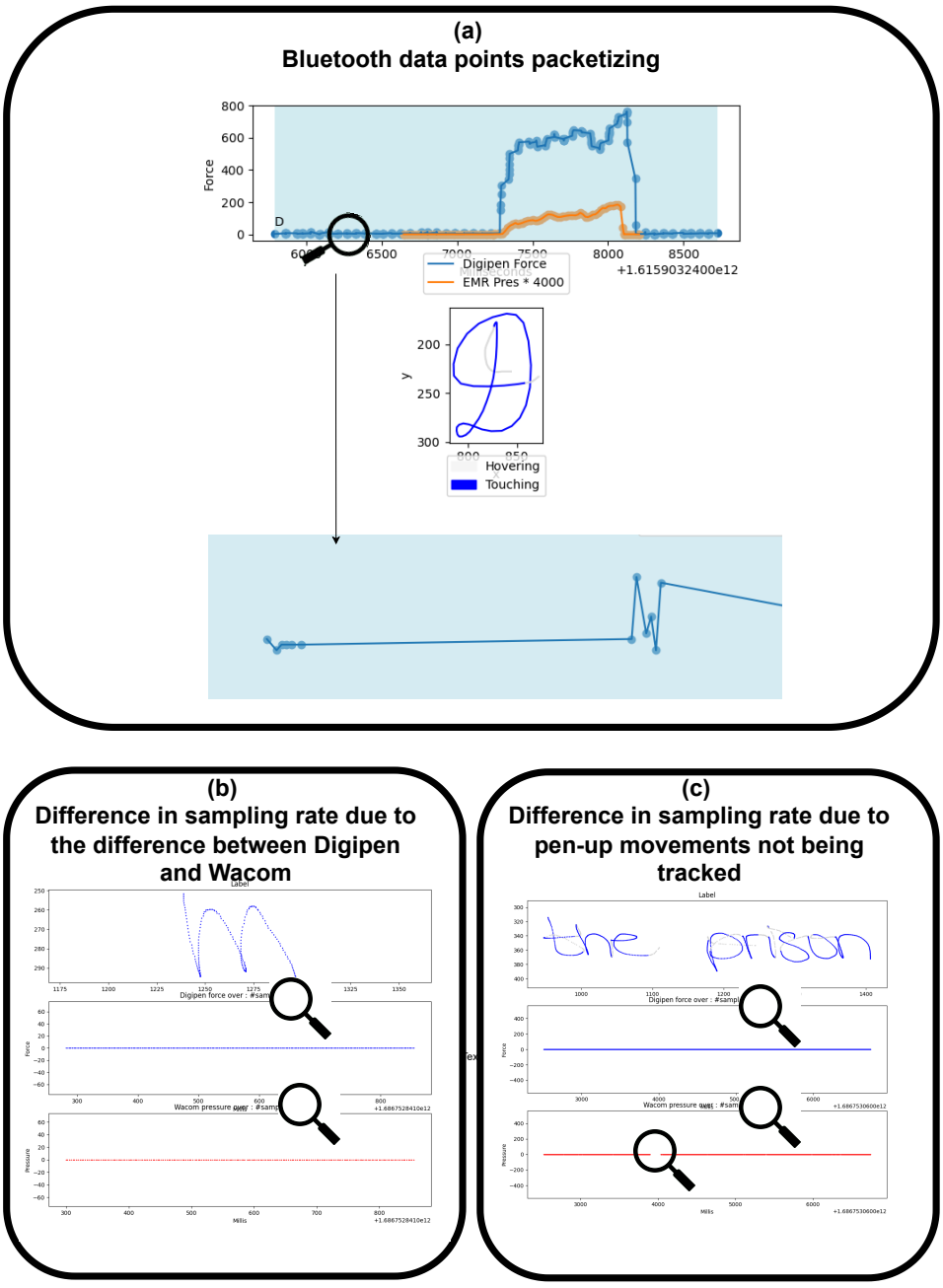


Figure 8.2 – Visualization of the challenges linked to the data, in particular the grouping for transfer by Bluetooth (a), and the difference in the number of points between the Digipen and the ground truth due to both a difference in acquisition frequency (b) and signals lost during pen-up movements (c). In figures (b) and (c) the first line corresponds to the ground truth (x, y), the second line corresponds to the distribution of Digipen 'pressure' points over time, the last line corresponds to the distribution of Wacom 'pressure' points over time.

Due to this mismatch, the recorded input and output signals have different lengths and are not synchronized. A naive approach would be to linearly interpolate the ground truth to match the number of points in the sensor sequence. [Wehbi et al., 2022] uses this method, and we will come back to this point in more detail in the following chapters. However, as illustrated in Figure 8.3, linear interpolation fails to preserve the dynamics of the writing. While linear interpolation evenly distributes points, it disrupts the natural linkage between the ground truth and the sensor signals. Consequently, the interpolated points no longer align accurately with the sensor signals, undermining the fidelity of the synchronization.

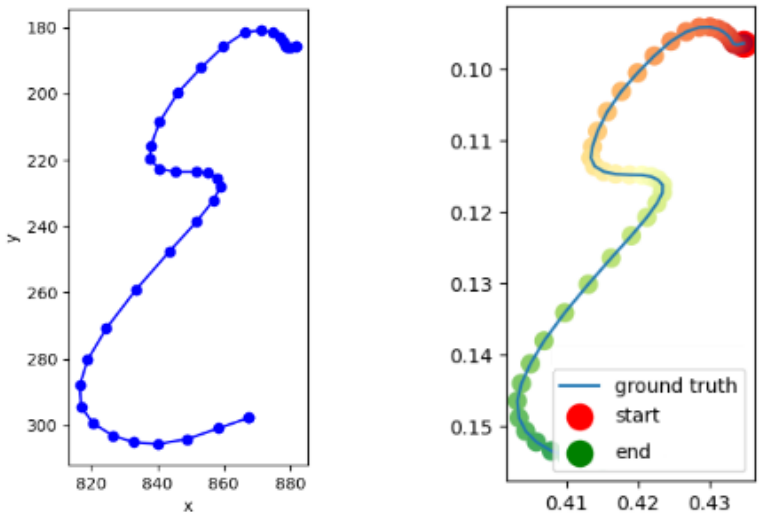


Figure 8.3 – Visualization of a raw and linearly interpolated character that alters the dynamics.

To respect the dynamics of writing as much as possible, we propose an alignment approach based on the Dynamic Time Warping (DTW) algorithm (Figure 8.1(C)). In practice, we observe that aligning the timestamps using the DTW algorithm is more effective than relying on the DTW alignment on force and pressure data. Figure 8.4 shows a comparison between the alignment based on linear interpolation and our proposed approach based on DTW. Alignments are presented on the force (pen sensor) signal against the pressure signal (from the tablet). Unlike linear interpolation, DTW matche points to the closest position in time in the trajectory. This results seems more fairthful and reliable representation of the data. The transmission time delay is between 10 and 40 milliseconds in the Digipen version 6.0. The average transmission time delay is subtracted

to the tablet data in order to approximately synchronize it with sensors data. Then, we use the DTW algorithm to find an alignment path between the timestamps of the stylus and the tablet. Since the sampling rate of the sensor data is higher than the one of the tablet data, we have made the choice to up-sample the tablet data to match the sensors data length. Indeed, We wanted to maintain the pen signals as they are to avoid affecting the input dynamics of the model.

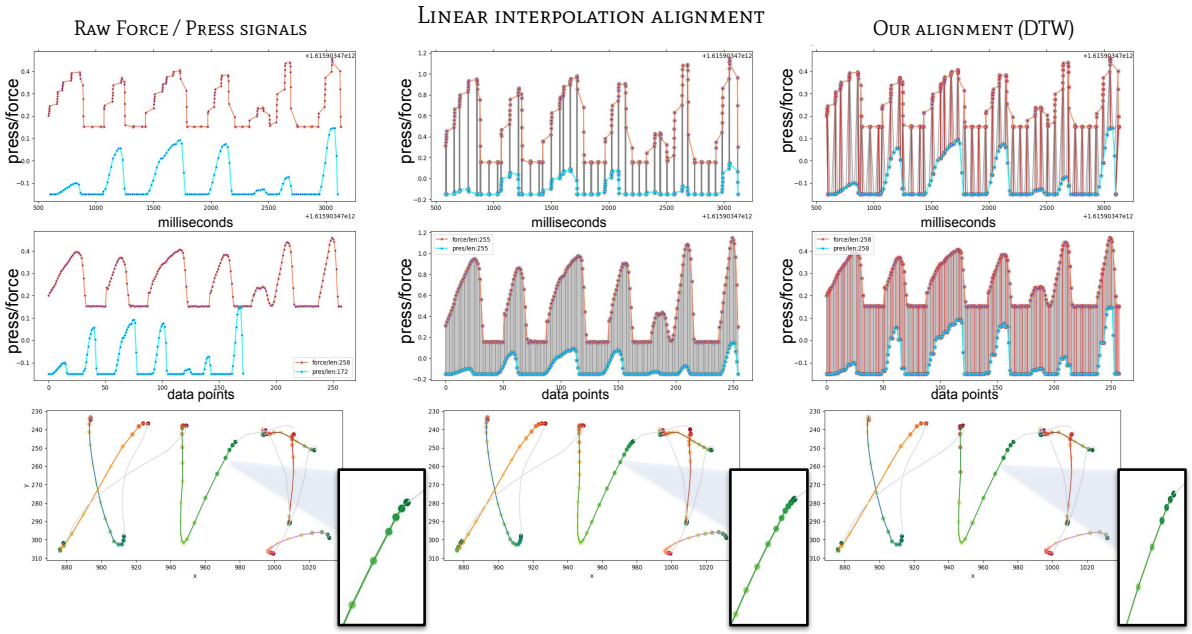


Figure 8.4 – Our proposed DTW-based alignment compared to the linear interpolation-based alignment. The left side image represents the raw force / pressure (sensor/online) signals, the middle image shows the linear interpolation alignment result and the right side image shows the DTW-based alignment results. The red lines connecting the two signals represent the duplicate points in the pressure signal and the grey ones represent the one-to-one alignment. The first line shows the 'pressure' of Digipen and Wacom over time. The second line shows Digipen and Wacom 'pressure' by number of points. The last line shows the handwriting trajectory used as ground truth.

8.3 Data formatting

8.3.1 Splitting into strokes

Splitting into strokes is an aspect of the training phase in our proposed pipeline. This process focuses on training the neural network using only touching stroke of handwriting. The primary motivation for that is to avoid losing trace of pen-up movements when the pen is lifted high (over 7mm [1]). Strokes are identified by analyzing the force/pressure signals from the sensors and the tablet. Specifically, we consider a stroke to be valid if it meets the following conditions: the force value of the Digipen must be greater than a predefined threshold of 0.01, and the pressure value of the tablet must be greater than 0. These conditions must be met simultaneously for a stroke to be preserved. This method ensures that the neural network is trained only on parts dedicated to the handwriting trace, enhancing the accuracy and effectiveness of the model in recognizing handwriting patterns.

8.3.2 Ground truth representation

Initially, one might consider using the absolute positions (x, y) of each point in the trace. However, learning to predict absolute positions from sensor data that represents relative changes (i.e., displacements due to acceleration, speed, and orientation of the Digipen) is mathematically impossible. The starting position is unpredictable and is therefore not worth predicting.

Consequently, we opt for the "displacement vector" representation. This approach entails using displacement vectors between successive points, meaning each point is defined relative to its predecessor. Given that the input sensor signals reflect relative changes, this representation logically aligns with the data characteristic. In the "displacement vector" representation, vectors (Δx, Δy) are calculated from the successive (x, y) coordinates of the handwriting trajectory. This method leverages the relative changes in position, making it a suitable choice for capturing the dynamics of the handwriting process.

1. http://tennojim.xyz/article/wacom_intuos_pro_l_guide

8.4 Conclusion

In conclusion, the preprocessing of sensor and tablet data is essential for ensuring the accurate alignment and effectiveness of deep learning models designed to analyze handwriting patterns. By addressing challenges such as varying sampling frequencies, transmission delays, and signal gaps caused by pen-up movements, we implement a robust pipeline that ensures synchronized input and ground truth data. Through steps like the removal of unnecessary sensor channels, signal normalization, DTW-based alignment, and stroke segmentation, the data is transformed to maintain the natural dynamics of handwriting.

As a result from this preprocessing, we get touching strokes as input of the neural network, and for each stroke we have two time series of equal size, the first coming from the Digipen sensors, the other being the Wacom ground truth. We reduced the number of channels for DigiPen signals from 14 to 10 and decreased the number of ground truth channels sides from 4 to 2, with a significant shift from absolute to relative position measurements. Comparative results between our proposed DTW-based alignment and the linear interpolation used in the state of the art [Wehbi et al., 2022] will be discussed in the next chapter.

Chapter 9

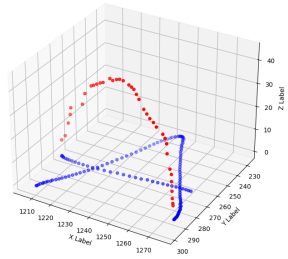
A Touching Expert Model (TEM)

Reconstructing the handwriting trajectory from IMU signals is a complex task for several reasons. Firstly, sophisticated algorithms are required to deal with the multi-dimensional nature of IMU data comprising acceleration and angular velocities and to synthesize it into a two-dimensional handwriting trajectory. Secondly, the challenge is further heightened by the need to mitigate sensor noise and bias, which can cumulatively lead to significant trajectory drifts. We need a model that can analyze signals in fine detail at the local level while also considering broader, global patterns to correct these biases. The approach needs to capture the intrinsic variability of human handwriting, which fluctuates not only between individuals but also within a single individual between different writings sessions, as well as the sensitivity to different frictions according to the surface and the size of the writing, which leads to different movements.

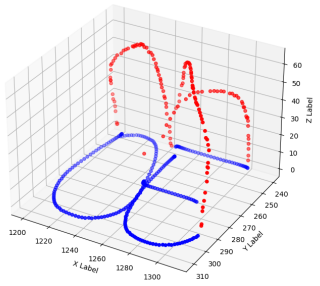
In fact, handwriting can be divided into two distinct phases: the act of writing, where the stylus touches the surface, and the pen-up movements, which occurs when transitioning to the next segment of writing. The first phase, involving direct contact with the surface, has a 2-dimensional ground truth based on the observable path of the stylus on the surface. The second phase, the pen-up movements, is to move the stylus through a 3-dimensional space to the next starting point for writing. This phase introduces a different set of dynamics and that there is more variance in the gesture, as the trajectory can be more random, with no ground truth if the stylus is raised too high during pen-up movements (Fig. 9.1).

We therefore decided to start by simplifying our problem, by focusing on the stylus touching parts, for which we have sensor information and ground truth for all sequences.

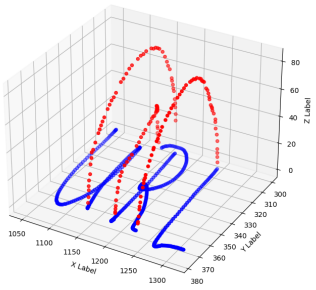
So, we design a neural network architecture to predict the displacement vectors of the handwriting trajectory from the tablet, given a sensors signal of the Digipen (Fig. 8.1(d)). The training and test steps of our proposed model are described in the following sub-sections. In a second step, we will present a more advanced version of our model,



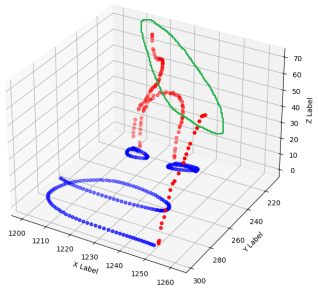
(a) "A" for ground truth with pen-up movements



(b) "OE" for ground truth with pen-up movements



(c) "URL" for ground truth with pen-up movements



(d) an "ë" for ground truth with pen-up movements, but part of it (in green) is not tracked

Figure 9.1 – Examples of data used as ground truth. They are composed of writing parts (in blue) and pen-up movements (in red) that are not always tracked (in green).

named TEM-C. The chapter will conclude with a comparison of experimental results, both with state-of-the-art methods and between our two variants, TEM and TEM-C.

9.1 A Touching Expert Model (TEM) based on Temporal Convolutional Network

In order to take into account the past and future states when mapping an input sequence toward an output signal, we propose to use a non-causal Temporal Convolutional neural Network (TCN) architecture inspired by [Bai et al., 2018]. We name this model TEM for touching expert model as it is trained on touching strokes only.

9.1.1 Architecture Choice

The choice of a TCN architecture is supported by the great success of the TCN for sequence-to-sequence tasks with few training samples, e.g. for weather prediction [Yan et al., 2020], traffic prediction [Dai et al., 2020] and sound event localization and detection [Guirguis et al., 2021].

A CNN architecture can capture the spatial features that refer to the arrangement of data points of a sequence, and the relationship between them within the sequence. However, the advantage of TCN over CNN is their ability to capture more distant context with less depth, thanks to dilated convolutions. This allows TCN to look further into the past while maintaining a shallower network architecture, thereby reducing the risk of encountering vanishing gradients that typically arise with deeper networks. TCN is designed to extract both local and global features, allowing for reconstruction that aims to minimize drift effects.

It has been shown that TCN architecture is most suitable to extract relevant spatial and temporal features for a sequence of frames describing an action compared to an LSTM recurrent network [Nan et al., 2021]. Indeed, recurrent architectures are known to suffer from vanishing gradient [Roodschild et al., 2020] and forgotten information problems in very long sequences, such as those produced by IMU sensors. Thus, TCN based systems outperforms their LSTM counterparts in different fields of application such as anomaly detection [Gopali et al., 2021] or skeleton-based action recognition [Nan et al., 2021].

While transformers are gaining popularity in many domains, they come with notable limitations. One major issue is their computational complexity, which increases quadratically with the sequence length. This quadratic dependency makes training Transformer models computation-heavy for very long sequences. In addition, a context limited to a fix number of tokens may not be sufficient to capture all relevant information [Zaheer et al., 2021]. Additionally, Transformers are known to demand a significant amount of data in training, which lack in the KIHT project.

9.1.2 Architecture details

Our TEM model, described in Fig. 9.2, is designed to handle 10 input channels of sensor data. The core of this architecture consists of Temporal Convolutional Network (TCN) layers.

The TCN part consists of four stacked inner blocks of non-causal and dilated 1D-convolutions with kernel size of 3. The dilation rate applied to each block is increased

with the depth of the network, from 1 to 2. Each convolutional layer is followed by a batch normalization layer, which standardizes the activations from the layer.

The network terminates with two dense layers (with a linear activation for the first), with a layer of batch normalization in between. The output of the second dense layer refers to the two channels $(\Delta x, \Delta y)$ representing the spatial displacements.

In this configuration, the receptive field of 49 and there are less than 900 000 parameters, which is a relatively small number compared to popular neural networks. This is consistent with the fact that the data sets are of limited sizes.

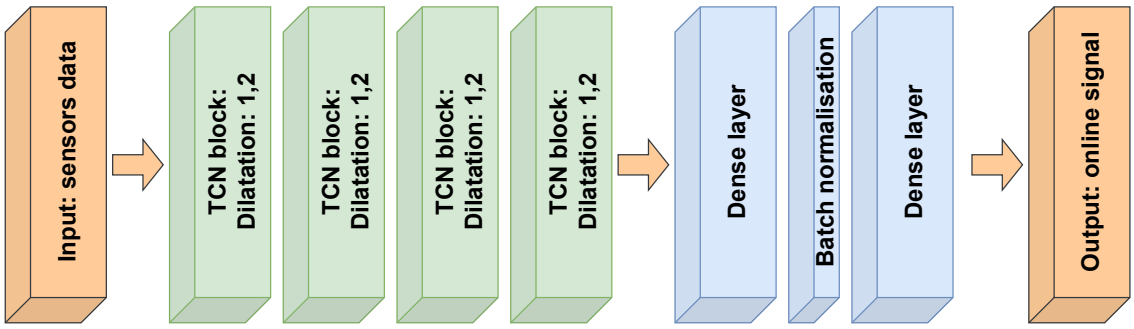


Figure 9.2 – Our TEM model for handwriting trajectory reconstruction. This model have 870452 parameters.

9.1.3 Model training and test

During the training phase, the model is trained to minimize the Mean Squared Error between the real and predicted displacement vectors of the handwriting trajectories. We uses the ADAM optimization, a batch size of 16 and a learning rate of $10e^{-3}$. As stated before, the 10 input channels of the sensor signal are the front and rear accelerators, gyroscope and force channels of the sensor signal, obtained after the cleaning and normalization process. During the test phase, a cleaned and normalized signal is given as input of the model that predict the corresponding displacement vectors of the handwriting trajectory signal.

The results of this first approach will be presented in section 9.4. In the following section we will present a first improvement of our TEM model.

9.2 TEM-C: Incorporating temporal context that reflects physics and dynamics to enhance the touching expert model

The TEM architecture is well-suited for stroke-level reconstruction, but cutting touching strokes leads to a theoretical weakness.

Indeed, in the reconstruction of a trajectory based on Inertial Measurement Units, the temporal context plays a crucial role in accurately capturing the dynamic movement of an object. IMU are sensor systems that measure specific forces and angular rates to determine the acceleration and orientation of an object. Mathematically, the integration of these measurements over time helps to reconstruct object trajectory. The importance of temporal context in trajectory reconstruction process is the result from the necessity to accurately model continuous variations in acceleration and angular velocities, which reflect object dynamic behaviors including acceleration and directional shifts. By integrating temporal information, the object motion can be more accurately reconstructed, accounting for these dynamic changes.

This weakness has been confirmed in our experiments. The prediction was less accurate (Fig. 9.3) for the initial points of a stroke due to a lack of dynamics.

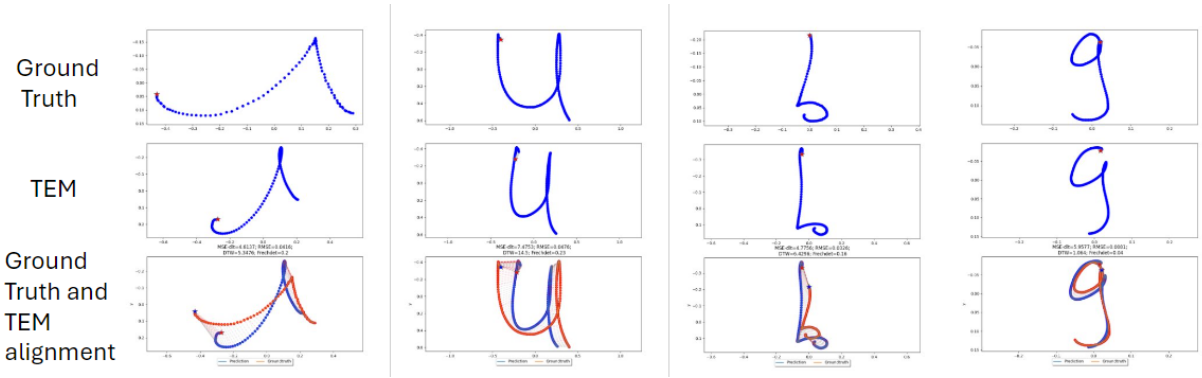


Figure 9.3 – On the first line the ground truth. On the second the TEM prediction. The observation is that the first points are less well reconstructed. On the last line the alignment between ground truth and the prediction.

To take this physical aspect into account, dynamic context is given to input during the network training phase. For that, pen-up movements preceding the touching strokes are added in the input sequences (in red on Fig. 9.4). The size of this pen-up movements

corresponds to half of the receptive field that can be captured by the model on the left border of the touching strokes, so that the model sees no padding to predict the positions associated with the first touching values. In addition, this will enable the network to see real signals instead of padding and thus have a better generalization capability.

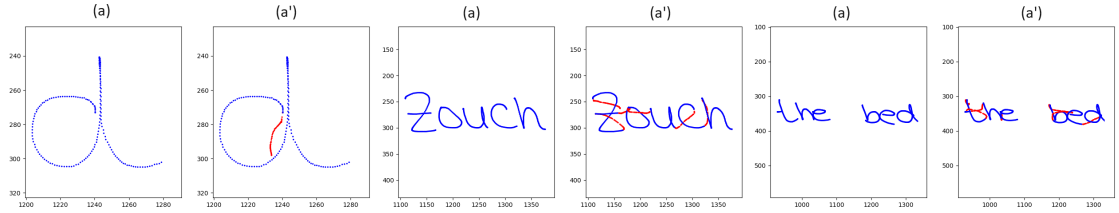


Figure 9.4 – Illustration of ground truth with and without context (in red) beginning each stroke. We add a 2D projection of the 3D pen-up movements captured by the tablet to replace the padding seen on the input with real data. This figure is a representation of the trace whose associated input is seen during learning.

9.3 Evaluation protocol

As discussed in Chapter 4, the Fréchet distance is a good metric to evaluate the trajectory reconstruction, due to its more intuitive measure of geometric shape similarity than DTW or MSE, which corresponds to our expectations in this work. This is why we use this metric to evaluate trajectory reconstructions.

The Fréchet distance measures the distance between curves, by taking into account the location and ordering of the points along the curves [Har-Peled et al., 2002]. This allows us to capture both local and global information accurately, and correlates well with qualitative assessment in practice.

Since we are interested in the shape of the reconstruction and not its size, we propose 2 additional steps before calculating the Fréchet distance (Fig. 9.5).

The first step is to find the longest dimension of the ground truth bounding box (resp. the reconstruction), and set its size to 1. The aim is not to give too much weight in the evaluation to the size of reconstruction, but to the overall quality of shape reconstructions. The second step consists in centering the centroids of the prediction and ground truth bounding boxes. As the Fréchet distance is not directly impacted by the length of sequences, we can have a quantitative analysis of the impact of sequence length on reconstruction quality.

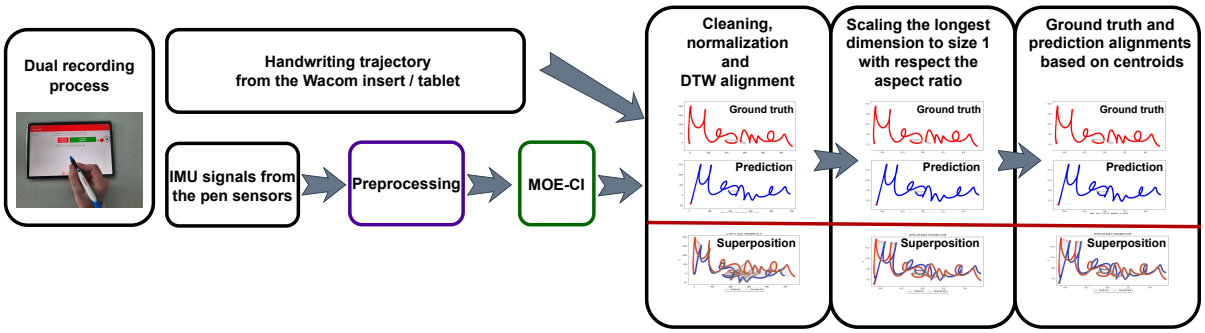


Figure 9.5 – Our evaluation pipeline, composed of four steps, dual acquisition of Digipen signals and ground truth, alignment with DTW to find identical sequence sizes, scaling and centering on centroids.

TEM/TEM-C are trained using touching strokes but are designed to predict both touching and pen-up strokes. During evaluation, we assess the full label, which includes both the touching and pen-up strokes. This approach allows us to evaluate the overall quality of the reconstruction by considering both the actual writing components (touching strokes) and any repositioning errors (pen-up strokes). In doing so, we ensure that the model performs well across the entire motion, not just the writing segments.

9.4 Experiment

9.4.1 Comparison with state-of-the-art

We compare the performance of our approach against the one of [Wehbi et al., 2022] on our own dataset, called IRISA-KIHT. We also provide results on the subset of the latter called IRISA-KIHT-S which is made publicly available and can hence be used as benchmark for future works. Since [Wehbi et al., 2022] provided a dataset called FAU-EINNS with their approach, we also compared ourselves on their dataset. It has the peculiarity of having been collected with Digiven v5.0, i.e. at 100Hz, as opposed to 400Hz for our Digipen v6.0. The FAU-EINNS dataset consists of words written by 6 writers, and we take the samples of the user numbers 1, 2, 3, 5 and 6 for training (1774 samples) and testing on the 344 samples of the user numbers 4. We applied the same evaluation protocol on both models and processing pipelines. Due to the comparatively smaller size of IRISA-KIHT-S, a 3 fold cross validation is used when the dataset is at stake. To

create the folds, we selected the 9 recordings from the test set of the IRISA-KIHT dataset and added one additional random recording. Additionally, we randomly chose 10 other recordings, each from a different writer, to complete the two other folds. Note that the test sets are not identical between these different datasets.

Table 9.1 – Average Fréchet distance of our pipeline compared to [Wehbi et al., 2022]. For IRISA-KIHT-S dataset on 3 folds with standard deviation.

	[Wehbi et al., 2022]	Our approach
FAU-EINNS	0.463	**0.277**
IRISA-KIHT	0.669	**0.404**
IRISA-KIHT-S	0.654 ± 0.048	**0.433 ± 0.057**

Results show that our pipeline (TEM and DTW based preprocessing) outperforms [Wehbi et al., 2022] (CNN model and linear interpolation as preprocessing) on every dataset, both quantitatively as seen in Table 9.1 on the average Fréchet distance and qualitatively as illustrated in Figure 9.6 and 9.7. The differences in alignment (using DTW vs. linear methods and TCN vs. CNN) appear to result in more accurate trajectory reconstructions. Additionally, Wehbi's approach exhibits a more pronounced drift phenomenon.

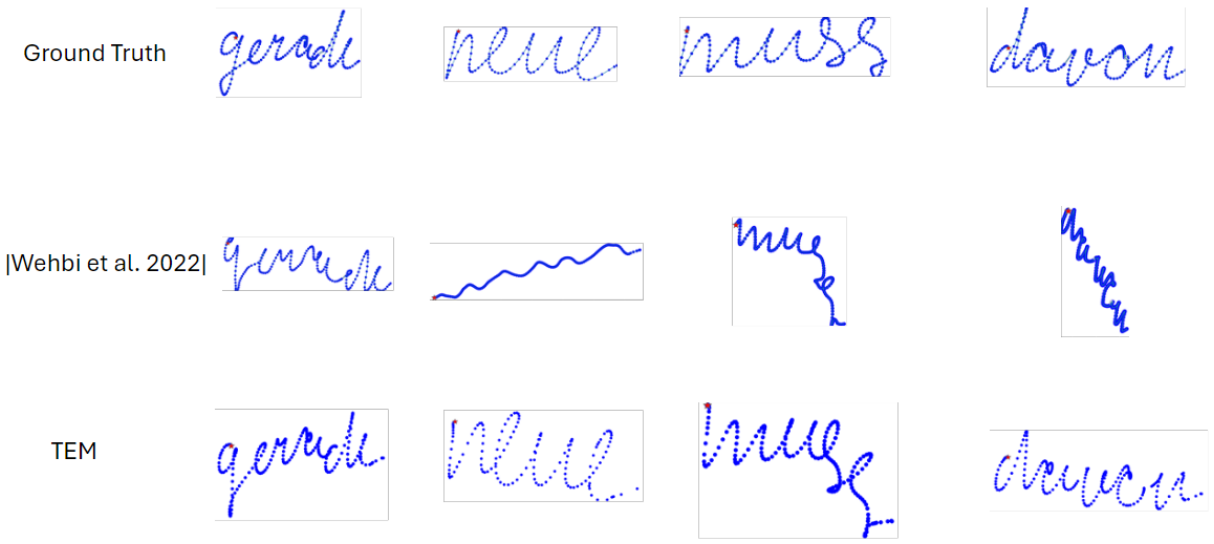


Figure 9.6 – Illustrations of handwriting reconstructions on the FAU-EINNS dataset. First line: the ground truth; Second line: the approach proposed by [Wehbi et al., 2022]; Thrid line: our proposed approach based on the TEM.

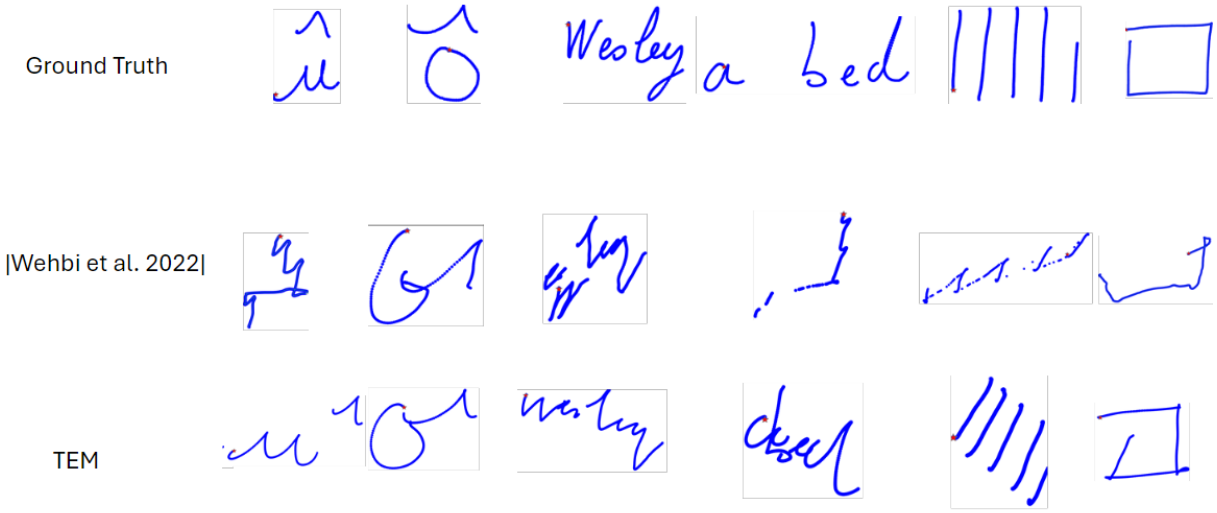


Figure 9.7 – Illustrations of handwriting reconstructions on the IRISA-KIHT dataset. First line: the ground truth; Second line: the approach proposed by [Wehbi et al., 2022]; Thrid line: our proposed approach based on the TEM.

Despite the difference in size between IRISA-KIHT and IRISA-KIHT-S datasets and different test set, the same conclusion can be makes for both datasets (Table 9.1). This confirms that IRISA-KIHT-S seems of sufficient size for the quantitative assessment of online handwriting trajectory reconstruction from IMU sensor data. In the next section, we will evaluate the impact of preprocessing or models on performance.

9.4.2 Ablation study

9.4.2.1 Alignment methods and models

In the following, we compare two effective processes of the handwriting trajectory reconstruction pipeline; (i) the input-output alignment process (linear interpolation [Wehbi et al., 2022] versus our DTW alignment) and (ii) the reconstruction model performance (CNN [Wehbi et al., 2022] versus TEM).

We provide a comparative evaluation scheme between the two alignment methods (linear interpolation and DTW based alignment methods) and the two reconstruction models (CNN and TEM) on the FAU-EINNS and IRISA-KIHT datasets. From Table 9.2, we observe that each step of our approach outperforms the counterpart from the [Wehbi

et al., 2022] method on both datasets. The results show that our alignment is better whatever the model on all the datasets. In contrast to the linear alignment, our alignment based on the DTW algorithm keeps the writing dynamic, which seems to be essential to reach quality trajectory reconstruction. Figure 9.9 shows examples of reconstruction trajectories using the CNN or TCN models, when using the linear interpolation and the DTW alignments on the two datasets. Moreover, the results are close and slightly better whatever the alignment method. On the other hand, the visuals show that the overall quality is better, which shows that the TCN model benefits from a larger receptive field and a deeper network to model more complex patterns and to be less sensitive to the noise.

Table 9.2 – Model and alignment method comparison between [Wehbi et al., 2022] and ours on both datasets.

	CNN model [Wehbi et al., 2022]		Our TEM	
	[Wehbi et al., 2022] alignment	Our DTW alignment	[Wehbi et al., 2022] alignment	Our DTW alignment
FAU-EINNS 100Hz	0.463	0.280	0.321	**0.277**
IRISA-KIHT 400Hz	0.669	0.410	0.586	**0.404**

9.4.2.2 Receptive field effect

To design the TEM architecture, we pay attention to the sampling rate of the input signal of sensors. The TCN based neural network may have the capacity to absorb the noise of such low quality and noisy signals, depending on the size of its receptive field.

In order to evaluate the receptive field effect, we trained and evaluated the model with different sizes of receptive fields equal to 49, 85, 168 and 373 on IRISA-KIHT dataset. These results are presented in the following table 9.3.

On our IRISA-KIHT dataset, the TCN model with the greatest receptive field (373) generally outperforms the other sizes of receptive field as seen in Table 9.3. With a bigger receptive field, a large context can be exploited to make the prediction and the signal noise of long pen-up movements can be absorbed, at the cost of a larger number of learnable parameters.

Table 9.4 illustrates the size of TCN models in terms of number of learnable parameters with regard to the receptive fields.

By comparing TCN-49 and TCN-373, during pen-up movements, the network may see the last and next touching points of two successive touching strokes, as illustrated in

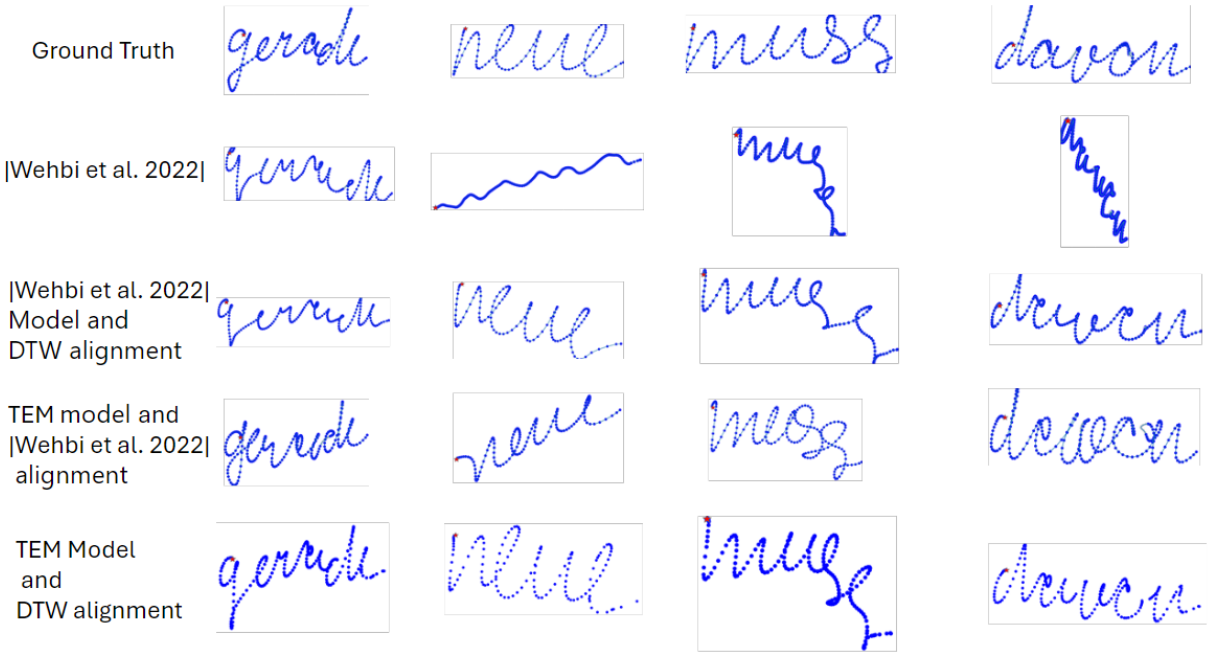


Figure 9.8 – Comparison between [Wehbi et al., 2022] and ours on FAU-EINNS dataset. Note that TEM model and DTW alignment is our comprehensive approach.

Figure 9.9 – Comparison between [Wehbi et al., 2022] and ours on IRISA-KIHT dataset. Note that TEM model and DTW alignment is our comprehensive approach.

Table 9.3 – Fréchet distance from TCN model with varied receptive fields on IRISA-KIHT dataset

Type	TCN-49	TCN-85	TCN-169	TCN-373
Global	0.404	0.442	0.404	**0.381**
Characters	0.476	0.505	**0.436**	0.440
Words	0.339	0.372	0.361	**0.321**
Equation	**0.344**	0.413	0.396	0.346
Shapes	0.481	0.507	0.476	**0.451**
Word groups	0.355	0.397	0.403	**0.327**

Table 9.3 with better results on word groups on repositioning the next stroke. These two models seem to yield fairly similar reconstructions (Fig. 9.10).

The one of the goal of the KIHT project is to create an autonomous pen, which necessitates that the reconstruction process is conducted internally within the Digipen. Therefore, we decided to keep the TCN-49 model to get a good trade-off between the model performance and the number of parameters (nearly 2 times fewer parameters).

9.4.2.3 Touching versus pen-up trajectories reconstruction

We focused on training the model using only the "touching" strokes, as outlined above. The goal was to evaluate the model's performance on both touching and pen-up strokes after being trained exclusively on touching strokes. To assess the model's effectiveness, we compared its performance when trained solely on touching strokes versus when trained on a combination of both touching and pen-up strokes. It's important to note that during training on the entire sequence, if the pen-up distance exceeds the threshold (where the trace is no longer tracked), the closest point identified by Dynamic Time Warping (DTW) is retained.

Table 9.5 shows that the strategy of learning on touching strokes only is generally better than the one of learning on touching and pen-up strokes upon the distance of Fréchet, and this is also the case for characters, words, equations and shapes categories. Results on the sentence category are very close. We think this is due to the correlated

Table 9.4 – TCN models' size with various receptive fields

# Params	TCN-49	TCN-85	TCN-169	TCN-373
Total	467,452	528,452	870,452	894,452
Trainable	464,152	524,752	866,752	888,352

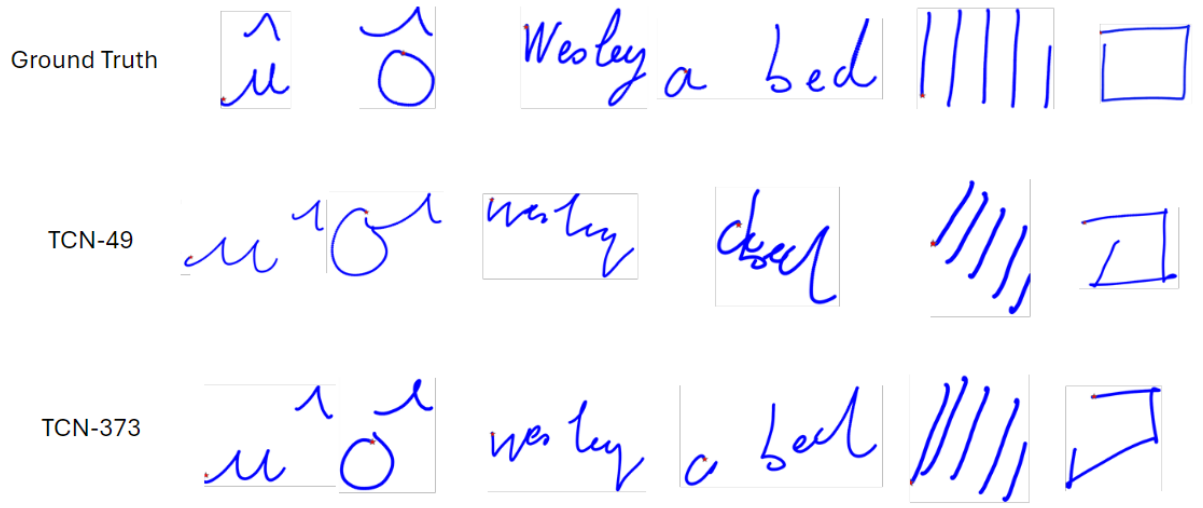


Figure 9.10 – Handwriting reconstructions with two TCN models with different receptive fields on the IRISA-KIHT datasets. First line, ground-truth. Second, the 49-receptive-field TCN model. Third line: TCN model with 373 receptive field.

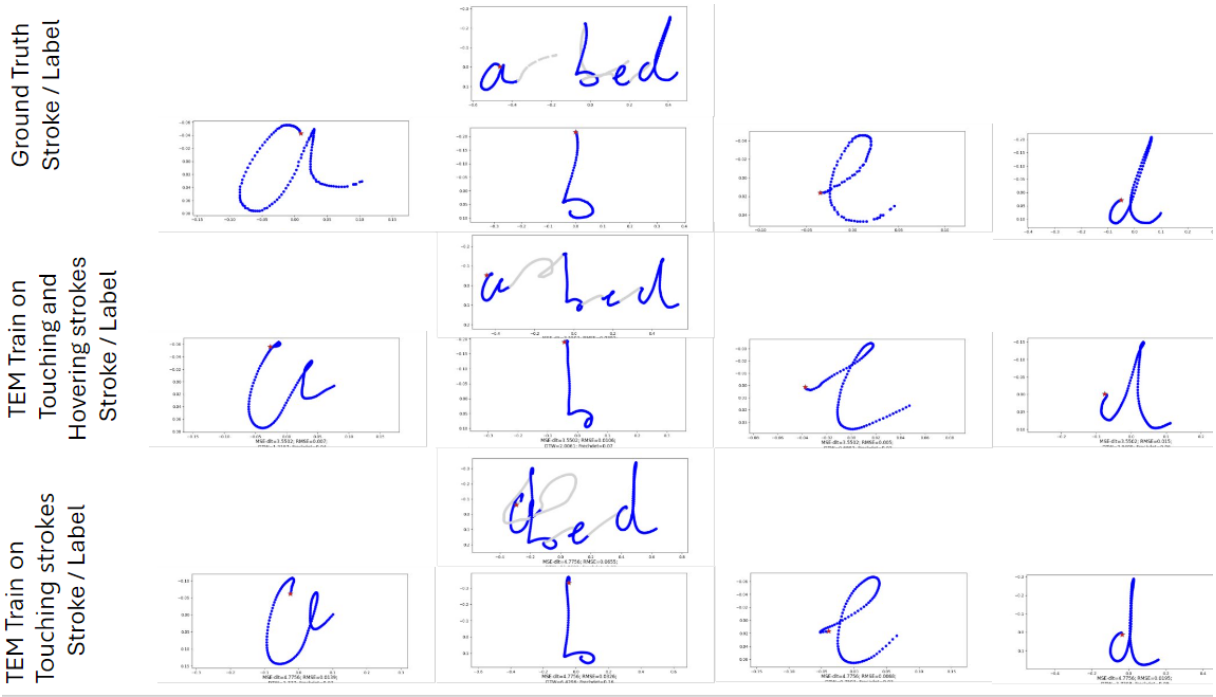


Figure 9.11 – TCN-49 model performance when training on pen-up and touching data (middle) and touching data only (bottom). The ground truth is on first line.

Table 9.5 – Training on touching strokes versus training on touching & pen-up strokes results

Type	Our TEM	
	Trained on touching & pen-up strokes	Trained on touching strokes only
Global	0.447	**0.404**
Characters	0.559	**0.476**
Words	0.350	**0.338**
Equations	0.348	**0.344**
Shapes	0.629	**0.481**
Word groups	**0.350**	0.355

effect of the untracked pen-up movements and the pen-movement hesitations of the writer between the words of the sentence.

However, by looking at Figure 9.11, we observe that the touching model reconstructs better touching strokes. The only part where it is not as accurate is on pen-up movements prediction. This is logical, as the model has not been trained on signals that include the third dimension.

Similarly, we observe that our model better reconstructs the touching parts of the word groups but it fails to find where to start the reconstruction of the next stroke. This may happen due to untracked pen-up movements (that represents movement hesitations about where to put the pen again on the screen) like in "a — bed" where there is a long untracked pen-up movements between "a" and "bed".

9.4.2.4 Temporal context integration in input of the touching expert (TEM-C)

As presented in section 9.2, we suggest to integrate temporal context in input to the Touching Expert Model. Results are shown at stroke level both quantitatively in Tab. 9.6 and qualitatively in Fig. 9.12. Adding temporal context significantly improves the model's ability to capture dynamic movement. By adding an earlier pen-up portion to the touching strokes inputs, the model can now account for past dynamics, providing a more comprehensive understanding of the trajectory. It also appears to enhance stroke reconstruction and improve repositioning, as the model can see during training sections of 3D signals.

Table 9.6 – TEM and TEM-C comparison at label and stroke levels using the Fréchet distance.

Evaluation	TEM	TEM-C
Label level	0.404	**0.308**
Stroke level	0.097	**0.091**

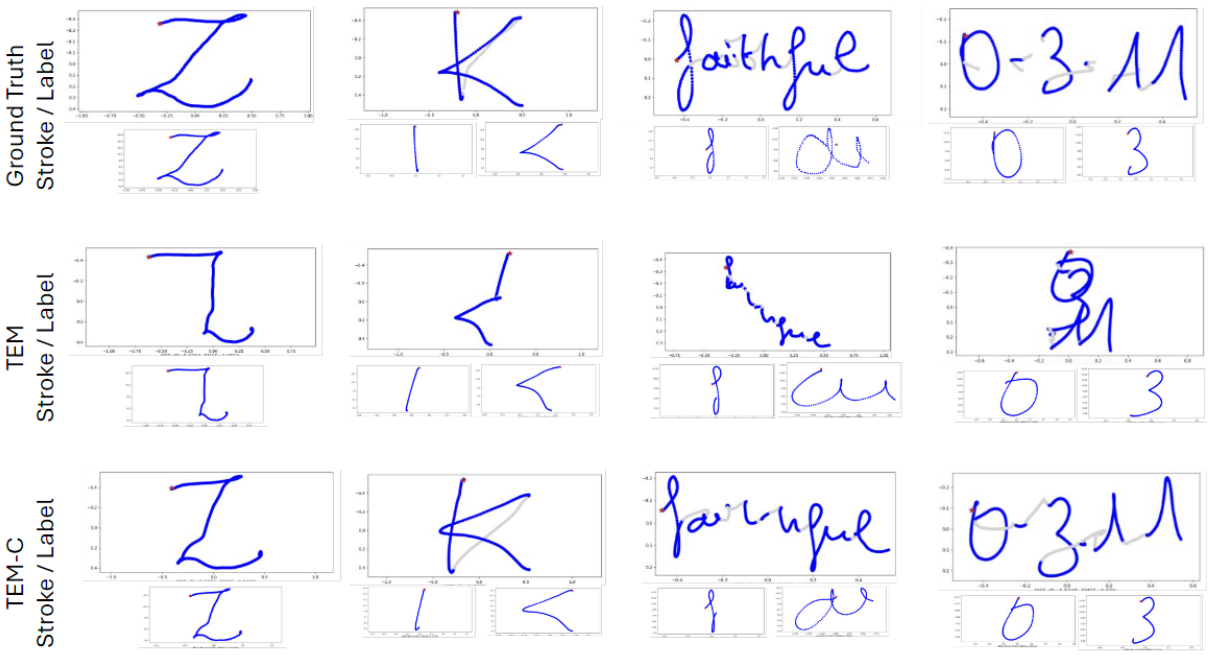


Figure 9.12 – Comparison between the touching expert model (TEM) and its variant (TEM-C), on the first line the ground truth at stroke (down) and label (up) level, on the second line the expert model dedicated to touching strokes, on the last line the TEM improved by adding temporal context (TEM-C).

9.5 Conclusion

We present two new models and demonstrated their performance, as well as the benefits of the preprocessing introduced in the previous chapter.

To address the discrepancies in sampling rates between the Digipen and the tablet, as well as synchronization issues, we proposed a Dynamic Time Warping (DTW) based alignment approach. Through experimentation the multiple datasets we have made as part of the KIHT project and one from the state-of-the-art, we demonstrate that the our approach surpasses the [Wehbi et al., 2022] proposal.

The first factor is the alignment with DTW, which preserves its dynamic nature and significantly enhances prediction accuracy. The second offers visually much better reconstruction thanks to the wider context.

While the TEM model achieved strong reconstruction results at the stroke level, the challenge of accurately predicting the initial points of a stroke where dynamic motion is cut. To address this, we proposed a solution to incorporate the preceding pen-up trajectory into the input data during the network training phase. By extending the input sequences to include pen-up signals, the network is exposed to a richer set of dynamic patterns, thereby improving its generalization capability and ability to predict the first touching values without relying on padding. Which helps to improve the accuracy of what we predict.

However, a significant limitation of our approach is modeling untracked and complex pen-up trajectories, which remains a challenge. We present a new proposal to meet this challenge in the following chapter.

9.6 Related publications

The preprocessing work presented in the previous chapter and the first TEM model gave rise to a scientific publication [Swaileh et al., 2023], a presentation at the SIFED symposium [Imbert et al., 2022], and served as a baseline for the KIT colleague working on the hardware part with whom we have produced a joint publication [Serdyuk et al., 2023].

Chapter 10

A MIXTURE OF EXPERT MODEL FOR BETTER COLLABORATION WITH TASK SPECIFICATION

Our TEM-C presented in the previous chapter, the third attempt to reconstruct handwriting trajectories from the Digipen using deep neural networks, has produced promising results in terms of handwriting reconstruction. It is dedicated to the reconstruction of touching parts (except for the contextual parts of TEM-C, which are quite marginal ie the 24 points are extremely small compared to the length of a stroke.), which are the only parts for which a reliable ground truth is available during training. Unlike touching trajectories, which are 2D trajectories where ground truth can be obtained using double Digipen-Wacom acquisitions (Fig. 2.1), pen-up trajectories are 3D trajectories where ground truth signals can only be recovered up to a 7mm height. TEM-C, learned only from 2D data where we have the ground truth, fails to process 3D trajectories that correspond to pen-up movements.

Given the positive outcomes of TEM-C regarding pencil touching, but the significantly degraded results during the pen-up phases (Fig. 10.1), we should consider adjustments to enhance performance. We want a more global approach to deal both touching and pen-up movements.

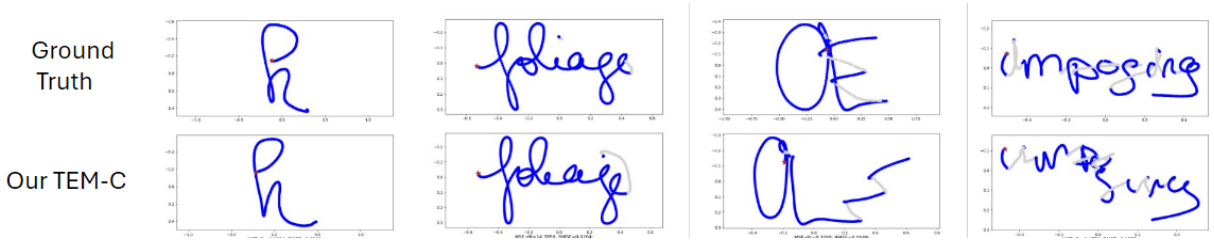


Figure 10.1 – Our TEM-C, which performs very well on touching strokes but has difficulty for predicting pen-up movements.

We propose an approach consisting of two expert networks to deal with the different nature of IMU signals. In the scenario of pencil-based handwriting, we observe two distinct phases: the pencil touching phase, which produces a 2-dimensional trajectory, and the pencil pen-up phase, which involves a 3-dimensional target trajectory as the pencil moves towards the next point of contact. These two phases differ in both the nature of the signals they generate and the associated ground truth data.

During the pencil touching phase, the pencil directly contacts the surface, resulting in a 2D trajectory that accurately captures the written path. This trajectory reflects the actual movements involved in writing, providing clear and reliable ground truth data.

However, once the pencil is lifted off the surface, the dynamics change significantly. In this pen-up phase, the pencil's motion becomes 3D, as it moves through space towards the next part of the writing. The target trajectory during this phase involves the pencil's position in 3D space, where height (z-axis) is introduced in addition to the usual 2D coordinates (x and y). Unlike the pencil touching phase, where the ground truth is the writing itself, the ground truth for this pen-up movement is more ambiguous. If the user raises the pencil too high, it can lead to a loss of reliable 3D ground truth. When the ground truth is lost, learning becomes either impossible or dependent on interpolated points that do not accurately reflect the real trajectory. Dynamic time warping (DTW) alignment can introduce discrepancies and disrupt the model's dynamics. Due to the absence of a ground-truth trajectory during a pen-up movement, DTW applied over time associates part of the elevated trajectory with the last point of the previous stroke, and another part with the first point of the following stroke. As discussed in the previous chapter, if we attempt to train a model using data that includes both touching and pen-up, we degrade the quality of the touching reconstruction. To address this, it is essential to keep a high-performance model specifically dedicated to touching interactions. By introducing an expert model for pen-up movements, the system can compartmentalize the learning tasks, ensuring that the touching model remains robust and unaffected by the inconsistencies and variability inherent in pen-up movements data. This approach preserves the integrity and performance of the touching model while addressing the unique requirements of pen-up movements interactions.

In this way, we propose a mixture-of-experts made of two expert neural networks, one for the touching trajectory, the other for the pen-up parts of the trajectory. This model is called MOE-C, combines the TEM-C seen before with the Pen-up Expert Model, called PEM. Additionally, we introduce a variant of the Pen-up Expert Model, called PEM-I, which accounts for the extra dimension associated with the pen's height during the

pen-up phases. This variant uses new, specialized training data to enhance the expert model dedicated to these pen-up parts. The resulting mixture-of-experts model is called MOE-CI.

These two approaches are described in detail in the following sections, followed by an experimental section.

10.1 MOE-C: a new mixture of expert model

We can formalize our problem as multitask learning, in that we have two linked tasks, the first being the prediction of the writing itself, the second the pen-up movement between these different parts. These two tasks differ in their dynamics and nature (2-dimensional vs 3-dimensional signals). In practice, this means combining two specific neural networks. Inspired by our previous work where the TEM-C is trained on touching strokes only producing degraded reconstructions on pen-up parts, we suggest to use the same network architecture as the expert model for touching strokes.

As a reminder and for reproducibility purposes, the TCN model (Fig. 9.2) is based on 4 blocks of a non-causal TCN followed by 2 dense layers. Each TCN block is composed of 2 convolutions with dilation 1 and 2 respectively and a kernel size of 3. To reconstruct pen-up strokes, we propose the same architecture but with different learning strategies.

We therefore turned our attention on how to train our PEM model, in particular with regard to the type of data given as input to the network. Addressing the specificity of pen-up strokes, we suggest to use the backbone model on complete sequences, because we believe that giving as much context as possible can be beneficial for pen-up prediction. The first reason for training our network on entire sequences, rather than isolating pen-up strokes, comes from the complex dynamics of pen-to-tablet interactions. By having the full sequence, the model gains insights into the transition patterns between active stylus contact and pen-up states. Another hypothesis comes from physics: we are in the process of integration of a signal, so we need information about the initial conditions.

The result is a global reconstruction of the handwriting where two models are trained in parallel. Thus, this approach acts as a mixture-of-experts, one on the touching that corresponds to the handwriting itself, the other trained on complete sequences is dedicated to predicting pen-up part. Contrary to standard MOE models, switching from one model to the other, our mixture is controlled through the pen's pressure sensor. In this way, each expert better captures its own specificities.

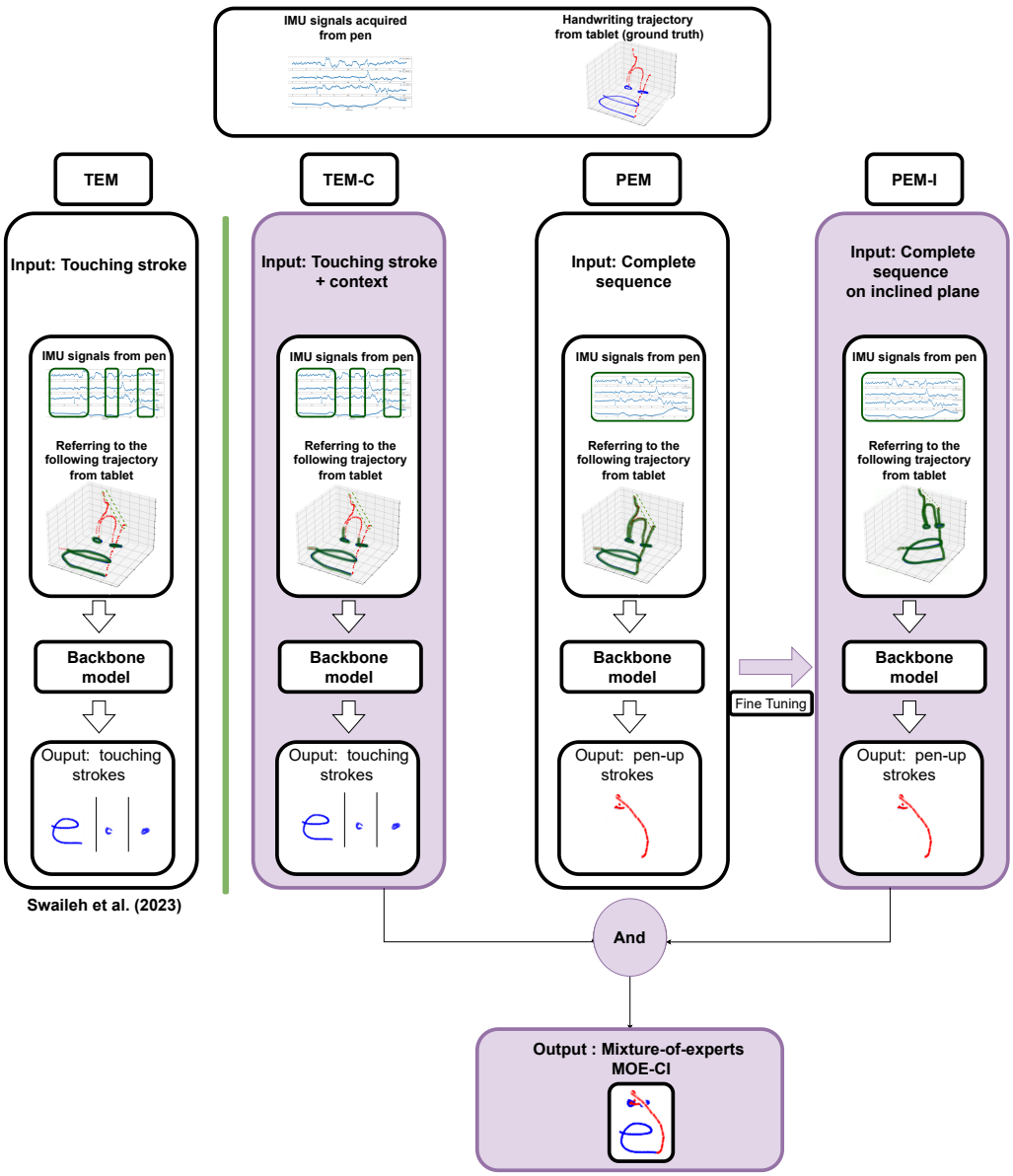


Figure 10.2 – Our MOE-C / MOE-CI approaches merge advancements for the two expert models: TEM-C, which improves the Touching Expert Model (TEM) by integrating pen-up context (in green), which gives more context and facilitates the transition from pen-up to touching, PEM-I, which enhances the Pen-up Expert Model (PEM) by a fine tuning on 3D data for a better understanding of this third dimension.

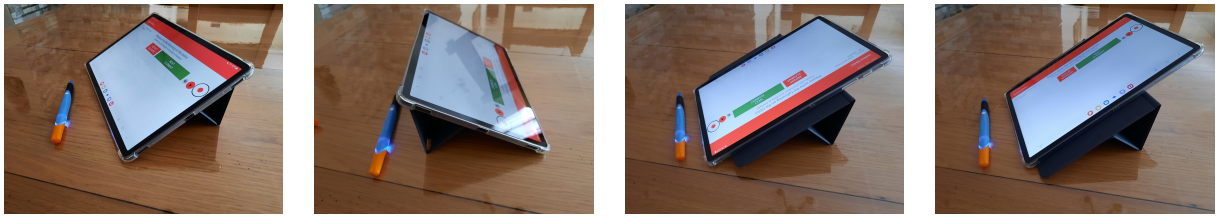
We call this approach MOE-C (Fig 10.2), which is the combination of the 2 following experts:

— the TEM-C for Touching Expert Model with pen-up Context: its corresponds to the backbone model trained with touching stoke with pen-up portions preceding the touching strokes, presented in section 9.2;

— the PEM for Pen-up Expert Model, its corresponds to the backbone model trained with complete sequence named Pen-up Expert Model (PEM).

10.2 MOE-CI: Training on 3D labeled samples

Handling the pen-up phase in trajectory reconstruction presents unique challenges due to distinct dynamics compared to writing segments. While the touching parts share a common 2D plane, pen-up movements introduces a third dimension, representing the height of the pen-up movements, adding complexity to sensor data as well as the variability of unconstrained pen-up trajectories. To address this variability, we refined our approach by fine-tuning a dedicated network using data acquired on inclined planes. In this way, we expect that the fine-tuned network demonstrates enhanced adaptability to variations in pen-up height, resulting in more robust predictions for pen-up segments. Acquisition protocol to acquire inclined examples (Fig. 10.3) includes several positions to introduce variability in the inclination of the writing surface. Four setups are considered, positioning the tablet horizontally with a 30-degree upward or downward inclination, and vertically with similar 30-degree inclinations upwards or downwards.



(a) Horizontal tablet inclined upward

(b) Horizontal tablet inclined downward

(c) Vertical tablet inclined upward

(d) Vertical tablet inclined downward

Figure 10.3 – Data acquisition protocol on inclined planes

We call this approach MOE-CI (Fig 10.2), which is the combination of the 2 following experts:

— the TEM-C for Touching Expert Model with pen-up Context: its corresponds to the backbone model trained with touching stoke with pen-up portions preceding the touching strokes, presented in section 9.2;

— the PEM-I for Pen-up Expert Model fine tuned on inclined data, its corresponds to the backbone model trained with complete sequence followed by a fine tuning on data acquired on inclined plane.

10.3 Experiments

First, we report a comparative analysis of our first approach based on touching strokes and the novel mixture-of-experts approaches (MOE-C / MOE-CI) against the established [Wehbi et al., 2022] methodology. We then proceed to an ablation study to analyze in detail the contributions. We will focus on the specific contribution of each expert and their impact on overall reconstruction. We do a comparative exploration of possible mixtures, in order to better understand the improvements of each expert in their collaboration together in the mixture-of-experts. Finally, we offer an overview of the approaches used on private and public databases, providing an understanding of the impact of data quantity and diversity. Note that the databases have been evolved and been enriched throughout the course of this thesis and the associated project. Unlike in the previous chapter, the experiments in this section are conducted using the KIHT-Private and KIHT-Public databases.

10.3.1 Comparison of our approach with state-of-the-art methods on the KIHT-Private dataset

Using the evaluation protocol and datasets that have been presented, we compared our TCN model learned on touching strokes only and MOE approaches to the work of [Wehbi et al., 2022], the reference in the field.

As a reminder, they proposed an approach based on a CNN model with a linear interpolation between sensor signals and ground truth. These results are presented qualitatively (Fig. 10.4) and quantitatively (Table. 10.1).

The Fréchet distances at label and stroke level are significantly better for our mixture-of-experts. We can see that on the touching parts, MOE-C and TEM-C performance is the same, as the same models are used on this part. Adding 3D data into the training of the Pen-up expert model, as proposed in our MOE-CI, slightly improves the efficiency. It can be observed that our mixture-of-expert retains the strengths of both models, resulting in a significant improvement of the reconstruction for both the touching and the pen-up part. The gain is particularly noticeable at label level, demonstrating a better estimation of the

pen-up trajectory for repositioning the next stroke. Note that differences in performance may occur at stroke level for the same expert as the bounding box normalizations are applied at label level.

Table 10.1 – Fréchet distance computed for our TEM-C, mixture-of-experts and the state-of-the-art methods trained on the KIHT-Private dataset, and evaluated on the test set.

	[Wehbi et al., 2022]	Our: TEM-C	Our: MOE-C	Our: MOE-CI
Label level	0.571	0.437	0.321	**0.312**
Stroke level	0.120	0.097	0.097	**0.091**

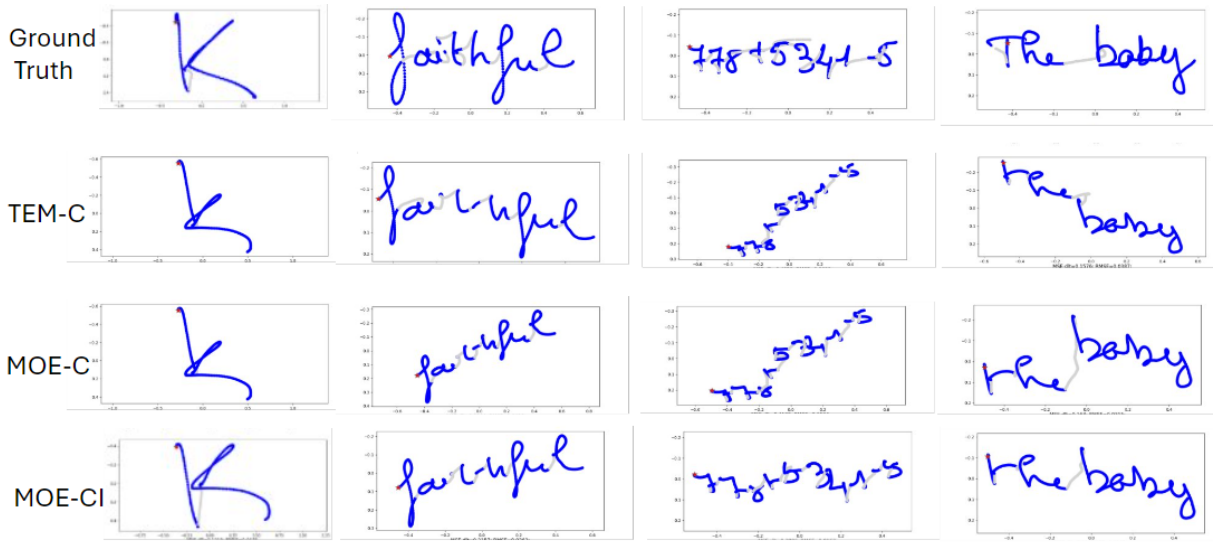


Figure 10.4 – Comparison of our approaches, on the first line the ground truth, on the second the reconstruction with the TEM-C, on the third and fourth lines ours mixture-of-experts approaches: MOE-C / MOE-CI.

10.3.2 Ablation study

In this section, we focus on the ablation study, to evaluate the impact of different contributions on the overall performance of our different approaches.

We focus on the impact of different contributions on the overall performance of our mixture-of-experts. We first explore the impact of the extra dimension on the pen-up expert (PEM vs PEM-I) (cf. 10.3.2.1). This ablation study was performed on the KIHT-Private dataset. As a reminder, the contribution of temporal context in TEM-C versus TEM was shown in Section 9.2.

10.3.2.1 Fine tuning on extra dimension for the pen-up expert (PEM-I)

We now evaluate the training of the pen-up expert on inclined data, including the extra dimension (discussed in section 10.2). Fine-tuning the PEM model on data acquired from inclined planes has yielded remarkable improvements in prediction accuracy. It can be seen both at the stroke and label level, quantitatively in Table 10.2 and qualitatively in Fig. 10.5. Our expert model seems to tackle the inherent variability in sensors during pen-up parts, that is due to different dynamics between writing and pen-up strokes and from having the height dimension varying in the data. Thus, using data acquired from inclined planes to simulate various pencil heights, we have successfully dealt with the height dimension in sensors inputs. The fine-tuned model (PEM-I) demonstrates enhanced adaptability to variations in pen-up height, resulting in a more robust expert model for predicting pen-up segments as shown on the Fréchet distance at label level. Additionally, improvements are observed at the stroke level, attributable to a more accurate orientation of characters in their reconstructions. This enhancement stems from a more precise understanding of spatial relationships.

Table 10.2 – PEM and PEM-I comparison at label and stroke levels using the Fréchet distance.

Evaluation	PEM	PEM-I
Label level	0.413	**0.350**
Stroke level	0.129	**0.114**

10.3.2.2 Comparison of model combinations into a mixture-of-experts

Now that each expert has been established, we evaluate the performance of each possible mixture combination (Fig. 10.6 and Table 10.3). We introduce some notation for the mixture-of-experts (MOE) :

— MOE: the combination touching expert model (TEM) & pen-up expert model (PEM);

— MOE-I: the combination touching expert model (TEM) & pen-up expert model fine tuned on inclined data (PEM-I);

— MOE-C: as a reminder, the combination touching expert model with temporal context (TEM-C) & pen-up expert model (PEM);

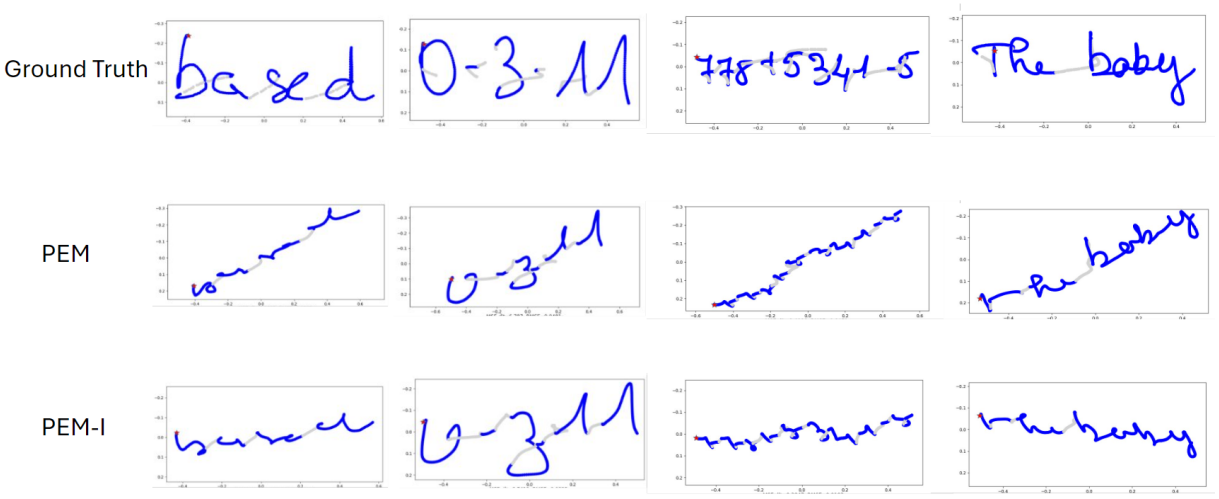


Figure 10.5 – Comparison between the Pen-up expert model without (PEM) and with (PEM-I) a fine-tuning on inclined data, on the first line the ground truth, on the second line the expert model dedicated to pen-up (PEM), on the last line the fine-tuned model on inclined data (PEM-I).

— MOE-CI: the combination touching expert model with temporal context (TEM-C) & pen-up expert model fine tuned on inclined data (PEM-I).

This evaluation shows us the benefits of different contributions within a mixture of models. The addition of context enables more accurate reconstruction of touching strokes, especially at their extremities. As a reminder, performance differences may occur at stroke level for the same expert as the bounding box normalizations are applied at label level. Fine-tuning on 3D data enables better repositioning of pen-up movements. In addition, the combination MOE-CI, which is the combination of the 2 improved experts, is actually the method that shows the best results, as expected. The MOE-CI is more often better than the others, which attests to the benefits of our contributions. Nevertheless, the MOE has correct performance due to good performances on short mono stroke examples, which are the easiest part to reconstruct.

Table 10.3 – Evaluation of possible mixtures of experts on the KIHT-Private dataset. The Fréchet distance is computed to evaluate the models at label and stroke level.

Evaluation	MOE	MOE-I	MOE-C	MOE-CI
Label level	0.344	0.321	0.333	**0.312**
Stroke level	0.096	0.091	0.096	**0.091**

Figure 10.6 – Comparison between the different combinations of mixture-of-experts, on the first line the ground truth, then from the top to bottom: MOE, MOE-C, MOE-I and MOE-CI.

10.3.2.3 Evaluation on the public dataset

We have released the KIHT-public dataset that will serve as a benchmark for futur research. We evaluate the performance of our mixture-of-experts (MOE-CI) and compare it both to the different expert combinations (MOE, MOE-C, MOE-I) to the state-of-the art approach [Wehbi et al., 2022] and our previous approach (TEM). The results (Table 10.4) are similar to those obtained on the private dataset. Indeed, the integration of an expert trained at the label level significantly improves performance on the related metric compared to the state-of-the-art, and thus improves the processing of repositioning the reconstructed handwriting after pen-up movements. The proposals associated to each expert improve the robustness of mixture-of-experts, both the temporal context that reflects physics and dynamics to enhance the touching expert model, or refining the pen-up expert model with 3D labeled samples for improved pen-up movements predictions. These consistent results shows that the public dataset is relevant to be used as benchmark.

Table 10.4 – Fréchet distance computed for our mixture-of-experts and state-of-the-art methods trained on the KIHT-Public dataset, and evaluated on the test set.

Evaluation	[Wehbi et al., 2022]	TEM-C	MOE	MOE-I	MOE-C	MOE-CI
Label level	0.583	0.354	0.336	0.331	0.321	**0.320**
Stroke level	0.127	0.096	0.101	0.100	0.097	**0.096**

10.4 Conclusion

We introduce a mixture-of-experts approach where two models are tailored on a specific task. The first expert model is designed to predict touching strokes, processing 2-dimensional signals input. The second expert model focuses on predicting trajectory repositioning between touching strokes, handling 3-dimensional inputs due to pen-up movements. To optimize this second expert for pen-up movements, we fine-tuned it using data collected on an inclined plane to leverage variations in the height dimension during training. Our experiments on two datasets show that our mixture-of-experts framework surpasses the performance of the two leading state-of-the-art methods. This is particularly true of long sequences such as equations or sentences, in which pen-up movement is the most common component. And a more marginal effect on characters that are mainly made up of touching.

This work is dedicated to the handwriting reconstruction from data written on a tablet using the Digipen. As the Digipen can be used for learning to write in classroom, we will be interested in processing data from children whose dynamics are different from those of adults. This will be the subject of the next chapter.

10.5 Related publications

This mixture-of-expert approach is in submission to the Pattern Recognition journal. We also present it as part of a global article presenting the project by the four partners [Harbaum et al., 2024].

Chapter 11

Domain adaptation methods to process children data

11.1 Introduction

Although the Digipen stylus can be used on any surface and by different users, our first approaches focused on the reconstruction of handwriting from data written by adults on tablets. In fact, while collecting labelled adult data on a tablet is not a problem, it's more complicated to collect data from children as it is necessary to contact schools and have all parents sign parental authorizations. This requires additional time compared to collecting from adults. This is why we have less data for children.

Using the Digipen in another experimental context, e.g. on data written by children, or on another surface, e.g. on paper, leads to different input signals. On the one hand, children's handwriting are of variable speed and hesitant gestures depending on the assertiveness in the handwriting. On the other hand, handwriting on paper produces noisier signals due to friction than when the user writes on a tablet.

The KIHT project's application objective is to help people learn to write with a pencil that embeds AI, and to have a generic model whatever the user. This means having a generic model capable of handling multiple cases. As different inputs can produce the same writing trace, so it makes sense to build a shared representation that can be used for reconstruction. This leads us to consider a domain adaptation method for dealing with the different domains of data and capture the differences to build a single common representation of the signals. To our knowledge, no domain adaptation method addresses handwriting trajectory reconstruction from different sources, e.g. from adults vs children.

In this context, we present a domain adaptation approach designed to enhance the adaptability of our model to deal with the different data sources. For the first adaptation approach, we will focus on our TEM-C model, which addresses the core of the writing.

Initially trained on data acquired from tablets by adults, our model aims to effectively handle an additional types of data: handwriting acquired from tablets by children.

11.2 DANN-based method for handwriting reconstruction

The variability of handwriting sources complicates the task of handwriting reconstruction. Children handwriting poses a unique challenge due to the ongoing development of graphomotor skills, resulting in dynamic and inconsistent handwriting patterns as children learn to write. This translates into longer signal sequences than adults and a wider range of possible values (Fig. 11.2). Moreover, children's writing is slower, with inconsistent fluency marked by pauses and irregular accelerations, larger letter sizes, and more frequent pencil lifts complicating the reconstruction using a model trained on tablet-acquired data.

Our goal is to establish a unified model for all users, requiring a method that enables a shared representation for both source and target features. Based on the state of the art, we have observed that Domain-Adversarial Training of Neural Networks (DANN) effectively achieve this and have been successfully applied across various fields. Domain-Adversarial Training of Neural Networks [Ganin et al., 2016] are specialized neural network architectures designed to address the challenge of domain shift, where a model trained on one domain (source) is expected to perform well on a different but related domain (target). These networks operate by learning features that are domain-invariant, meaning they are useful and generalizable across both domains. This is typically achieved through a shared feature extractor on which additional components are built: a domain classifier and a task-specific classifier. The domain classifier is trained to determine the domain of the input data, whereas the task-specific classifier focuses on predicting the label of the dedicated task. In our context, the task-specific classifier is trained to reconstruct the handwriting trajectory.

The training process employs adversarial techniques by reversing the domain classifier's gradients during backpropagation, prompting the feature extractor to produce domain-invariant features that enhance performance on the target domain without needing extensive labeled data.

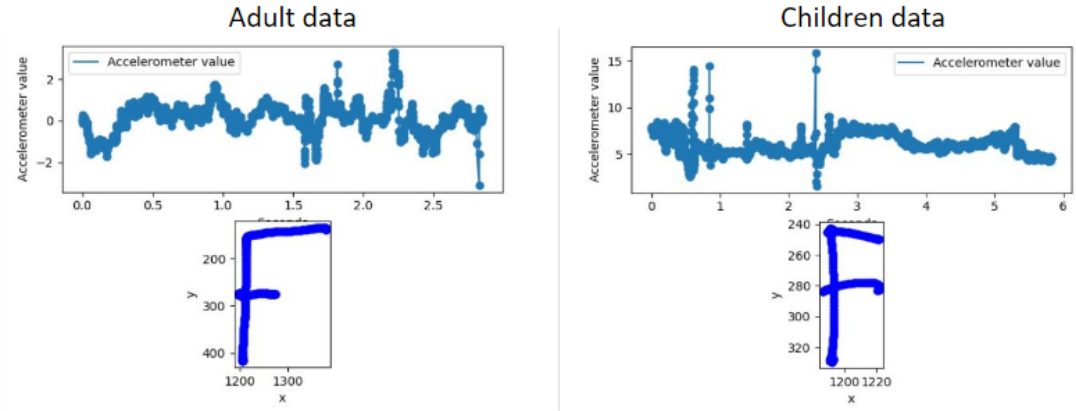


Figure 11.1 – Visualization of the x component of the Digipen's rear accelerometer over time in seconds, from left to the right: a F from adult on tablet, a F from children on tablet. We notice that the same pattern is written but not with the same fluency due to the different level of automation of handwriting (adult vs child), which results in slower writing for children and a jerky gesture that results in greater acceleration amplitude.

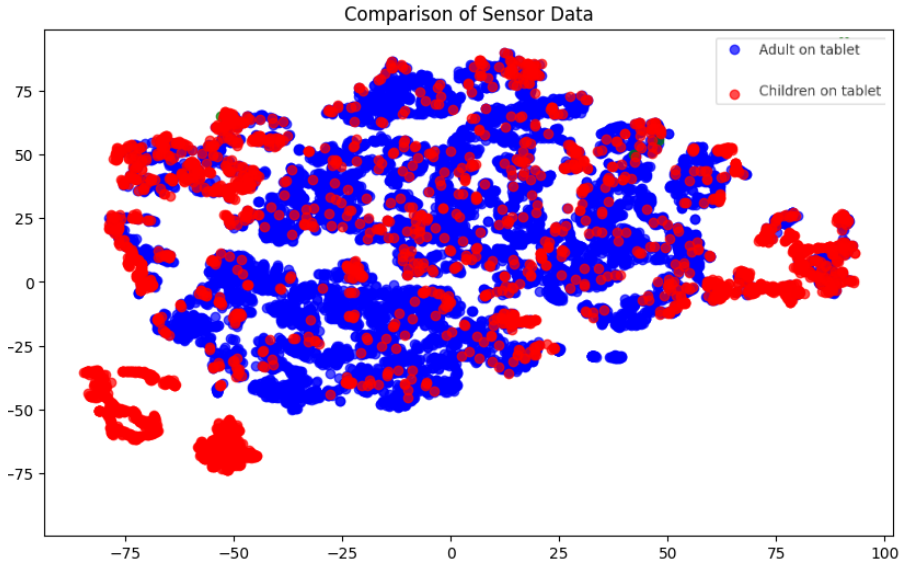


Figure 11.2 – Data visualization with the Multidimensional Scaling (MDS) method, we can see that children's data on tablets (in red) takes on a wider range of values due to their handwriting which is still being learned.

11.2.1 Application

In this work, the reconstruction part of the DANN is the TEM-C from the previous chapter 9. Then we have slice the baseline model as follows: the 4 blocks of the non-causal TCN is the feature extractor (in green in Fig. 11.3). Each TCN block is composed of 2 convolutions with dilation 1 and 2 respectively and a kernel size of 3. The next two dense

layers refer to the label predictor (in blue in Fig. 11.3) which in our context corresponds to the trajectory reconstruction. The domain classifier (in pink in the Figure 11.3), is made up of a max polling layer followed by a dense layer with 256 neurons using the ReLU activation function, and then another dense lyaer with a single neuron using a sigmoid activation function.

In our case study, we want to process both adult handwriting on a tablet (source) and children handwriting on a tablet (target). During training, we provide mixed matches to two model branches: (1) the feature extractor and the label predictor (green + blue in Fig. 11.3) pre-trained on adult data, and (2) a feature extractor and domain classifier (green + pink in Fig. 11.3).

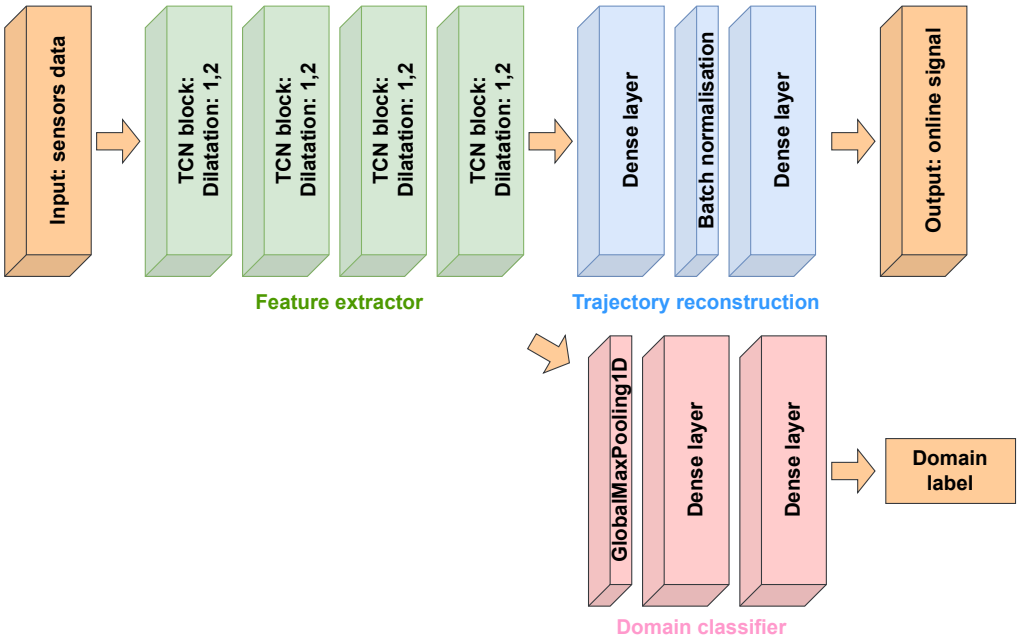


Figure 11.3 – DANN in our use case.

11.3 Experimental results

We experiment our approach on two datasets, one for adult tablet data (9629 samples), another for children tablet data (3910 samples). Each dataset contains characters, words, word groups, equations, and shapes. We compute the Fréchet distance to evaluate the quality of reconstruction on children's handwriting. We trained a DANN with children

and adult data with a Feature extractor and trajectory reconstruction model pretrained on adult data. We compared the DANN qualitatively (Fig. 11.4) and quantitatively (Table 11.1) to the following methods: baseline model trained on adult data and fine-tuned on children data and the baseline model trained from scratch on children data.

Table 11.1 – Comparison of reconstruction methods between training from scratch, fine tuning and domain adaptation method on **children test** data with the Fréchet distance.

Model	TEM-C			DANN
Pretraining Data	∅	∅	Adult	Adult
Training Data	Adult	Children	Children	Adult & Children
Fréchet distance	0.470	**0.345**	0.348	0.349

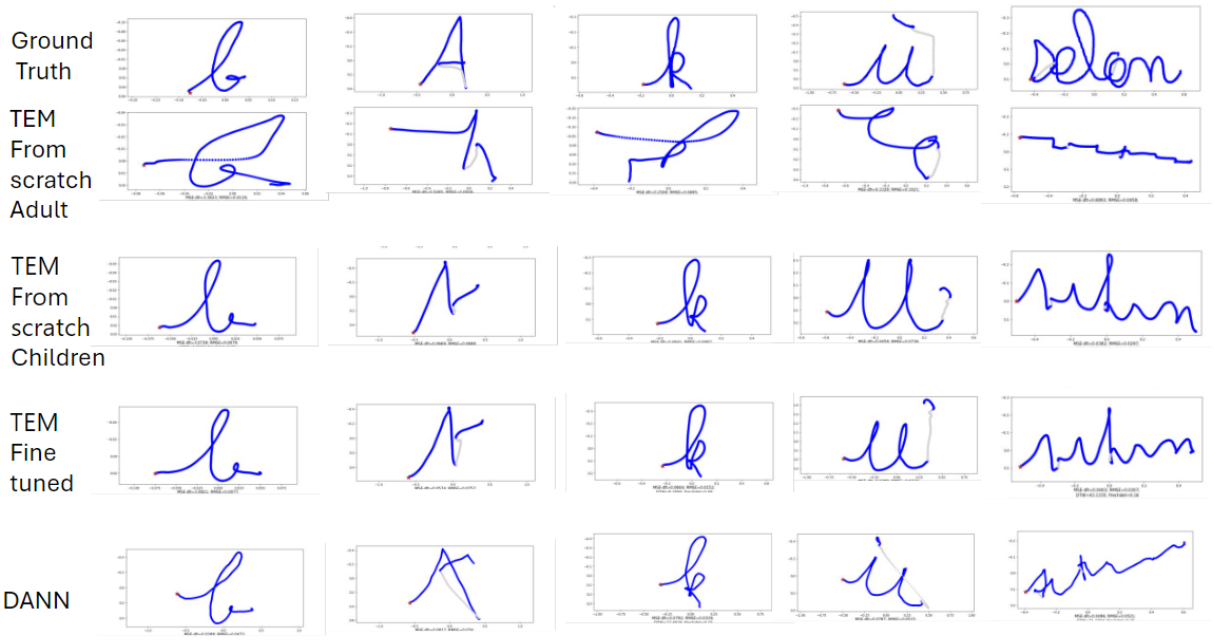


Figure 11.4 – Comparison of reconstruction methods on children test data, on the first line the ground truth, on the second the Baseline Model trained on adult data, on the third the Baseline Model trained on child data, then the Baseline Model trained on adult data and fine-tuned on child data, on the last line the DANN pretrain on adult data.

Table 11.1 shows that methods integrating children data perform the best overall. Specifically, all the methods trained on children's data improve the Baseline Model trained only on adult data. The advantage of DANN is that it keeps a common representation of adult and children's features, and table 11.2 shows that it performs well in both areas, making it a 2-in-1 solution, unlike from scratch and fine tuning models (Table 11.1).

Table 11.2 – Comparison of reconstruction methods between training from scratch, fine tuning and domain adaptation method on **adult test** data with the Fréchet distance.

Model	TEM-C			DANN
Pretraining Data	∅	∅	Adult	Adult
Training Data	Adult	Children	Children	Adult & Children
Fréchet distance	**0.332**	0.378	0.386	<u>0.364</u>

Table 11.2 shows that the DANN performs better than the two other approaches, making it a 2-in-1 solution. Regarding the qualitative analysis (Fig. 11.4), we observe that the trajectory reconstruction using the DANN is quite satisfactory and it seems closest to the ground truth than using the other approaches on those examples. The unique representation shared by the two data sources seems help the model in the trajectory reconstruction, especially on the pen-up part. Additionally, this work is still in progress and the DANN has potential for further improvement.

11.4 Conclusion

This study shows the benefits of retaining knowledge from one domain (adults on tablet) and moving on to a second (children on tablet). Future improvements to this study include several key areas for exploration. First, we plan to study the impact of padding strategies in a batch, which may affect performance by ensuring uniformity across inputs. Additionally, investigating how to create and structure batches optimally could lead to more efficient training and better generalization. We also aim to explore adaptability as a function of age, to assess the gradation of difficulty in adapting to a domain, as measured by the writing expertise gap. . Finally, the role of the lambda parameter will be further studied, particularly focusing on its impact on model performance and devising an adaptive lambda approach that dynamically adjusts the weight of each branch over time for enhanced DANN performance. We also want to investigate domain adaptation on various handwriting surfaces, such as data from tablets to data written on paper.

11.5 Related publications

This initial work on adapting domain to data led to a publication at ICDAR's ADAPDA Workshop [Imbert et al., 2024], and a poster at the SIFED symposium [Imbert et al., 2023].

General conclusion

Summary of contributions and results

The first challenge of this thesis was acquiring data, as none with the current Digipen version was available at the start. Our initial contribution was to establish a data acquisition protocol using the digital pen developed by the Stabilo company, called Digipen and the acquisition app developed again by Stabilo. During this thesis we collected two private datasets and make public parts of them to public [1] two datasets to the research community that allows for testing and refinement of handwriting reconstruction algorithms for future research. These datasets are valuable due to their variability in character words, sentences, equations, and geometric shapes, enhancing their utility beyond merely the quantity of data. We were able to train our models on 11811 data, while the public database on which we also produced results contains 2761 data.

Once the data was collected, the next step was to determine how best to utilize it. The primary goal of the project is to develop a low-cost pencil for use in classrooms to aid in learning to write. However, "low cost" also implies noisy signals to deal with. Additionally, constraints on model training required the sensor and ground truth sequences to be of equal length. We have proposed a complete preprocessing chain to prepare the data that will be provided as input to a neural network. To address the discrepancies in sampling rates between the Digipen and the tablet, as well as synchronization issues, we introduced a Dynamic Time Warping (DTW) based alignment as a preprocessing step to align a sensor signal and a trajectory trace in time.

To create a complete pipeline, we needed to supplement our preprocessing with a neural network. Given our context, the network needed to handle a limited amount of data and be lightweight enough to be embedded easier in the Digipen. Therefore, we opted for a Temporal Convolutional Network (TCN)-based architecture. Our experiments across several datasets demonstrated that the TEM-C architecture trained on touching strokes

1. https://www-shadoc.irisa.fr/irisa-kiht-s-and-kiht-public-datasets/

with context outperformed the Convolutional Neural Network (CNN) architecture proposed in [Wehbi et al., 2022], while the DTW preprocessing improves the performance of both the CNN and TCN compared to a linear alignment [Wehbi et al., 2022].

One of the biggest challenges was modeling untracked and complex pen-up trajectories, where ground truth is lost when the pen is raised too high. To address this, we proposed an approach of a mixture of experts—one specialized for the touching phase and another for the pen-up phase. We refined the pen-up expert model by a fine tuning on data acquired from inclined planes. This fine-tuning allowed the network to better adapt to variations in pen-up height, leading to more robust predictions for the pen-up segments and more qualitative reconstructions.

Another key challenge involved the variability between writers. Writing is a complex motor task requiring coordinated movements of the fingers, wrists, and arms. There are inherent differences between the handwriting gestures of children and adults. Children's handwriting tends to be slower and less precise compared to adults, largely due to developmental factors as they are still refining their motor skills and coordination. To develop a unique model capable of processing data from both adults and children, we began exploring domain adaptation techniques, and especially a Domain-Adversarial Training of Neural Networks (DANN), to learn a common representation for both groups. We propose a DANN network based on the TEM-C proposed previously with an additional head for the domain classifier. The preliminary results shows that DANN outperforms the other two approaches when both child and adult data are considered, making it a 2-in-1 solution.

The objective of our work was to design a model capable of reconstructing the trajectory of the Digipen digital pen, with the flexibility to be used by both adults and children on any type of support.

We focused on adult data collected from a tablet, as it is easier to acquire and allowing us to get a ground truth. Through our various studies, we developed a comprehensive preprocessing pipeline and explored various models, investigating different research avenues such as mixtures of experts and domain adaptation.

We successfully provided a solution to the problem and demonstrated the effectiveness of the proposed approaches in comparison to the current state of the art. These approaches can be adapted to other contexts, such as for children or on paper, with appropriate training data.

Perspective

Our current use of MSE as a loss function offers simplicity in terms of calculus and enables fast training and inference times, allowing the network to be embedded in something as compact as a pencil. However, it does not account for the dynamic nature of handwriting. It would be worthwhile to explore optimized versions of DTW like losses, as these could better capture the temporal dynamics of handwriting while maintaining reasonable computational efficiency. I'm thinking in particular of a Fréchet loss (differentiable Fréchet distance) like what has been done for the DTW (the soft DTW [Cuturi et al., 2018]. Any performance gains achieved through improved loss functions could benefit our mixture-of-experts and domain adaptation methods.

The second area of focus is the treatment of pen-up movements, which represents the most complex aspect of the handwriting process. Exploring alternative approaches to pen-up movements could lead to more effective solutions.

Looking ahead, several promising strategies could enhance the effectiveness and robustness of our handwriting trajectory reconstruction approach. One key improvement would be integrating 3D data into our neural network models. Currently, we rely solely on the projection in the (x, y) plane as the ground truth, but by incorporating a 3-dimensional ground truth (x, y, z). Incorporating full 3D kinematic information could provide a more comprehensive representation of the handwriting process, particularly during pen-up movements. This enhancement has the potential to significantly improve the quality of pen-up movements predictions.

Additionally, reconstructing the entire pen-up trajectory poses significant challenges. Focusing instead on the repositioning vector between strokes may offer a more practical solution. By accurately modeling these repositioning movements without addressing the full complexity of pen-up movements, we can simplify the reconstruction process while still capturing essential aspects of the pen's motion. This approach could streamline the model and reduce computational overhead without compromising the quality of the reconstructed handwriting.

A final area of focus is domain adaptation. While we have started exploring adult-to-child adaptation, the next step would be to refine the two branches of the mixture-of-experts. We are continuing with the adult/children model and have begun exploring an adult-tablet/adult-paper model. Next, we aim to develop a child-paper model and

potentially even a comprehensive model that classifies across four categories: adult vs child, and tablet vs paper.

PERSONAL PUBLICATIONS

Article in an international peer-reviewed journal (+ presentation at ICDAR 2023 – Journal Track)

Swaileh, Wassim et al. (2023), « Online Handwriting Trajectory Reconstruction from Kinematic Sensors using Temporal Convolutional Network », *in*: *International Journal on Document Analysis and Recognition*, DOI: 10.1007/s10032-023-00430-1, URL: https://inria.hal.science/hal-04076399.

Conference Paper

Harbaum, Tanja et al. (2024), « KIHT: Kaligo-based Intelligent Handwriting Teacher », *in*: *DATE 2024*, Valencia, Spain, URL: https://inria.hal.science/hal-04359877.

Imbert, Florent et al. (2024), « Domain adaptation for handwriting trajectory reconstruction from IMU sensors », *in*: *ICDAR 2024 Workshops, ADAPDA*, Athènes, Greece, URL: https://hal.science/hal-04605593.

Serdyuk, Alexey et al. (2023), « Towards the on-device Handwriting Trajectory Reconstruction of the Sensor Enhanced Pen », *in*: *IEEE 9th World Forum on Internet of Things*, Aveiro, Portugal, URL: https://inria.hal.science/hal-04358219.

Communication at the SIFED symposium in France without proceeding

Imbert, Florent et al. (2022), « Toward Deep Neural Network for Pen Trajectory Reconstruction from Kinematic Sensors », *in*: *Symposium International Francophone sur l'Ecrit et le Document (SIFED'2022)*, Rennes, France, URL: https://hal.science/hal-03895960.

Conference poster at the SIFED symposium in France without proceeding

Imbert, Florent et al. (2023), « Domain Adaptation for Pen Trajectory Reconstruction from Kinematic Sensors », *in*: Poster, URL: https://inria.hal.science/hal-04125711.

BIBLIOGRAPHY

Réferences

Alaei, Fahimeh and Alireza Alaei (2022), « Handwriting Analysis: Applications in Person Identification and Forensic », *in*: *Breakthroughs in Digital Biometrics and Forensics*, ed. by Kevin Daimi, Guillermo Francia III, and Luis Hernández Encinas, Cham: Springer International Publishing, pp. 147–165, ISBN: 978-3-031-10706-1, DOI: 10.1007/978-3-031-10706-1_7, URL: https://doi.org/10.1007/978-3-031-10706-1_7.

Alonso, J-M. and Y. Chen (2009), « Receptive field », *in*: *Scholarpedia*.

Askvik, Eva Ose, F. R. (Ruud) van der Weel, and Audrey L. H. van der Meer* (2020), « The Importance of Cursive Handwriting Over Typewriting for Learning in the Classroom: A High-Density EEG Study of 12-Year-Old Children and Young Adults », *in*: *Frontiers in Psychology*.

Atilola, Olufunmilola et al. (2014), « Mechanix: A natural sketch interface tool for teaching truss analysis and free-body diagrams », *in*: *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 28.*2*, 169–192, DOI: 10.1017/S0890060414000079.

Azimi, H., S. Chang, J. Gold, et al. (2022), « Improving Accuracy and Explainability of Online Handwriting Recognition », *in*: *arXiv preprint arXiv:2209.09102*.

Babichev, S. A., J. Ries, and A. I. Lvovsky (2002), *Quantum scissors: teleportation of single-mode optical states by means of a nonlocal single photon*, Preprint at https://arxiv.org/abs/quant-ph/0208066v1.

Bai, S., J Z. Kolter, and V. Koltun (2018), « An empirical evaluation of generic convolutional and recurrent networks for sequence modeling », *in*: *arXiv*.

Barreto, Laura, Paul Taele, and Tracy Hammond (May 2016), « A Stylus-Driven Intelligent Tutoring System for Music Education Instruction », *in*: pp. 141–161, ISBN: 978-3-319-31191-3, DOI: 10.1007/978-3-319-31193-7_10.

Beneke, M., G. Buchalla, and I. Dunietz (1997), « Mixing induced CP asymmetries in inclusive B decays », *in*: *Phys. Lett.* B393, pp. 132–142, arXiv: 0707.3168 [gr-gc].

Bonneton-Botté, Nathalie et al. (2020), « Can tablet apps support the learning of handwriting? An investigation of learning outcomes in kindergarten classroom », *in*: *Computers & Education* 151, p. 103831, ISSN: 0360-1315, DOI: https://doi.org/10.1016/j.compedu.2020.103831, URL: https://www.sciencedirect.com/science/article/pii/S0360131520300336.

Broy, M. (1992), « Software engineering—from auxiliary to key technologies », *in*: *Software Pioneers*, ed. by M. Broy and E. Denert, New York: Springer, pp. 10–13.

Bu, Y., L. Xie, Y. Yin, et al. (2021), « Handwriting-Assistant: Reconstructing Continuous Strokes with Millimeter-level Accuracy via Attachable Inertial Sensors », *in*: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*.

Campbell, S. L. and C. W. Gear (1995), « The index of general nonlinear DAES », *in*: *Numer. Math.* 72.*2*, pp. 173–196.

Chen, Z., D. Yang, J. Liang, et al. (2022), « Complex Handwriting Trajectory Recovery: Evaluation Metrics and Algorithm », *in*: *Proceedings of the Asian Conference on Computer Vision*, pp. 1060–1076.

Choi, Haegyeom et al. (2023), « Hand-Guiding Gesture-Based Telemanipulation with the Gesture Mode Classification and State Estimation Using Wearable IMU Sensors », *in*: *Mathematics* 11.*16*, ISSN: 2227-7390, DOI: 10.3390/math11163514, URL: https://www.mdpi.com/2227-7390/11/16/3514.

Chung, Junyoung et al. (2014), *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*, arXiv: 1412.3555 [cs.NE].

Chung, S. T. and R. L. Morris (1978), *Isolation and characterization of plasmid deoxyribonucleic acid from Streptomyces fradiae*, Paper presented at the 3rd international symposium on the genetics of industrial microorganisms, University of Wisconsin, Madison, 4–9 June 1978.

Cuturi, Marco and Mathieu Blondel (2018), *Soft-DTW: a Differentiable Loss Function for Time-Series*, arXiv: 1703.01541 [stat.ML], URL: https://arxiv.org/abs/1703.01541.

Dai, R., S. Xu, Q. Gu, et al. (2020), « Hybrid spatio-temporal graph convolutional network: Improving traffic prediction with navigation data », *in*: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Derrode, S., H. Li, L. Benyoussef, et al. (2018), « Unsupervised pedestrian trajectory reconstruction from IMU sensors », *in*: *TAIMA 2018: Traitement et Analyse de l'Information Méthodes et Applications*, Hammamet, Tunisia, URL: https://hal.archives-ouvertes.fr/hal-01786223.

Digo, Elisa et al. (2022), « Real-time estimation of upper limbs kinematics with IMUs during typical industrial gestures », *in*: *Procedia Computer Science* 200, 3rd International Conference on Industry 4.0 and Smart Manufacturing, pp. 1041–1047, ISSN: 1877-0509, DOI: https://doi.org/10.1016/j.procs.2022.01.303, URL: https://www.sciencedirect.com/science/article/pii/S187705092200312X.

Diniz, P., D A D. Junior, J. Diniz, et al. (2022), « Time2Vec transformer: a time series approach for gas detection in seismic data », *in*: *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*.

Du, N., H. Dai, R. Trivedi, et al. (2016), « Recurrent marked temporal point processes: Embedding event history to vector », *in*: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.

Ehteram, Mohammad et al. (2024), « Gaussian mutation–orca predation algorithm–deep residual shrinkage network (DRSN)–temporal convolutional network (TCN)–random forest model: an advanced machine learning model for predicting monthly rainfall and filtering irrelevant data », *in*: *Environmental Sciences Europe* 36, DOI: 10.1186/s12302-024-00841-9.

Ganin, Yaroslav et al. (2016), *Domain-Adversarial Training of Neural Networks*, arXiv: 1505.07818 [stat.ML].

Geddes, K. O., S. R. Czapor, and G. Labahn (1992), *Algorithms for Computer Algebra*, Boston: Kluwer.

Ghifary, Muhammad et al. (2016), *Deep Reconstruction-Classification Networks for Unsupervised Domain Adaptation*, arXiv: 1607.03516 [cs.CV], URL: https://arxiv.org/abs/1607.03516.

Gopali, S., F. Abri, S. Siami-Namini, et al. (2021), « A Comparative Study of Detecting Anomalies in Time Series Data Using LSTM and TCN Models », *in*: *arXiv preprint arXiv:2112.09293*.

Graves, A et al. (2006), « Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks », *in*: *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376.

Guirguis, K., C. Schorn, A. Guntoro, et al. (2021), « SELD-TCN: Sound event localization & detection via temporal convolutional networks », *in*: *2020 28th European Signal Processing Conference (EUSIPCO)*, IEEE.

Hamburger, C. (1995), « Quasimonotonicity, regularity and duality for nonlinear systems of partial differential equations », *in*: *Ann. Mat. Pura. Appl.* 169.*2*, pp. 321–354.

Hao, Z. et al. (2014), *Global integrated drought monitoring and prediction system (GIDMaPS) data sets*, figshare https://doi.org/10.6084/m9.figshare.853801.

Har-Peled, S. et al. (2002), « New similarity measures between polylines with applications to morphing and polygon sweeping », *in*: *Discrete & Computational Geometry*.

Hochreiter, Sepp (Apr. 1998), « The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions », *in*: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6, pp. 107–116, DOI: 10.1142/S0218488598000094.

Hochreiter, Sepp and Jürgen Schmidhuber (1997), « Long Short-term Memory », *in*: *Neural computation* 9, pp. 1735–80, DOI: 10.1162/neco.1997.9.8.1735.

Huang, H., D. Yang, G. Dai, et al. (2022), « AGTGAN: Unpaired Image Translation for Photographic Ancient Character Generation », *in*: *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5456–5467.

James, K. H. (2017), « The importance of handwriting experience on the development », *in*: *Current Directions in Psychological Science*.

Kazemi, S M., R. Goel, S. Eghbali, et al. (2019), « Time2vec: Learning a vector representation of time », *in*: *arXiv preprint arXiv:1907.05321*.

Klaß, A., S. M Lorenz, M. Lauer-Schmaltz, et al. (2022), « Uncertainty-aware evaluation of time-series classification for online handwriting recognition with domain shift », *in*: *arXiv preprint arXiv:2206.08640*.

Kreß, F., A. Serdyuk, T. Hotfilter, et al. (2022), « Hardware-aware Workload Distribution for AI-based Online Handwriting Recognition in a Sensor Pen », *in*: *2022 11th Mediterranean Conference on Embedded Computing (MECO)*, IEEE, pp. 1–4.

Krichen, O., S. Corbillé, É. Anquetil, et al. (2022), « Combination of explicit segmentation with Seq2Seq recognition for fine analysis of children handwriting », *in*: *International Journal on Document Analysis and Recognition (IJDAR)*.

Lecun, Y. et al. (1998), « Gradient-based learning applied to document recognition », *in*: *Proceedings of the IEEE* 86.*11*, pp. 2278–2324, DOI: 10.1109/5.726791.

Li, Y., N. Du, and S. Bengio (2017), *Time-Dependent Representation for Neural Event Sequence Prediction*, DOI: 10.48550/ARXIV.1708.00065, URL: https://arxiv.org/abs/1708.00065.

Liao, Chunyuan et al. (Jan. 2008), « Papiercraft: A gesture-based command system for interactive paper », *in*: *ACM Trans. Comput.-Hum. Interact.* 14.

Liu, Y., K. Huang, X. Song, et al. (2020), « MagHacker: Eavesdropping on Stylus Pen Writing via Magnetic Sensing from Commodity Mobile Devices », *in*: *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, Mo-

biSys '20, Toronto, Ontario, Canada: Association for Computing Machinery, ISBN: 9781450379540, DOI: 10.1145/3386901.3389030, URL: https://doi.org/10.1145/3386901.3389030.

Mai, Norbert and Christian Marquardt (1994), « A Computational Procedure For Movement Analysis In Handwriting. », *in*: *Journal Of Neuroscience Methods* 52, pp. 39–45.

Marquardt, Christian, Wolfram Gentz, and Norbert Mai (1996), « On the Role of Vision in Skilled Handwriting », *in*: ed. by M. L. Simner, C. G. Leedham, and J. W. M. Thomassen, vol. Handwriting and Drawing Research, IOS Press, pp. 87–97.

Marvasti-Zadeh, Seyed Mojtaba et al. (2022), « Deep Learning for Visual Tracking: A Comprehensive Survey », *in*: *IEEE Transactions on Intelligent Transportation Systems* 23.*5*, 3943–3968, ISSN: 1558-0016, DOI: 10.1109/tits.2020.3046478, URL: http://dx.doi.org/10.1109/TITS.2020.3046478.

McIntosh, J., A. Marzo, and M. Fraser (2017), « SensIR: Detecting Hand Gestures with a Wearable Bracelet Using Infrared Transmission and Reflection », *in*: *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, Québec City, QC, Canada, ISBN: 9781450349819, DOI: 10.1145/3126594.3126604, URL: https://doi.org/10.1145/3126594.3126604.

Miyagawa, T. et al. (2000), « Handwritten pattern reproduction using pen acceleration and angular velocity », *in*: *IEICE Trans.*

Mohamed Moussa, Elmokhtar, Thibault Lelore, and Harold Mouchère (2023), « Point to segment distance DTW for online handwriting signals matching », *in*: *ICPRAM 2023*, Lisbonne, Portugal: SCITEPRESS - Science and Technology Publications, pp. 850–855, DOI: 10.5220/0011672600003411, URL: https://hal.science/hal-03914726.

Mueller, Pam A. and Daniel M. Oppenheimer (2014), « The Pen Is Mightier Than the Keyboard: Advantages of Longhand Over Laptop Note Taking », *in*: *Psychological Science*.

Mustafid, A., J. Younas, P. Lukowicz, et al. (2022), « IAMonSense: Multi-level Handwriting Classification using Spatio-temporal Information », *in*.

N. Mai and C. Marquardt (2002), « Registrierung und Analyse von Schreibbewegungen: Fragen an den Schreibunterricht. », *in*: *Einblicke in den Schriftspracherwerb*, ed. by Speck-Hamdan A. Huber L. Kegel G., Westermann, pp. 83–99.

Nan, M., M. Trăscău, A M. Florea, et al. (2021), « Comparison between recurrent networks and temporal convolutional networks approaches for skeleton-based action recognition », *in*: *Sensors* 21.*6*, p. 2051.

Neurauter, Rene and Johannes Gerstmayr (Dec. 2022), « A novel motion-reconstruction method for inertial sensors with constraints », *in*: *Multibody System Dynamics* 57, DOI: 10.1007/s11044-022-09863-8.

Nguyen, H T., T. Nakamura, C T. Nguyen, et al. (2021), « Online trajectory recovery from offline handwritten Japanese kanji characters of multiple strokes », *in*: *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, pp. 8320–8327.

Oord, Aaron van den et al. (2016), *WaveNet: A Generative Model for Raw Audio*, arXiv: 1609.03499 [cs.SD].

Ott, F., D. Rügamer, L. Heublein, et al. (2022a), « Benchmarking Online Sequence-to-Sequence and Character-based Handwriting Recognition from IMU-Enhanced Pens », *in*: *arXiv preprint arXiv:2202.07036*.

— (2022b), « Domain adaptation for time-series classification to mitigate covariate shift », *in*: *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5934–5943.

— (2022c), « Joint Classification and Trajectory Regression of Online Handwriting using a Multi-Task Learning Approach », *in*: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 266–276.

Ott, F., M. Wehbi, T. Hamann, et al. (2020), « The onhw dataset: Online handwriting recognition from imu-enhanced ballpoint pens with machine learning », *in*: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.*3*, pp. 1–20.

Oviatt, Sharon et al. (2012), « The Impact of Interface Affordances on Human Ideation, Problem Solving, and Inferential Reasoning », *in*: *ACM Transactions on Computer-Human Interaction (TOCHI)* 19.

Pablo, Vásconez Juan et al. (2022), « Hand Gesture Recognition Using EMG-IMU Signals and Deep Q-Networks », *in*: *Sensors* 22.*24*, ISSN: 1424-8220, DOI: 10.3390/s22249613, URL: https://www.mdpi.com/1424-8220/22/24/9613.

Pan, T-Y., C-H. Kuo, and M-C. Hu (2016), « A noise reduction method for IMU and its application on handwriting trajectory reconstruction », *in*: *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6, DOI: 10.1109/ICMEW.2016.7574685.

Pan, T-Y., C-H. Kuo, H-T. Liu, et al. (2018), « Handwriting trajectory reconstruction using low-cost imu », *in*: *IEEE Transactions on Emerging Topics in Computational Intelligence* 3.*3*, pp. 261–270.

Pathour, Teja et al. (Apr. 2023), « Hemoglobin Microbubbles and the Prediction of Different Oxygen Levels Using RF Data and Deep Learning », *in*: vol. 12470, p. 15, DOI: 10.1117/12.2655121.

Peñaloza, V. (2020), « Time2Vec Embedding on a Seq2Seq Bi-directional LSTM Network for Pedestrian Trajectory Prediction. », *in*: *Res. Comput. Sci.*

Porterfield, M K. (Feb. 2020), « Accelerometer Drift Study », *in*: DOI: 10.2172/1601376, URL: https://www.osti.gov/biblio/1601376.

Roodschild, Matías, Jorge Gotay Sardiñas, and Adrián Will (2020), « A new approach for the vanishing gradient problem on sigmoid activation », *in*: *Progress in Artificial Intelligence* 9.*4*, pp. 351–360, ISSN: 2192-6360, DOI: 10.1007/s13748-020-00218-y.

Sakoe, H. and S. Chiba (1978), « Dynamic programming algorithm optimization for spoken word recognition », *in*: *IEEE transactions on acoustics, speech, and signal processing*.

Salvador, S. and P. Chan (2007), « Toward accurate dynamic time warping in linear time and space », *in*: *Intelligent Data Analysis*.

Schreibmotorik Institut in Kooperation mit dem Verband Bildung und Erziehung (VBE Bund) und den 16 VBE Landesverbänden (2019), *STEP 2019: Studie über die Entwicklung, Probleme und Interventionen zum Thema Handschreiben*, Technical Report, Schreibmotorik Institut, URL: https://www.schreibmotorik-institut.com/images/STEP_Studie_2019.pdf.

Seymour, R. S., ed. (1981), *Conductive Polymers*, New York: Plenum.

Shen, Jian et al. (2018), *Wasserstein Distance Guided Representation Learning for Domain Adaptation*, arXiv: 1707.01217 [stat.ML].

Simonnet, D., N. Girard, E. Anquetil, et al. (2019), « Evaluation of children cursive handwritten words for e-education », *in*: *Pattern Recognition Letters*.

Slifka, M. K. and J. L. Whitton (2000), « Clinical implications of dysregulated cytokine production », *in*: *J. Mol. Med.* 78, pp. 74–80, DOI: 10.1007/s001090000086.

Smith, S. E. (1976), « Neuromuscular blocking drugs in man », *in*: *Neuromuscular junction. Handbook of experimental pharmacology*, ed. by E. Zaimis, vol. 42, Heidelberg: Springer, pp. 593–660.

Taranta, Eugene et al. (July 2016), « A Dynamic Pen-Based Interface for Writing and Editing Complex Mathematical Expressions With Math Boxes », *in*: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6, p. 13, DOI: 10.1145/2946795.

Vaswani, Ashish et al. (2017), *Attention Is All You Need*, arXiv: 1706.03762 [cs.CL].

Vayer, T., L. Chapel, N. Courty, et al. (2020), « Time series alignment with global invariances », *in*: *arXiv preprint arXiv:2002.03848*.

Wang, J.-S, Y.-L. Hsu, and J.-N. Liu (2009), « An inertial-measurement-unit-based pen with a trajectory reconstruction algorithm and its applications. », *in*: *IEEE Transactions on Industrial Electronics*.

Wang, Yifeng and Yi Zhao (2024), « Handwriting Recognition Under Natural Writing Habits Based on a Low-Cost Inertial Sensor », *in*: *IEEE Sensors Journal* 24.*1*, pp. 995–1005, DOI: 10.1109/JSEN.2023.3331011.

Wegmeth, L., A. Hoelzemann, and K. Van Laerhoven (2021), « Detecting Handwritten Mathematical Terms with Sensor Based Data », *in*: *arXiv preprint arXiv:2109.05594*.

Wehbi, M., T. Hamann, J. Barth, et al. (2021), « Towards an IMU-based Pen Online Handwriting Recognizer », *in*: *International Conference on Document Analysis and Recognition*, Springer, pp. 289–303.

Wehbi, M., D. Luge, T. Hamann, et al. (2022), « Surface-Free Multi-Stroke Trajectory Reconstruction and Word Recognition Using an IMU-Enhanced Digital Pen », *in*: *Sensors*.

Wilhelm, M., D. Krakowczyk, and S. Albayrak (2020), « PeriSense: Ring-Based Multi-Finger Gesture Interaction Utilizing Capacitive Proximity Sensing », *in*: *Sensors*, DOI: 10.3390/s20143990.

Wöllmer, M., M. Al-Hames, F. Eyben, et al. (2009), « A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams », *in*: *Neurocomputing*.

Yan, J., L. Mu, L. Wang, et al. (2020), « Temporal convolutional networks for the advance prediction of ENSO », *in*: *Scientific reports*.

Zaheer, Manzil et al. (2021), *Big Bird: Transformers for Longer Sequences*, arXiv: 2007.14062 [cs.LG].

Zhu, Yuanping et al. (2024), « Handwriting Trajectory Reconstruction of Chinese Characters Based on the Font Imitate Network », *in*.

COLLEGE DOCTORAL BRETAGNE
MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

Titre : Conception d'une architecture de réseaux de neurones profonds dédiée à la synthèse d'écriture manuscrite à partir de capteurs cinématiques d'un stylo numérique.

Mot clés : Écriture manuscrite en ligne, reconstruction de trajectoire, stylo numérique, Inertial Measurement Units, réseau de neurones profonds, adaptation de domaine.

Résumé : Cette thèse vise à reconstruire la trace de l'écriture manuscrite en ligne à partir d'un stylo numérique Stabilo équipé de capteurs cinématiques. Nous proposons un pipeline de traitement associant les signaux des capteurs à la trajectoire d'écriture, utilisant le Dynamic Time Warping pour l'alignement et une architecture inspirée des Temporal Convolutional Networks. En outre, nous présentons une approche de mélange d'experts (MOE) pour améliorer la compréhension de chaque aspect de l'écriture manuscrite, comprenant un modèle d'expert pour les touchés de crayon et un modèle d'expert pour les trajectoires plume haute. La variation des signaux capturés entre les adultes et les enfants, due aux différences de vitesse et de confiance dans les gestes d'écriture manuscrite, constitue un défi important. Nous y remédions par une approche d'adaptation au domaine. Par ailleurs, nous fournissons un nouvel ensemble de données de référence publiques pour soutenir les recherches et les comparaisons futures dans le domaine de la reconstruction de l'écriture manuscrite.

Title: Design of a deep neural network architecture dedicated to handwriting gesture synthesis from kinematic sensors coming from a digital pen.

Keywords: Online Handwriting, Trajectory Reconstruction, Digital Pen, Inertial Measurement Units, Deep Neural Network, domain adaptation

Abstract: This thesis focuses on a digital pen equipped with kinematic sensors, and its aim is to reconstruct the in-line trace of handwriting. We introduce a new processing pipeline that associates pen sensor signals with the corresponding writing trajectory. Based on Dynamic Time Warping to align the signals and an architecture inspired by Temporal Convolutional Networks Additionally, we present a Mixture-Of-Experts (MOE) approach to enhance the focus and understanding of each aspect of handwriting, comprising a touching expert model for pencil touches and a pen-up expert model for pen trajectories. A significant challenge is the variation in captured signals between adults and children, due to differences in speed and confidence in handwriting gestures. We address this through a domain adaptation approach. Furthermore, we introduce a new public benchmark dataset to support future research and comparisons in the field of handwriting reconstruction.