MOMINA MOETESUM
01-284142-001

Deformation Estimation and Classification of Graphomotor Impressions - An Application to Neuropsychological Assessments

A Dissertation submitted to the Department of Computer Science, Bahria University, Islamabad in the partial fulfillment for the requirements of a Doctoral degree in Computer Science.

Supervisor: Dr. Imran Siddiqi

Department of Computer Science
Bahria University, Islamabad

September 2020

Abstract

Graphomotor skills of an individual can provide useful insight regarding his/her mental health and emotional state. Assessing these skills enables neuropsychologists to target areas of dysfunction and to design appropriate plans for rehabilitation. To assess graphomotor skills, clinical practitioners employ a variety of pen-and-paper based graphomotor tasks involving handwriting and drawings. Performance of an individual in these tasks is measured by using extensive scoring criteria that determines the presence/absence of various motor, perceptual, and cognitive deformations, by estimating deviations from expected stimulus. Nevertheless, manual scoring by human experts is time consuming and prone to inter-scorer bias. Computerized analysis of responses produced by at-risk subjects has high potential to address the aforementioned limitations of manual assessment. Furthermore, computerized analysis can also facilitate test standardization and validation, treatment efficacy assessment and disease progression monitoring.

While a number of techniques are presented in the literature, there is a constant need to explore effective methods to translate domain knowledge into computational feature space. In this research, we propose a novel deformation modeling and estimating method, that can model a variety of visual-motor and visual-perceptual deformations from both online and offline samples of patient responses. The proposed scheme suggests feature extraction by means of pre-trained Convolutional Neural Networks (CNNs) for rich visual representation of samples of a particular deformation. By employing pre-trained ConvNets, we overcome the limitations of data scarcity and feature insufficiency that are common characteristics in this domain. To further enhance representation deformation-specific augmentation is employed. These enhanced visual features are then used to train classical machine learning classifiers to predict the presence or absence of particular deformation(s) in the test sample. The proposed method can be employed in a wide variety of scenarios to analyze a neuropsychological response.

The performance of our proposed deformation estimation and classification method is evaluated in two empirical settings, that are popularly targeted by the relevant research community as well i.e. (a) Early detection of a neurodegenerative disease and (b) Scoring of a neuropsychological test with an extensive scoring standard. Our first scenario involves the identification of visual-motor deformations like tremor and micrographia from the graphomotor responses of elderly for the prediction of Parkinson's Disease (PD). We employ a popular benchmark dataset 'Parkinson's Disease Handwriting (PaHaW)' database, that comprises multiple graphomotor tasks performed by subjects suffering from PD and healthy controls (HC). To highlight fine imperfections caused due to associated motor dysfunctions (like tremor), we propose two non-linear transformations of the raw images using median and edge enhancing filters. For feature extraction, we employ the convolutional base of a pre-trained ConvNet and combine the extracted features of different representations to provide further enhancement. The combined feature vectors from each task are then employed to train a task-specific Support Vector Machine (SVM) classifier that predicts the response as belonging to either of the two classes (PD/HC). From our evaluations, it is observed that each task has a different impact on the classification accuracy. Due to this reason, decisions of

all tasks performed by a subject are combined by applying majority voting. The ensemble approach not only improved the overall classification results (83%) but also mitigated the negative impact of a task on the predictive potential of the extracted features.

The second study targets the identification of eleven visual-perceptual deformations outlined in the Lacks' scoring standard for the assessment of a Bender Gestalt Test (BGT) response. Perceptual deformations are challenging to model due to the insufficiency of features and reliance on extensive heuristics. We apply our proposed deformation modeling and classification method to identify Lacks' eleven indicators of perceptual dysfunction from samples of children with learning difficulties. Due to lack of relevant datasets, a customized dataset is employed for the evaluation purposes. Unlike conventional sketch recognition, where intra-shape class variations are diminished and inter-shape class variations are enhanced, our proposed methodology enhances deformation-specific intra-shape class variations and generalizes inter-shape class similarities. This has not been attempted previously and enables the identification of same deformation across multiple shapes and different deformations with in same shape class. Once again, deformation-specific transformations are employed to ensure representation of the missing classes as well as to enrich features. Several combinations of pre-trained ConvNets and binary classifiers are assessed to determine the best combination. Results of our experiments show that best classification rates (i.e. mean accuracies ranging from 79.1% to 97.6%) are achieved across all deformations when features extracted from ResNet101 are used to train Linear Discriminant Analysis (LDA) classifier. Decisions from different deformation-specific classifiers are combined to quantify errors as required by the scoring standard. From the results of our experiments, we found that the nature of the deformation contributes the most in the performance of the classifier. This finding is coherent with that observed during manual scoring, as there exists greater inter-scorer difference for some deformations as compared to others. Nonetheless, the outcomes of both scenarios highlight the effectiveness of our proposed methodology in terms of reliability and robustness and support its potential for providing a solid basis for relevant end-to-end systems that can easily be integrated into the mainstream clinical settings to facilitate practitioners in diagnostic decision making.

Acknowledgments

In the name of Allah Almighty, who is the most merciful and the most beneficent. He is the One who gave me the courage and determination to carry on when all seemed impossible.

It is my great fortune to have pursued my research under the guidance of my adviser, Prof. Dr. Imran Ahmed Siddiqi, who introduced me to the problem area, and guided me at every step of the way with his knowledge and experience. His prompt and detailed feedback greatly aided me throughout my research and inspired me to explore deeper. Our weekly meetings always kept me focused and helped me meet deadlines.

I would like to thank Dr. Uzma Masroor (former Head of Department of Professional Psychology, Bahria University, Islamabad), for her support. Despite her exhausting schedule, she has always shown keen interest in my research. Due to my Computer Science background, I was unfamiliar with several neuropsychological concepts that were essential for my research. Dr. Uzma Masroor provided vital guidance in this regard and answered all my queries with patience. Dr. Uzma and her team of clinical psychologists helped me with sample collection and ground truth preparation.

My special gratitude to Prof. Nicole Vincent and Dr. Florence Cloppet, who were my research collaborators at LIPADE, Paris Descartes University, France, during the Hubert Curien Franco-Pak PERIDOT Research project. Their valuable feedback during the two year research project helped shape the research objectives and direction for my PhD work.

Research is always a team work and therefore, I would like to extend my appreciation to my team members Haris Bin Nazar and Osama Zeeshan, who provided technical assistance during experimentation and result compilation. I would like to thank Dr. Shoaib Ehsan and Klaus McDonald-Maier (University of Essex, UK) for introducing me to the world of deep learning. I would also like to appreciate the assistance of Dr. Khurram Khurshid (Institute of Space Technology, Pakistan) for providing access to technical resources at his university and Dr. Chawki Djeddi (Larbi Tebessi University, Algeria) for his feedback regarding the statistical analysis employed in my research.

During any research, there comes a time when one feels discouraged by dead ends. It is the time when the support of family and friends boosts hope and provides much needed encouragement. I am privileged to have had the support and patience of my dearest husband Moetesum Khurshid and my beloved children Haya and Taha during my demanding routine and occasional frustrations. I am grateful to my dearest friend Mrs. Anaum Saqib and my spiritual guide Mrs. Saima Jawad for their positive pep talks during my desperate days.

Finally I would like to extend my gratitude to the Department of Computer Science and Bahria University, Islamabad, for facilitating my research by providing access to technical resources and online archives. A special thanks to my instructors during the course work and to Mr. Fazal Wahab, Prof. Dr. Faisal Bashir and Dr. Muhammad Muzammal (Head of Department) for the unobstructed execution of official requirements from the commencement to the completion of my degree.

MOMINA MOETESUM
Department of Computer Science,
Bahria University Islamabad,
Pakistan

September 2020

'Unlike success and failure, contribution has no other side.
It is not arrived at by comparison.'

Benjamin Zander

Acronyms and Abbreviations

AD	Alzheimer's Disease
ADHD	Attention deficit hyperactivity disorder
AI	Artificial Intelligence
ANN	Artificial Neural Networks
AUC	Area Under the Curve
BDT	Bangor Dyslexia Test
BGT	Bender Gestalt Test
BiSP	Biometric Smart Pen
CAD	Computer-Aided Diagnosis
CD	Critical Difference
CDT	Clock Draw Test
CNNs	Convolutional Neural Networks
CT	Computer Tomography
DAP	Draw-A-Person
DT	Decision Trees
EER	Equal Error Rate
HTP	House Tree Person
KNN	K-Nearest Neighbour
LDA	Linear Discriminant Analysis
fMRI	Functional Magnetic Resonance Imaging
MCI	Mild Cognitive Impairments
MDD	Major Depressive Disorder
MLP	Multi-Layer Perceptron
MMSE	Mini Mental Status Examination
MRI	Magnetic Resonance Imaging
MRT	Mean Relative Tremor
NB	Naiive Bayes
OPF	Optimum-Path Forest
PaHaW	Parkinson's Disease Handwriting Database
PD	Parkinson's Disease
PET	Positron Emission Tomography
RBF	Radial Basis Function
R-CNNs	Regions with Convolutional Neural Networks
ReLU	Rectified Linear Unit
RF	Random Forest
R-FCNs	Region Based Fully Convolutional Networks
RNN	Recurrent Neural Networks
ROCF	Rey-Osterrieth Complex Figure

SMOTE	Synthetic Minority Over-Sampling Technique
SPECT	Single Photon Emission Computed Tomography
SSD	Single Shot Multibox Detector
SVM	Support Vector Machine
UPDRS	Unified Parkinson's Disease Rating Scale
VSN	Visuo-Spatial Neglect
WRAT-4	Wide Range Achievement Test-4

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Artificial Intelligence (AI) has gradually paved its way into the healthcare industry over the past few decades due to its potential advantages [1]. AI-based solutions can provide efficient retrieval of relevant information from large volumes of electronic healthcare data [2]. It can facilitate proactive public health surveillance like epidemic prediction by formulating real-time inferences from social networks [3]. Currently, researchers are focusing on the development of reliable intelligent diagnostic systems to assist doctors in the interpretation of various test samples [4]. The primary objective of such a system is to provide optimization and standardization of conventional medical practices by reducing time, cost, and diagnostic errors [5]. This can benefit domains like neuropsychology that are facing an increasing stress on the traditional one-to-one clinical contact time-based services [6]. Consequently, neuropsychologists are becoming less skeptical towards adapting emerging technologies in an attempt to provide cost-effective and time-efficient services to masses [7].

Neuropsychology is an established discipline that investigates the brain–behavior relationships [8]. A neuropsychologist attempts to determine the presence of brain dysfunctions by examining abnormal behavioral patterns exhibited by a potential at-risk individual [9]. Dysfunctional processing of brain can either result from an underlying neurological disorder (e.g. traumatic brain injuries, dementia, Alzheimer's or Parkinson's disease, etc.) or a psychiatric imbalance (e.g. Schizophrenia, Attention deficit hyperactivity disorder (ADHD) or learning disability, etc.). In both cases, early detection and regular progression monitoring during treatment has a significant impact on the rehabilitation process. Invasive methods like 'Single Photon Emission Computed Tomography (SPECT)' [10] or 'Functional Magnetic Resonance Imaging (fMRI)' [11] are mostly costly and may cause discomfort to the patient. Due to this reason, clinical neuropsychologists employ several non-invasive test batteries to screen indications of a particular disorder before recommending for further assessment.

Neuropsychological test batteries [12] comprise a number of performance-based tests that require individuals to perform various verbal and non-verbal tasks. These tests are non-intrusive,

easy to administer and are designed to assess various cognitive, perceptual and motor skills of an individual. A comprehensive neuropsychological assessment like the *'Halstead-Reitan Neuropsychological Test Battery (HRB)'* [13] and *'Luria-Nebraska Neuropsychological Battery (LNNB)'* [14], can prove effective in explaining the consequential changes in an individual's behavior, emotions and executive functioning due to an underlying dysfunction. By employing specific methodological procedures sensitive to specific functional changes, a neuropsychologist can correlate a particular cognitive or behavioral impairment and suggest proper rehabilitation needs. Similar procedures can also be used to measure treatment efficacy [12]. In this regard, a neuropsychologist provides consultancy services to both neurologists and psychiatrists.

1.1.1 Graphomotor Based Neuropsychological Assessments

A popular category of neuropsychological assessments comprises the *'Pen-and-Paper'* tests. Most of these tests include certain graphomotor-based tasks involving drawing or handwriting. Drawing is a process of producing a graphic plan, while handwriting is a procedure of forming letters and numeric symbols. Both are widely employed means for recording thoughts or conveying experiences. Nevertheless, according to experts like occupational therapists, developmental psychologists, and neuropsychologists, drawing and handwriting are much more than mere tools of self expression [12]. Instead, these are complex multi-componential activities that require necessary graphomotor skills like visual-perceptual maturity, orthographic coding, motor planning and execution, kinesthetic feedback, and visual-motor coordination [15]. Studies [16, 17], support that dysfunction of any of these skills due to associated brain disorders affects the drawing and handwriting performance of an individual. Based on this assumption, graphomotor impressions (i.e. drawing and handwriting) have been employed as psychometric tools for the detection of a variety of neuropsychological and neurological disorders such as apraxia, visuo-spatial neglect (VSN), dysgraphia, and dementia etc. [18]. The impact of a writer's emotional state on his/her handwriting has also been established in some studies [19]. Some of the popularly employed graphomotor tests include the following:

- Rey-Osterrieth Complex Figure (ROCF) Drawing Test (Figure 1.1-a), is used for the assessment of visuo-spatial abilities, executive planning, working memory, effects of brain injury, dementia, and to study the degree of cognitive development in children [20].
- Clock Draw Test (CDT) (Figure 1.1-b), is popularly employed to assess stages of dementia in Alzheimer's disease (AD) [21].
- Bender Gestalt Test (BGT) (Figure 1.1-c), is used for assessing the visuo-perceptual maturity in children and the secondary effects of brain lesions due to trauma and injury in adults [22].
- Archimedean Spiral (Figure 1.1-d), is a commonly used screening test for early detection of Parkinson's disease (PD) in potential at-risk individuals [23].
- Draw-A-Person (DAP) (Figure 1.1-e), is a popular projective test for the assessment of the intellectual and emotional health of a child [24].

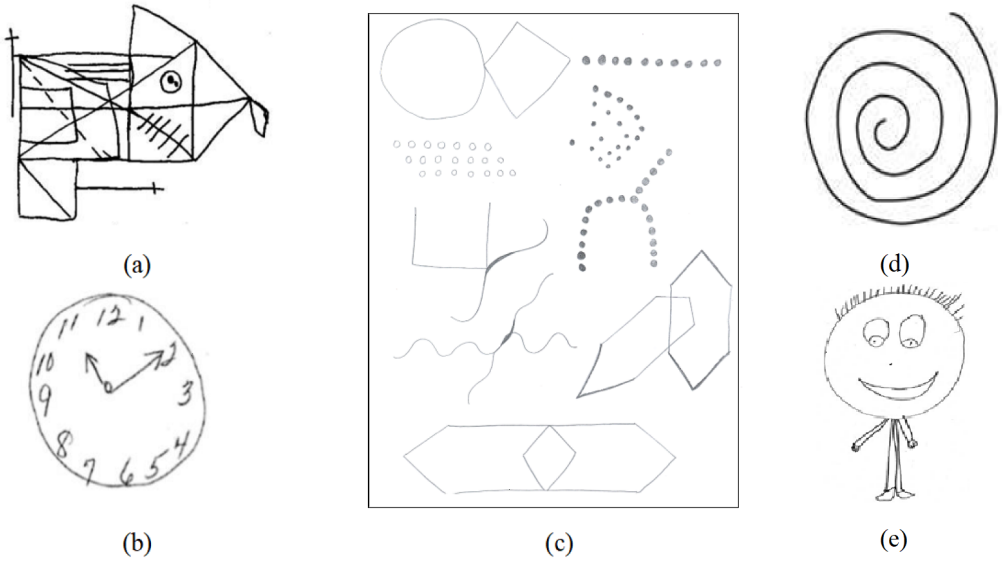


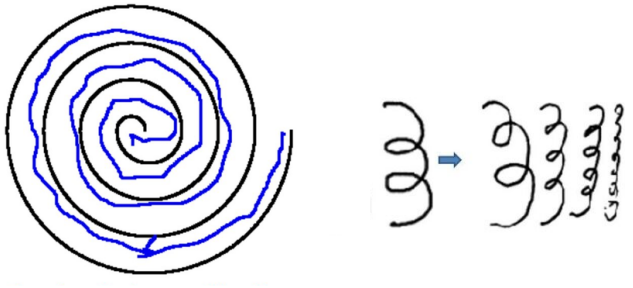
Figure 1.1: Samples of graphomotor-based neuropsychological tests: (a) ROCF drawing sample [25], (b) CDT drawing sample [26], (c) BGT drawing sample [27], (d) Spiral drawing sample [28], (e) DAP drawing sample [29]

1.1.2 Conventional Test Conduction and Scoring Protocol

Conventional test conduction protocol requires the subjects to draw or write on a page using a pen or a pencil as a medium. The test may include copying or recalling a visual stimuli (reproducing method), completing a partially drawn/written task (completion method) , or projecting a concept in words or graphics (projective method). These tests can be conducted individually or in a group setting. Once the subjects produce a response, these are then visually examined by the domain expert with the objective to identify indicators of specific brain dysfunctions, using standard scoring manuals [30–32]. The results obtained from these assessments are further correlated with other clinical findings to diagnose associated disorders and then suggest adequate rehabilitation.

The scoring manuals designed to interpret these tests can be quantitative [33] or qualitative [32] in nature. Quantitative methods assess the graphomotor response for the presence/absence of particular deformation(s), while qualitative ones attempt to describe the degree of deformation(s). Quantitative methods are mostly psychometric in nature while qualitative are inductive [34]. Due to their strong psychometric characteristics, quantitative methods are usually preferred for the initial screening purposes. The outcome of a quantitative method is usually a numeric score that is considered to have some intrinsic meaning upon which assumptions are made as to the existence of a brain lesion, its location, and associated deficits [12].

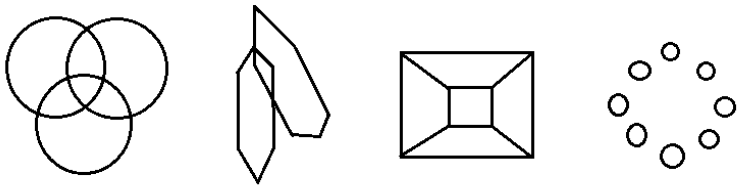
Deformations are determined by measuring the extent of deviation(s) from the standard template(s). High degree of deformation indicates various motor, perceptual and cognitive disorders. For instance, to assess signs of motor deficits like ‘Micrographia’ (tightness) [35] and ‘Tremor’ (irregularity) [36], potential PD patients are instructed to copy/trace spirals and repetitive loops, as shown in Figure 1.2-a, b, respectively. Similarly, visual-perceptual development of an individual is



(a) (b)

Figure 1.2: Irregularity and tightness of loops and spirals as biomarker for motor dysfunctions

measured by instructing him/her to draw geometrically inspired shapes that include components of linearity, circularity, curvilinearity and angularity, as shown in Figure 1.3. Common defor-



(a) (b) (c) (d)

Figure 1.3: Examples of shapes used in various Visual-Perceptual assessments; (a) 2-D rings (b) Wertheimer's hexagons (c) Tapered box (d) 8-dot circle

.

mations considered by clinical practitioners while assessing drawn responses to these stimuli for visual-perceptual disorders include rotation, fragmentation, cohesion and perseveration etc. For instance, drawings of individuals suffering from frontal lobe injury are prone to *perseveration* which means that such individuals may repeat a particular pattern over and over again. Closure difficulty or inability to meet the joining parts of a shape is linked with an indication of constructional apraxia [37] and VSN [38]. Similarly, rotation of a complete figure through 80° to 180°, has been associated with signs of focal brain lesions and dementia in elderly [39].

1.1.3 Need for Computerized Analysis of Graphomotor-Based Tasks

Despite the significance of graphomotor-based neuropsychological assessments, there has been a gradual decline in their use over the last decade [40]. Several factors contribute to this scenario. Nonetheless, two main contributing factors are listed below:

- **Extensive Scoring:** Scoring of graphomotor-based tasks is extensive and time-consuming due to lengthy scoring standards. An average scoring manual consists of approximately 30 to 50 scoring points against which the graphomotor response is to be evaluated. For instance, the *Goodenough-Harris scoring system* [30] assesses a DAP [29] response against 51 point criteria. In case a test consists of multiple templates, like in BGT [22], then the scoring

complexity further increases. According to *Lacks' scoring system* [31], each of the nine BGT responses are assessed for multiple deformations. The same deformation can also be assessed differently for different templates. In a typical clinical setting, assessment of a patient's response may take hours [41].

- **Inter-Scorer Variability and Test-Retest Reliability:** Interpretation of these tests requires extensive training and clinical experience. Nevertheless, at times a scorer's experience and bias, and a patient's profile can introduce undesirable factors like inter-scorer variability and test-retest reliability [42], leading to the lack of precision, accuracy and standardization.

Increased availability of commercial scanners and high resolution cameras has enabled digitization of pen-and-paper samples. This has facilitated the AI community to process such digitized samples for the purposes of automated analysis. An automated analysis of a complete test or portions of a test battery can address the above mentioned problems to much extent. Several commercial and research associated benefits can be attained by realization of such a system. Some of these are outlined below:

- **Time-Efficiency:** A computerized system that can score deformations in the graphomotor samples of potential patients can provide efficiency by reducing load of psychologists. Non-experts can be trained to use such system, who can then compile results for the clinician to interpret. In this way, psychologists can focus more on the patient and treatment instead of scoring and report formulation.

- **Score Reliability and Standardization:** Based on a basic premise that a screening device should not have an adverse impact on the overall assessment of the patient due to extrinsic factors like human biasness, an automated system can provide standardization. It can also help to reduce issues like inter-scorer variability.

- **Distance Treatment:** Such automated systems can also facilitate distance treatment where psychologists can recommend screening tests of a potential patient prior to conducting an actual face-to-face session. During the current pandemic situation, we have seen a tremendous need for such systems to ensure safety of both the doctors and the general public.

- **Test Validation and Improvisation:** Automation of such tests can not only preserve their usability but also encourage the practitioners to improve them. It can provide the collection of normative data that can later be compared with existing normative databases.

1.1.4 Existing Systems and Open Issues

Motivated by the advantages of computerized analysis of graphomotor-based tasks, several attempts have been made in this regard by researchers from the AI and more specifically pattern recognition community. Categorization of the existing systems can be done based on several criteria. For instance, these can be grouped together based on their objectives (i.e. *'feature validation'* or *'disease diagnosis'*), or their mode of data acquisition (i.e. *'offline'* or *'online'*). However, we have

opted for a categorization criterion based on the mode of analysis i.e. whether the system analyzes the visual feedback of the completed graphomotor response or assesses the procedure involved during its creation. Thus, terming them as *'Visual analysis based techniques'* [27, 43–52] and *'Procedural analysis based techniques'* [53–62, 62–81].

The visual analysis based techniques encompass studies that attempt to analyze completed responses of neuropsychological graphomotor-based tasks like the Necker's cube [43], ROCF [44], BGT [27], CDT [47] and Archimedean Spiral [49]. These systems rely on static geometric and spatial features (like size, angle, orientation or pixel-wise distance) extracted from digitized offline samples (and in some cases online samples [43, 52], as well) of drawings or handwriting responses. These features are then compared with those of the expected templates either by means of template-matching [43] or extensive domain-specific heuristics [52]. Due to the insufficiency of static features, researchers are now exploring potential alternatives like dynamic analysis of handwriting and drawing by means of specialized electronic devices (e.g. digitizer tablets and smart pens etc.). This has led to a shift in paradigm from visual analysis based techniques to procedural analysis based techniques. Novel biomarkers like kinematic, pressure and temporal features are being evaluated for discriminating between samples of patients and healthy subjects [78, 82, 83]. Despite its apparent advantage over static visual analysis, lack of clinical standards for dynamic analysis can lead to ambiguity. For instance, a number of studies [50, 84] have already suggested modifications of traditional templates to enhance feature representation. Both approaches have their inherent strengths and weaknesses, that has led to the identification of several open issues that need to be considered while designing an effective system. Some of these are outlined below to provide a basic perspective:

- **Mode of Sample Acquisition:** Techniques relying on the analysis of traditional paper-based samples require digitization by means of a scanner or a digital camera. The digitized images are then preprocessed to make them suitable for an automated analysis. Similar to any handwritten document analysis system, preliminary tasks of localization, segmentation and recognition must be performed before analysis. On the other hand techniques advocating online sample acquisition require the modification of conventional sample acquisition modality. Due to this reason, such systems may face hesitancy from the target users i.e. clinical practitioners and patients.

- **Data Paucity:** Like most health related problems, the lack of sufficient training data is a major limiting factor in the design of a computerized analysis system for any neuropsychological test. Being a highly domain-specific problem, sample acquisition and ground truth labeling has to be performed by a domain expert. Inability to do so may question the validity of the system.

- **Component-Level versus Gestalt-Level Analysis:** While handwriting is based on stroke formations, drawing-based tests are inspired by the 'Gestalt theory' [85]. The gestalt school of thought suggests that *'the whole is greater than the sum of its parts'*. Based on this assumption, the shapes of the neuropsychological drawing tests like CDT and BGT are analyzed not only

based on independent shape-based deformations but also based on the spatial organization of primitive shape components with respect to the neighbouring components. This is one of the fundamental design concerns involved.

- **Domain Knowledge Representation:** Each test has a distinct and in most cases, an extensive scoring criteria that requires effective translation of clinical manifestations (domain knowledge) into computable features to make an inference similar to the clinical practitioner. Two common approaches are adopted in the literature, these include *hand-crafted heuristics* and *supervised machine learning*. Heuristics are mostly exhaustive and rigid, and may not prove sufficient in most real life applications due to highly unconstrained nature of the responses. On the contrary machine learning-based approaches can generalize a wide variety of situations but require a large amount of training data that is already an issue in this domain. Keeping this in view, deeper exploration of both approaches is required before application for the problem under consideration.

1.2 Problem Statement

This research targets the problem of deformation estimation and classification for computerized analysis of graphomotor-based neuropsychological tasks. More specifically, two types of deformations that are commonly assessed by practitioners while studying cognitive dysfunctions are considered in our study. These include visual-motor and visual-perceptual deformations. Visual-motor deformations are early indicators of several neurodegenerative diseases that affect the nerve cells controlling the motor coordination and executive planning. Visual-perceptual disorders, on the other hand, result from trauma, injury or underdeveloped cognitive skills. The study on visual-motor deformations is carried out using samples of Parkinson's disease patients while for identification of visual-perceptual deformations, samples of BGT responses drawn by children are employed.

The current state-of-the-art for identification of visual-motor deformations is primarily dominated by procedural analysis based techniques. Such methods rely on procedural attributes (like kinematics, pressure, and time) which not only require specialized hardware but also result in modification of the conventional test conduction and scoring protocols. This in turn leads to a hesitancy of domain experts in accepting computerized solutions in their daily practices. There is a need for rich visual features that can sufficiently represent the target deformations and assist in early disease detection without modifying the sample acquisition modality. Unlike visual-motor dysfunctions, visual-perceptual skills cannot be modeled by kinematic or temporal features and require spatial and shape-based features for representation. Furthermore, their scoring typically relies on extensive heuristics that may prove insufficient to represent all possible deviations. Consequently, investigation of rich visual representations that can generalize a wide variety of deviations without the reliance on extensive heuristics remains a challenging research problem.

1.3 Research Objectives

Our research is primarily aimed at designing an algorithmic approach that can effectively translate relevant domain knowledge into computational features. More specifically, we target two types of deformations which are commonly employed in neuropsychological clinical practices. These include *visual-motor* and *visual-perceptual* deformations. Keeping in view the problem at hand, we have outlined some major objectives of this research as listed in the following.

- To develop computational methods to model and estimate graphomotor deformations assessed by clinical practitioners while analyzing neuropsychological test responses
- To explore effective and efficient computational representation of neuropsychological domain knowledge in the context of visual-motor and visual-perceptual deformations
- To investigate the potential of visual information in graphomotor impression in discriminating samples of diseased subjects and healthy controls
- To study the impact of various conventional and non-conventional graphomotor tasks on the performance of the proposed visual-motor deformation estimation and classification methodology
- To identify common visual-perceptual deformations across multiple shapes and multiple deformations co-existing in a single shape without extensive heuristics

1.4 Significance of Research

The findings of this research are expected to contribute towards the application of artificial intelligence techniques in the health care industry in general and neuropsychological assessments in particular. Research and development in health care systems and access to easy and affordable medical facilities for everyone is one of the major focuses of the Sustainable Development Goals (SDGs) outlined by the United Nations. From the perspective of developing countries like ours with limited facilities and opportunities for diagnosis, treatment and rehabilitation of patients with neuropsychological disorders, developing intelligent systems which can perform early screening can significantly reduce the load of clinical practitioners. Such computerized solutions will not only facilitate domain experts in assessing larger population in significantly less amount of time but can also be used to train non-medical experts to carry out initial screening and refer the suspected cases to the experts. For third world countries like Pakistan, where importance of mental health is relatively undermined and public support systems are near non-existent, the proposed system can prove useful for providing public level awareness and effective rehabilitation at an early stage.

From the perspective of visual-perceptual disorders, identifying such problems at an early age can lead to effective and timely rehabilitation and can serve to prevent further escalation of mental health deterioration in children. In this regard, academic institutes are the direct beneficiaries as

they can easily integrate our proposed system for early screening of their students. Since intellectual development is directly related with academic performance, potential ID students can be given special treatment like extra tutoring session, modified curriculum and regular counselling sessions, to nurture their learning and social skills.

1.5 Thesis Contributions

In pursuit of the aforementioned objectives, several contributions have been made in this thesis. The prominent ones are outlined as follows:

- An important contribution of this research is the adaptation of deep learning methods, more specifically, the Convolutional Neural Networks (CNNs), for the computational representation of neuropsychological domain knowledge, despite the scarcity of data. By employing cross-domain transfer learning and deformation-specific augmentation, we are able to model a number of visual-motor and visual-perceptual deformations from limited graphomotor samples. We evaluated the performance of several ConvNet architectures pre-trained on ImageNet [86] for feature extraction. The impact of depth and width of the model on the predictive potential of the extracted features is analyzed. Various combinations of ConvNet feature extractors and supervised machine learning algorithms are also evaluated. It is an important exploratory contribution for future studies as such analysis has not been reported previously. In this way, the results of our experiments can serve as a baseline for future researchers interested in this area.

- We presented a robust and generic framework to model and classify a wide variety of deformations in samples that are acquired by both online and offline means. For this purpose, we designed two case studies (explained in Section 1.6). In the first case study, we employed an online dataset comprising of eight graphomotor tasks performed by Parkinson's patients and healthy controls. A technique is presented for generating offline images from the online signals captured by a digitizer tablet. The proposed technique not only enabled us to extract visual features from the samples but also provides means of combining static and dynamic attributes to generate dynamically enhanced images for future analysis. The second case study employs offline samples of a multi-template BGT test, that is scored using Lacks' scoring manual. The samples required segmentation of individual shapes for analysis for which two segmentation techniques are presented. The first technique comprises three steps where perceptually grouped shapes are segmented hierarchically to allow de-cluttering. The second technique proposes the application of convolutional object detectors for the localization and segmentation of the nine BGT shapes. To the best of our knowledge, convolutional object detectors have not been employed on handwriting or drawing samples previously. Once prepared, the samples in both scenarios are then used to extract deep visual features that are then used to train deformation-specific classifiers. Extensive empirical analysis

validates the effectiveness of both methodologies in estimating and classifying the target deformations. It is important to mention that modeling human perception is a challenging task and therefore, has not been attempted extensively in the relevant literature. However, in this thesis we attempt to model eleven visual perceptual deformations, that is a unique effort in this direction.

- Several deformation-specific augmentation techniques are presented for enriching feature representation in both scenarios. For visual-motor deformations two non-linear transformations of raw data are suggested to capture fine imperfections caused by associated motor dysfunctions. Both empirical and statistical analysis validates the success of the proposed representations. This enables early detection of Parkinson's disease, where motor deficit symptoms are not yet severe. For visual-perceptual deformations several techniques are proposed to generate near-realistic data that was validated by the domain expert. Empirical results demonstrate that the features extracted from augmented data can effectively represent the specific deformation. This can help facilitate future integration of deep learning based solutions in this domain.

- Our proposed visual-motor deformation estimation and classification methodology performs effectively on conventional graphomotor tasks. This is contrary to the popular procedural analysis based techniques presented in the literature, that do not perform well on conventional templates and suggest non-conventional tasks. Our proposed approach does not require original template modifications.

- Our proposed visual-perceptual deformation estimation and classification methodology can effectively identify a particular deformation across multiple templates. It can also identify deformations with in the same shape class. This enables shape-invariant deformation modeling, that has not been attempted before, since CNNs have primarily been used as explicit *shape recognizers* only.

- By training independent deformation-specific classifiers, we introduce a method to identify multiple deformations present in the same task independently. This avoids the need for extensive heuristics that are otherwise commonly used in such scenarios.

1.6 Visual-Motor and Visual-Perceptual Deformation Identification– Case Studies

As discussed earlier, we target two types of graphomotor deformations in our study, visual-motor and visual-perceptual. Visual-motor deformations are studied using Parkinson's disease as a case study while for visual-perceptual deformations, we employ the BGT responses of children. For completeness, we provide preliminary details of each of these case studies in the following.

1.6.1 Overview of Parkinson's Disease Associated Disorders

Parkinson's disease (PD) is a neurodegenerative disorder that affects the coordinated movements of a person due to loss of dopamine producing neurons in substantia nigra [87]. According to studies [88, 89], it is one of the most prevalent neurological diseases after Alzheimer's, with an average onset age of 60-65. The term 'Parkinsonism' encompasses common cardinal symptoms (mostly experienced by PD patients), such as tremor, slow movement, and muscle stiffness [90]. Handwriting abnormalities represent one of the major symptoms in PD and thus can be employed as an objective screening tool [91, 92]. Three most established clinical manifestations of PD targeted in these studies are 'Micrographia', 'Bradykinesia' and 'Tremor'.

- 'Microgrphia' or abnormal reduction in writing size, is a commonly observed event associated with PD. According to studies [35, 93, 94], it becomes difficult for a patient to maintain the size and alignment of the produced graphomotor impression (handwriting and drawing) due to the impaired control of the extension of the wrist. Consequently, templates that involve wrist movements in all four directions, such as the Archimedean spiral [95] and Luria loops [14] (as shown in Figure 1.4-a and Figure 1.4-b respectively), are best suited to capture micrographic effects. In handwriting based tasks, it is observed [96], that reduction of letter size is not much explicit in languages consisting of words with variable sized letters. On the contrary, reduction of length of a longer sentence or a long string of repetitive letters or characters (as shown in Figure 1.4-c and Figure 1.4-d respectively), can provide a better insight into the micrographic tendencies.

- 'Bradykinesia' or slowness of movement (either due to motor or cognitive dysfunction) [97, 98], causes a potential PD patient to complete a graphomotor task in more time than usually required. Dual tasking based tests, like copying an unfamiliar pattern, helps in identification of bradykinesia resulting from cognitive dysfunctions [99]. This is due to the fact that individuals with potential PD symptoms exhibit difficulties in anticipation capability (i.e. to plan forthcoming strokes while writing current sequences). This can be observed by measuring the pauses between writing characters [100] and drawing the upcoming component of a complex pattern [101].

- 'Tremors' are involuntary to and fro movements that can result from a PD associated condition called 'Akinesia' (i.e. loss of voluntary muscular movements) [36]. This can result in an irregular formation of characters and drawings due to the jerking movements of hands while following a smooth pattern in a particular direction.

The aforementioned visual-motor deformations can either coexist or are present independently, depending upon the type and progression of the disease. Recently a term 'PD Dysgraphia' has been suggested to summarize many of the behavioral, clinical and physiological impacts of PD [103]. Successful identification of any one of these indicators from the graphomotor samples of patients can assist in early prediction of the disease. However, automatic quantification of visual-motor

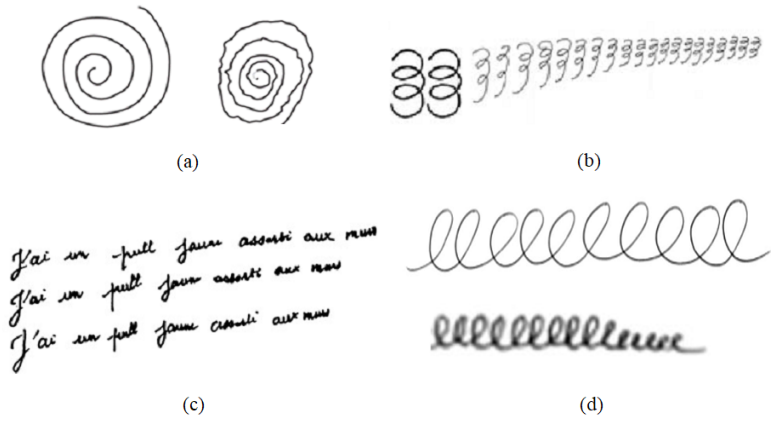


Figure 1.4: Original Templates and their responses drawn by PD patients demonstrating micrographia; (a) Archimedean Spiral [102] (b) Luria Loop [93] (c) Handwritten sentence [35] (d) Repetitive 'l's [102]

.

features associated with PD is a challenging task, that is highly dependent on the choice of sample acquisition mode, selection of template and attribute representation. In the literature, both static (spatial and geometric) [49, 50] and dynamic (kinematic, pressure, temporal, non-linear dynamics and neuromotor) [69, 75, 78, 81–83, 104, 105] features have been extracted from various graphomotor tasks (spirals, meanders and handwriting etc.) produced by PD patients to characterize one or more associated visual-motor deformations. It is important to mention that a significant number of studies rely on dynamic features for the characterization of PD-related deformations. As mentioned earlier, capturing dynamic information requires specialized hardware and also results in modifying the data acquisition protocols. We, consequently, aim to seek enriched static (visual) representations that can perform comparatively to these dynamic features.

1.6.2 Overview of Bender Gestalt Test (BGT)

Bender Gestalt Test (BGT) is a popular drawing based psychometric test employed by clinical psychologists for the screening and differential diagnosis of various neuropsychological and neurological disorders [106–110]. The test comprises a set of nine different templates or gestalts as shown in Figure 1.5-a. The test conduction protocol requires the subject to copy each template on a single sheet of paper using a pen/pencil (Figure 1.5-b). Since BGT is primarily a visual-perceptual assessment test, therefore focus is on the outcome rather than the procedural strategy involved. Several scoring systems [22, 32, 111–113] have been proposed for the estimation of the deformations in BGT drawings, however, 'Lacks' scoring system' [31] is popularly employed by practitioners and is being considered in this study as well.

Lacks' scoring system is based on eleven perceptual discriminators of brain dysfunction inspired by the Gestalt psychology. It determines the presence/absence of these deformations using some or all of the nine BGT templates. The deformations and the templates on which they are applicable

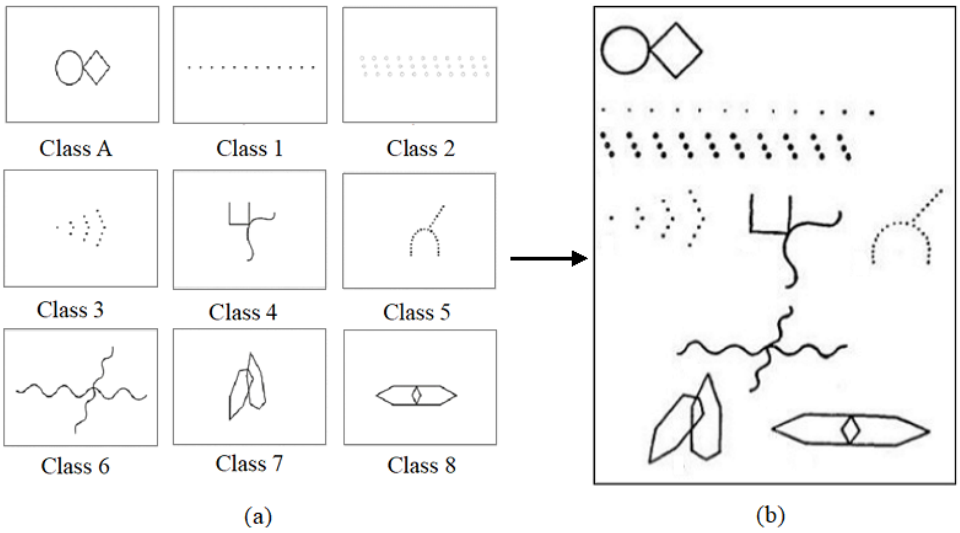


Figure 1.5: BGT test protocol (a) Each card is shown individually (b) Subjects draw samples on a single sheet of paper

are presented in Table 1.1, followed by a brief description of each. Figure 1.6 shows some of the examples of Lacks' deformations, scored in different BGT shapes.

Table 1.1: Scoring Sheet using Lacks' Scoring System

	BGT Shape Class								
Deformations	**A**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Rotation	✓	✓	✓	✓	✓	✓	✓	✓	✓
Overlap	X	X	X	X	X	X	✓	✓	X
Simplification	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fragmentation	✓	✓	✓	✓	✓	✓	✓	✓	✓
Retrogression	✓	✓	✓	✓	X	✓	X	✓	✓
Perseveration	X	✓	✓	✓	X	✓	X	X	X
Collision	✓	✓	✓	✓	✓	✓	✓	✓	✓
Closure	✓	X	X	X	✓	X	X	✓	X
Motor Incoordination	✓	✓	✓	✓	✓	✓	✓	✓	✓
Angulation	X	X	✓	✓	X	X	X	X	X
Cohesion	✓	✓	✓	✓	✓	✓	✓	✓	✓

1. **Rotation:** Lacks suggests that if a subject introduces rotation beyond a certain degree (i.e. more than 80° and less than 180°) in the reconstructed drawing, then it is marked as rotation error. It is observed across all nine BGT templates, however, two examples of rotation in BGT template A are shown in Figure 1.6-a.

2. **Overlap:** One major indicator of perceptual disorder is the inability to overlap portions of a shape correctly (Figure 1.6-b.). Lacks' scoring manual terms it as overlapping difficulty. It includes various conditions like overlapping at wrong places, the omission of portions which overlap, simplification, reworking, and distortion at points of overlap. It is only marked in BGT shapes 6 and 7.

3. **Simplification:** Although marked across all nine BGT shapes, yet simplification is highly shape dependent. For instance, for BGT template 1, a substitution of circles for dots is considered as simplification (Figure 1.6-c.), whereas in BGT templates 6 and 7, significant separation of overlapping portions is considered as simplification. Similarly, significant gaps between the two joining parts of BGT templates A, 4 and 8, are also considered as simplification errors. In BGT template 5, it is the joining of dots into continuous connected strokes.

4. **Fragmentation:** If the subject misses primitive components of a shape or changes their organization in a way that destroys the original template, then it is classified as fragmentation, as shown in Figure 1.6-d. This error is also scored for all BGT shapes.

5. **Retrogression:** Retrogression is marked when a primitive shape is substituted in place of an original one. For instance, in BGT shape 2, if circles are substituted with persistent loops or if dashes are substituted for dots, in shapes 1 and 5, it is considered as retrogression. In BGT shapes (like A, 7 and 8), which involve polygons, the substitution of a primitive polygon for an advanced one (e.g. triangle or square for diamond or triangle or rectangle for hexagon, as shown in Figure 1.6-e), is also considered as retrogression.

6. **Perseveration:** Perseveration is marked if the number of dots in BGT template 1 exceeds 14 or number of columns of BGT template 2 exceeds 13 (Figure 1.6-f) or if another row is added in BGT shapes 2 and 3 respectively. It is also marked if dots of BGT templates 3 and 5 are replaced by circles or, circles of BGT template 2 are replaced by dots.

7. **Collision:** As discussed earlier, BGT test responses are drawn randomly, by subjects, on a single sheet of paper. Although we are considering individual segmented drawings, nevertheless, at times the drawn shapes are either colliding (Figure 1.6-g) or are extremely close to each other. This condition is termed as collision and it is considered in this study i.e. after the drawing is segmented, if it still contains parts of a neighboring shape then it is marked as collision error.

8. **Closure:** Difficulty in reconstruction (e.g. open ends, gaps, overlapping, pressure difference, distortion, etc.) of closed shapes like circles, diamond, and polygons, of BGT shapes A, 7 and 8 or, inability to join adjacent parts of BGT shape 4, is considered as closure difficulty. Some examples are shown in Figure 1.6-h.

9. **Motor Incoordination:** Motor incoordination is indicated by irregular, tremored lines with increased pressure as shown in Figure 1.6-i. It is marked across all templates.

10. **Angulation:** BGT shapes 2 and 3 are specifically designed to assess the angulation maturity of a subject. The inability to produce angulation of parts of the templates (Figure 1.6-j) or reconstruction of the whole shape at angles greater than 45° but less than 80°, is considered as angulation error.

11. **Cohesion:** Significant size disparity between different parts of a template (Figure 1.6-k), is scored as cohesion. Generally the size difference of one-third between parts is considered as deformation. Cohesion is also marked when there is a one-third increase or decrease in the size of a whole shape as compared to the dimensions used to draw the other shapes. Nevertheless, considering conditions where relevance with other shapes is made to mark the presence or absence of an error, is beyond the scope of this study.

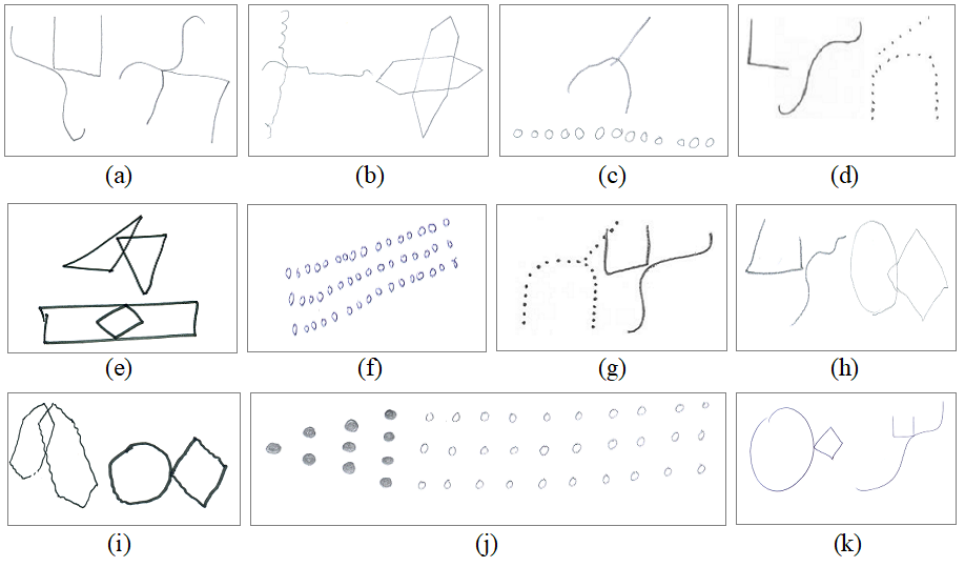


Figure 1.6: Examples of deformations (a) Rotation in BGT template 4 (b) Overlap difficulty in BGT template 6 and 7 (c) Simplification error in BGT template 5 and 1 (d) Fragmentation in BGT shapes 4 and 5 (e) Retrogression in BGT shapes 7 and 8 (f) Perseveration in BGT template 2 (g) Collision of BGT shapes 5 and 4 (h) Closure difficulty in BGT shape 4 and A (i) Motor Incoordination in BGT template A and 7 (j) Angulation in BGT shapes 2 and 3 (k) Cohesion in BGT templates A and 4

A detailed study of the scoring system under consideration reveals the complexity of the problem at hand. Some of these are outlined below.

- Several conditions determine the presence or absence of a particular deformation in a shape.
- Deformations are highly shape dependent and may be measured differently across different templates. For instance, in BGT shape 1, retrogression error is scored when a subject substitutes circles with loops, whereas in shape 2, it is scored when there is a substitution of dashes for dots. Similarly, in BGT shape A, substitution of triangle or square in place of diamond is marked as retrogression, whereas in shape 8, it is the substitution of rectangle for hexagon.
- Some conditions may be considered as an error in one shape while normal in another. For instance, if BGT shape 2 and 3 are rotated at an angle greater than 45° but less than 80°, it is considered as angulation, whereas, in all other BGT shapes any rotation which is less than 80° is considered normal.

- Same characteristics may be considered as one type of error in one shape while another type of error in a different shape. For instance, substitution of circles for dots in BGT shape 1 is considered as simplification, while the same condition when applied on BGT shapes 3 and 5 is considered as perseveration.
- Multiple deformations can co-exist in a single shape.

Based on these observations, it is evident that modeling the BGT deformations is a highly challenging task. Image analysis based techniques would require extensive heuristics which will still be insufficient to cover the large variety of possible scenarios. Consequently, richer feature representations need to be sought which could generalize across multiple shapes and a variety of errors.

1.7 Thesis Organization

Brief description of the main chapters of the thesis is provided as follows.

- Chapter 2 reviews the existing techniques employed in the domain of computerized analysis of neuropsychological drawings and handwriting. Based on the literature survey, it proposes a taxonomy to categorize the state-of-the-art. It also identifies the gaps in the existing literature, that need to be addressed.
- Chapter 3 provides a detailed description of the proposed methodology for deformation modeling and estimation using CNNs. The chapter first provides the theoretical support for employing CNNs. It then highlights the challenges that have hampered the use of deep learning-based solutions for the problem under consideration till now. Finally, it explains how we have addressed these challenges and adapted deep-learning based algorithms to model graphomotor-based deformations.
- Chapter 4 presents the application of our proposed deformation estimation and classification process for the identification of visual-motor deformations associated with Parkinson's disease. A benchmark dataset is employed for this purpose. Several experimental scenarios are described and the results and their detailed analysis are also reported. Key findings and comparison with the state-of-art are also provided in the chapter. Some interesting observations regarding the impact of different templates on the performance of the computed features are also discussed.
- Chapter 5 assesses the performance of the proposed method when applied to identify visual-perceptual deformations. A case study of automatic scoring of the Bender Gestalt Test (BGT) based on the Lack's scoring manual is presented in this chapter. This chapter is significant as it attempts to model human perception computationally. This is a challenging task and the proposed technique shows promising results. Due to non availability of relevant dataset, scored BGT samples are collected from the Department of Professional Psychology, Bahria

University, Islamabad, Pakistan. The samples are then employed to evaluate the proposed methodology. Details of the samples and demographics of the participants are also presented in this chapter.

- Chapter 6 concludes the thesis with a comprehensive analysis of the research under taken during the PhD. It lists the key findings and the limitations of the research. Furthermore, it provides directions for future endeavors as well.

- A list of relevant research published during the PhD is provided in the Appendix section. Most of the published research contributions constitute the main chapters of the thesis. However, some publications are exploratory in nature and could not provide meaningful insight in the main flow of the thesis. Nonetheless, the exploratory studies and experiments were significant for selecting appropriate directions and to derive important conclusions during the research. Due to this reason, the results of these experiments have also been reported in the Appendix section. The Appendix section also provides brief details of the relevant funded projects and research grants acquired during this period.

Chapter 2

Review of Computerized Methods for Graphomotor Deformation Analysis

2.1 Introduction

Characterization of graphomotor deformations is a relatively new area of research in the domain of computerized analysis of handwriting and drawings. This chapter aims to provide a comprehensive review of the related work with the objective to highlight the contributions of relevant studies and to outline the challenges discussed. An introductory overview is presented in Section 2.2, followed by the categorization of the work done in this domain. The state-of-the-art has been grouped into 'Visual Analysis Based Techniques' and 'Procedural Analysis Based Techniques', based on the mode of analysis employed. Section 2.3 discusses the visual analysis based techniques that assess a graphomotor response after completion. It also highlights some of the limitations of the existing methods. Section 2.4 discusses the techniques that focus on the procedure involved in producing a graphomotor response rather than the completed outcome. The section also highlights the challenges faced by such systems. Since the prime objective of both techniques (visual and procedural) is to represent domain knowledge effectively, therefore, Section 2.5 outlines the popularly employed deformation representation and estimation methods presented in the literature and discusses their strengths and weaknesses. Lastly, we conclude the chapter by outlining the overall research concerns this thesis aims to address.

2.2 Characterization of Handwriting and Drawing as Biomarkers for Brain Dysfunctions

Over the past few decades, computerized analysis of handwriting and hand drawn shapes has remained an active area of research in the AI and pattern recognition community. Consequently, research in domains like handwriting recognition [114–118], binarization [119], segmentation [120], keyword spotting [121], manuscript dating [122], signature verification [123, 124], writer identification [125] and writer demographics prediction [126–128], has matured considerably and is

being practically applied in fields like *Forensic investigations* [129], *Document preservation* [130] and *Information retrieval* [131]. Contrary to these popular applications (Figure 2.1), automatic analysis of handwriting and hand drawn shapes for the assessment of mental health of an individual or for the prediction of different neurological disorders calls for further attention from the relevant community.

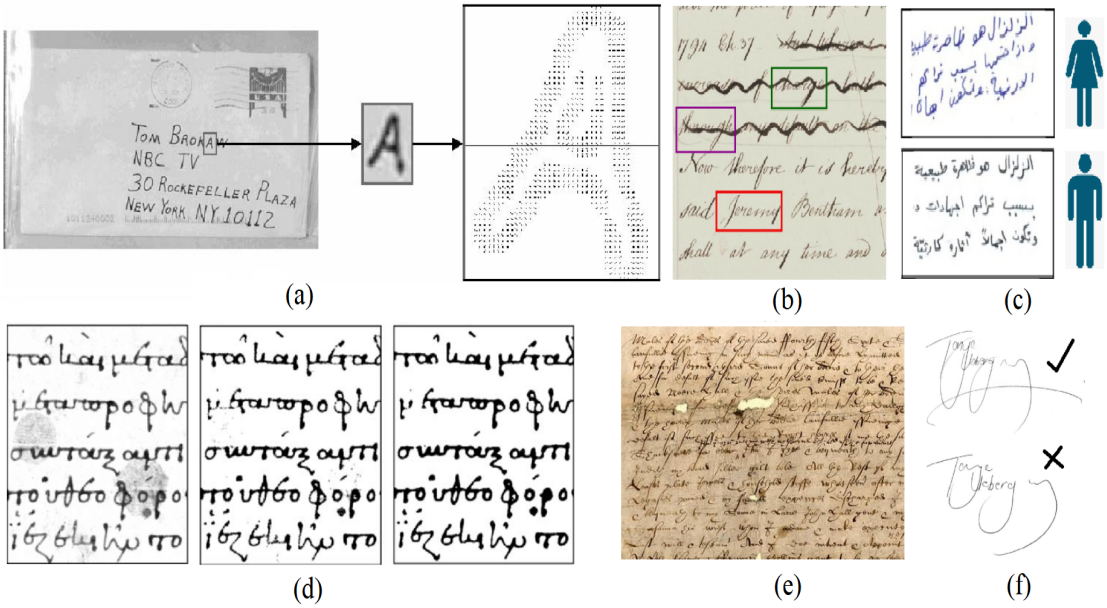


Figure 2.1: (a) Handwriting recognition [132], (b) Keyword spotting [133], (c) Writer demographics [134], (d) Binarization [135], (e) Historical manuscript dating [136], (f) Signature verification [137]

Despite its significance, there are limited contributions is this direction. The highly domain-specific nature of the problem is a major contributing factor in this regard. Therefore, the first step towards developing an effective solution, is the comprehension of the relevant fundamentals. Experts believe that a skilled graphomotor task such as handwriting or drawing, is an interplay among the individual, the task and the environment (as shown in Figure 2.2) [138]. The subsequent sections briefly explain the role of each.

2.2.1 Role of the Individual

The nature of the produced graphomotor impression varies depending upon the developmental maturity, motor-learning and experience of an individual [16]. An individual must possess essential graphomotor skills to produce a complex response like drawing or handwriting. These skills include mature cognitive, perceptual and motor abilities. Motor theorists describe the brain controlled coordinated movements involved in handwriting or drawing as a two-phased loop system [139]. The first phase consists of a *closed-loop*, where the central nervous system receives afferent feedback from vision and other sensory perceptions (finger pressure, hand and wrist movement), regarding the control of subsequent actions required to complete a graphomotor task. During early school

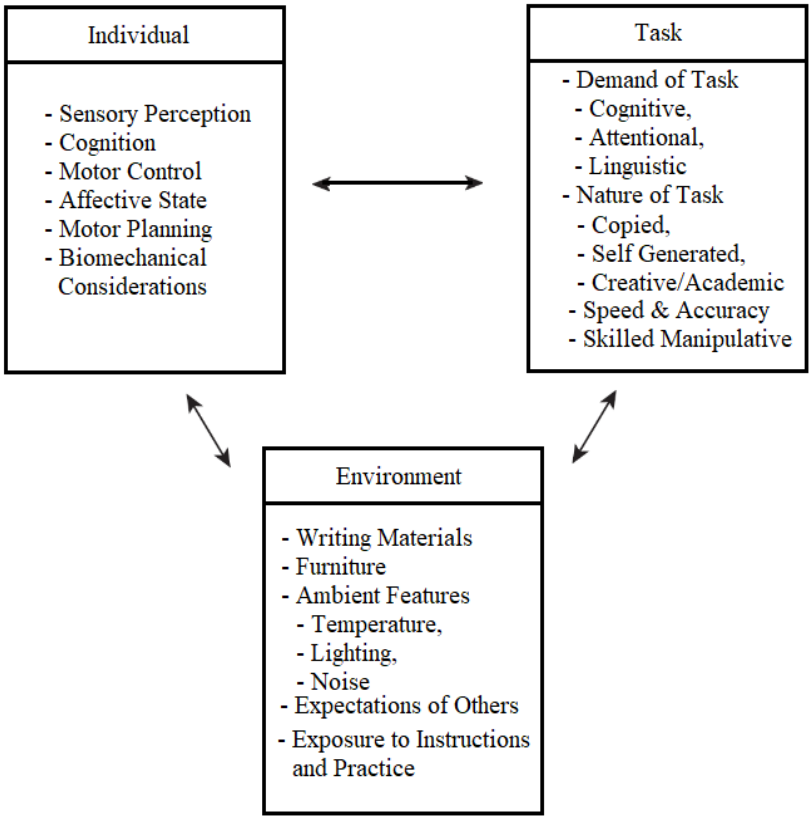


Figure 2.2: Relationship between individual, task and environment [138]

years, children acquire handwriting and drawing skills in a similar manner. However, once the skills are acquired, the control of these coordinated movements shifts to an *open-loop* system, where the central nervous system no longer relies on the afferent feedback [140]. The movements then get pre-programmed in the memory and predominate while producing a familiar graphomotor impression [141]. The transition of motor-learning from closed-loop to open-loop in children with learning difficulties is slower than their peers [142]. This results in different movement patterns between both groups due to an increased reliance of the former on visual feedback [143]. Same can be observed in elderly with dementia while trying to perform a functional task [144].

In addition to visual feedback, the role of kinesthetic information received from muscles, joints, and skin while performing a graphomotor task, is also widely discussed in the literature [145, 146]. An impairment of kinesthesis may influence the refinement of fine motor skills, as observed in patients with neurodegenerative diseases that affect the peripheral nervous system like Parkinson's disease [87]. Consequently, the reduced capability of the brain to coordinate the fingers, wrist and wrist flexion, adversely affects the graphomotor task performance [147]. Another factor impacting the quality of a graphomotor product is the grasp of the writer on the writing tool [148]. Grasp is measured by the nature of contact between the finger or palm with the writing tool and consequently, the movement of the tool on the writing surface. Based on research [149], experts have identified several patterns of grasp suitable for proficient handwriting and drawing. Most popular amongst

these is the 'Dynamic Tripod' grasp [150]. All these attributes impact the quality of a graphomotor impression, and can be assessed independently by employing specific tasks and tools.

2.2.2 Role of The Task

Based on the nature of the underlying graphomotor skill to be assessed, several methods have been designed over the decades. One of the earliest methods employed to measure the developmental level of the drawer/writer, is by means of visual reproduction. For instance, BGT [22] that assesses the visual-motor, visual-perceptual and visual-spatial abilities of an individual [31], requires the participants to reproduce a number of visual stimuli. Reproduction can be performed under two modes i.e. copying or recalling. Another category of tasks belongs to the *completion* method, where parts of the tasks are deliberately left missing and the participants are instructed to complete it. CDT [21], is one of the prominent examples, where the clock circumference is pre-drawn and the participants are required to draw numbers and clock hands. Such tests are usually employed to assess memory and cognition. A recent approach is the *projective* method that attempts to unveil the emotional and perceptual insight of the drawer. One of the frequently employed tests used in this category of tasks is Draw-A-Person (DAP) test [30], that characterizes the intelligence of children based on the Goodenough scoring criteria.

Drawing-based templates are a popular tool for the assessment of various developmental and graphomotor skills. Drawings are well tolerated (among individuals from all age groups) and are not limited by linguistic barriers or educational background. Simple to complex geometric shapes have frequently been used to study various motor and cognitive impairments based on the study objective. Shapes like spirals have been employed to assess motor control deficits like tremors (irregularities due to involuntary mascular movements) [36, 95], in PD patients. Similarly, studies like [94], suggest that templates like spirals and repetitive loops [14] that involve wrist movements in all four directions, can better capture conditions like micrographia (another common PD manifestation linked with the control of the extension of the wrist) [35, 93]. Pentagons are employed to assess apraxia in patients suffering from Mild Cognitive Impairment (MCI) and AD [151]. Similarly, in projective tests like House-Tree-Person (HTP) [152] and DAP, children feel more comfortable in expressing than in a verbal assessment. In addition to drawings, handwriting is also a commonly employed tool in neuropsychology. Clinical practitioners rate the legibility and speed of handwriting in school children to assess learning difficulties [153]. In potential PD patients, handwriting is also employed to capture micrographia by analyzing the abnormal reduction in letter or sentence size. Bradykinesia (slowness of movement due to motor deficits) [97, 98], has also been assessed using handwriting templates like long words and sentences [70, 92]. In MCI and AD, handwriting based functional tasks like *signature* and *writing a grocery list* are commonly employed to assess the underlying cognitive impairments like memory loss, attention deficits and poor executive planning [58].

Selection of the task requires careful consideration as it can lead to ambiguity of results. For instance, due to the extensive use of signature, it often requires a minimal conscious effort to reproduce, and therefore, can prove to be a weak indicator of cognitive impairment. Similarly,

studies like [96] suggest that micrographia is best captured by a spiral or loop template than a word-based template with variable sized letters.

2.2.3 Role of The Environment

Demographic attributes (like age and gender and educational background etc.) of an individual, severity of the disease and duration of treatment (if any) are important factors that can impact the quality of a graphomotor response [58, 154, 155]. For instance, researchers in [43] have highlighted the difference in the graphomotor skills of children under the age of 11 years and those older than 11 (both without any developmental disorders). Similarly, authors in [155] have discussed the handwriting changes in the adult population exclusive of neurodegenerative diseases. This suggests a need for baseline results (or control groups) for excluding the impact of environment related extraneous factors, while assessing the handwriting or drawings of a potential at-risk group. Also in PD, patients under the influence of L-Dopa medication, have known to show difference in some motor aspects of a graphomotor task as compared to those without medication [96].

The nature of the writing implement and the writing surface are also significant influences. With the advent of latest technology, a multitude of devices are now available for sample acquisition. An increasing demand of dynamic handwriting analysis has shifted paradigm towards acquiring data by using digitizing tablets with electronic pens. These devices record handwriting and drawing signals in terms of x- and y-coordinates in different time stamps. Pen trajectories, and pen pressure are also recorded by such devices. An interesting aspect of these devices is the capture of in-air trajectories, where pen is not touching the surface. Despite their apparent advantages, these devices may face reluctance from some groups, especially the elderly subjects. One of the reason is that in some cases, these devices may not provide an immediate visual feedback to the writer and require fixing a sheet of paper on the screen to facilitate users [58]. In cases, where visual feedback is available, getting used to the feel and texture of the writing surface may require some pre-test practice [91]. Some studies [105, 156] have also employed a sensor-based electronic pen called *Biometric Smart Pen (BiSP)* to measure grip on the writing tool. The pen has sensors that record various pressure-based signals. Some practice might be required to hold the pen correctly in order to stimulate the target sensors while writing.

Graphomotor tasks coupled with secondary tasks like tone counting [157], increase the cognitive load of the writer/drawer. In such cases, individuals with certain cognitive impairments, tend to perform poorly due to increased cognitive load and poor executive planning [91, 99]. All these factors must be taken into consideration while designing a research plan. An unintentional extraneous factor can adversely impact the outcome of the experiments and can weaken the confidence of the proposed solution.

2.2.4 Characterizing Graphomotor Deformations in Neurodegenerative Diseases

Neurodegenerative diseases (NGD) affect the peripheral nervous system which controls the muscular movements [158]. Lack of healthy nerve cells causes various movement dysfunctions due

to impaired muscle control. Most neurodegenerative diseases are incurable, but early screening and identification can reduce this 'diagnostic odyssey' to a great extent [159]. Due to the inherent properties of a skilled graphomotor task like handwriting and drawing, it can be employed to assess various cognitive consequences of age-related neurodegenerative diseases [80, 92]. As a result, a significant portion of the work done [43, 47–52, 58, 61, 69, 70, 75, 77–79, 81, 82, 104, 105, 160–162] in computerized analysis of graphomotor tasks targets at finding effective biomarkers for neurodegenerative diseases (like Parkinson's and Alzheimer's disease), with the objective to develop an automatic tool for discriminating between the samples of healthy controls and diseased subjects. Such a tool can provide a low-cost, non-intrusive complementary approach to pathological evaluation performed by the expert clinicians. An extensive amount of work done in this direction is the characterization of visuo-motor deformations in samples of PD patients [49, 50, 61, 69, 70, 75, 78, 79, 81, 82, 104, 105, 160]. PD (Parkinson's Disease) is a neurodegenerative disorder that affects the coordinated movements of a person due to loss of dopamine producing neurons in substantia nigra [87]. Traditional diagnostic methods for PD include invasive procedures like 'Single Photon Emission Computed Tomography (SPECT)' and 'Positron Emission Tomography (PET)' scans, which measure low dopamine levels in the brain [10]. Other diagnostic procedures like 'Magnetic Resonance Imaging (MRI)' and 'Functional Magnetic Resonance Imaging (fMRI)' have also been employed to observe changes in various regions of brain resulting from the impact of PD [11]. However, sophisticated procedures like these require expensive equipment and trained practitioners, facilities that are neither omnipresent nor easily accessible. In addition to cost and accessibility issues, another aspect that is common with most invasive procedures, is the discomfort to patients which consequently leads to diagnostic delays, due to patient hesitancy and resistance. Since early detection is vital for rehabilitation, therefore finding low-cost, easy to administer, and non-invasive alternatives to avoid unnecessary delays, is a worthy research contribution. Any indication of parkinsonian symptoms in these non-invasive tests can then further be verified by invasive ones, if required.

Another neurodegenerative disease addressed in the relevant literature, is Alzheimer's disease (AD). Several graphomotor-based tests have been designed to assess the cognitive dysfunctions indicating AD. The popularly employed ones include the pentagon drawing task in 'Mini Mental Status Examination (MMSE)' [163, 164], the clock drawing task (CDT) [26] and functional writing tasks. The objectives of these tests are to assess the degradation of memory and visuo-spatial abilities of an individual at risk of developing AD. Since neurodegenerative diseases progress with time, the severity of the underlying symptoms may change too, for instance, during the early stages of AD, an individual can develop several Mild Cognitive Impairments (MCI). Studies like [47, 48, 51, 52, 58, 77, 161, 162] attempt to characterize these early indicators of AD from different handwriting samples.

2.2.5 Characterizing Graphomotor Deformations in Neurodevelopmental Disorders

Studies in developmental psychology [165], reveal that the periodic progression of drawing ability in children can give useful insight regarding their physiological and psychological growth and

development. The performance of a child in a drawing task can be linked with his academic capabilities and social skills, that are common characteristics of his developmental age [166]. Based on these assumptions, several drawing based tasks like BGT [22] and DAP [30], have been designed to assess the perceptual-motor abilities and emotional state of children and adolescents. These tests can be integrated into mainstream academic systems for regular assessment of children's mental and emotional health and development. Nevertheless, such integration requires trained professionals and lengthy one-to-one sessions with each and every child. This protocol is both cost and time inefficient. A computerized assessment of children's drawings can increase the efficiency of such programs. In an attempt to provide an automated tool for the analysis of children's drawings (and handwriting), several studies [43, 46, 56, 57, 60, 62, 68, 74] have proposed techniques to characterize the related graphomotor difficulties. However, characterizing human perception is more challenging than characterizing motor difficulties and therefore, requires attention from the relevant community.

2.3 Visual Analysis Based Techniques

This section discusses prominent studies that analyze a graphomotor response after completion, either to discriminate samples of healthy and diseased subjects, or to characterize certain graphomotor deformations. As discussed earlier, we have termed such techniques as the *visual analysis based techniques*. The inspiration of analyzing a graphomotor task after its completion is derived from the conventional clinical procedures of scoring a graphomotor-based test. Such techniques attempt to classify a given sample (as healthy/diseased or error/no-error) by estimating the extent of deviation(s) from the expected visual template(s). For instance, Figure 2.3 shows two responses of templates drawn by a healthy subject and a patient suffering from visuo-spatial neglect (VSN). As expected, the samples drawn by the patient are severely deformed as compared to the ones drawn by the healthy subject.

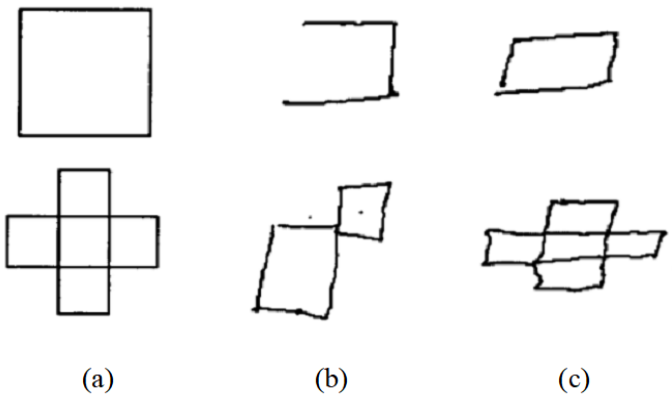


Figure 2.3: (a) Intended templates (b) Response samples drawn by a patient suffering from VSN, (c) Response samples drawn by a healthy subject [46]

Although an easy task for a human expert, it is a significantly difficult one for the machine to determine the degree of deformation based on visual analysis. Nevertheless, in an attempt to design a system to achieve this, two common approaches have been employed in the literature.

- Techniques that assess the deviations at the local level (i.e. primitive component level)
- Techniques that assess the response as a whole by extracting shape-based global features

We will be discussing some of the prominent works carried out in these directions in the subsequent sections.

2.3.1 Component-Level Analysis

The basic notion of such techniques is to assess the quality of the whole figure by estimating the deformations in its constituent parts independently. In one of the pioneering works by Smith and Hiller [43], authors employ analysis of simple cube drawings in an attempt to characterize grapho-motor performance of subjects with visuo-spatial dysfunctions. The study participants are grouped into *Healthy controls (32)*, *Elderly suffering from stroke (6)* and *Children with under-developed graphomotor abilities (30)*. Authors propose a number of static visual features to characterize the samples of the three groups. These include *Line segment distortion (LSD)*, *Parallelism of oblique sides (POS)*, *Angle of oblique lines (AOL)* and *Front panel disproportion (FPD)*. To compute these features, components (like sides and corners) of the drawn cubes are first localized and geometric attributes like pixel-by-pixel distance, angle and orientation are extracted from them, as shown in Figure 2.4. Template matching is then applied to estimate the deviations from the expected stimulus. The effectiveness of the proposed features is then evaluated independently. Preliminary results of the experiments show that POL and LSD outperform the other two in successfully discriminating between the control group and the one with visuo-spatial difficulties (i.e. young children and stroke patients).

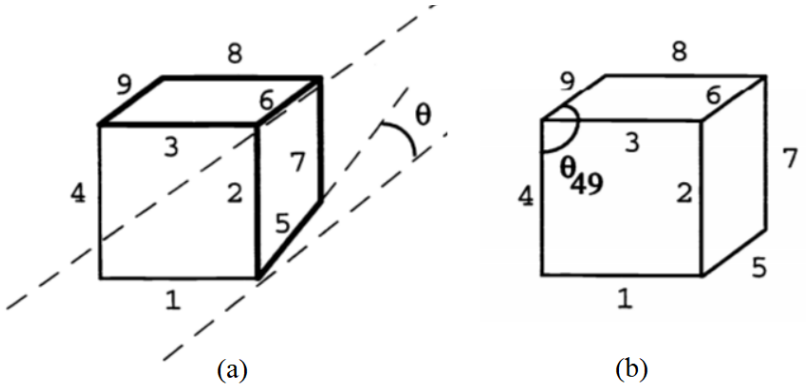


Figure 2.4: (a) Measuring parallelism (b) Measuring oblique angles [43]

In a series of related studies [44, 45], Canham et al. employ a similar technique to score the responses of the Rey-Osterrieth Complex Figure (ROCF) drawing test. According to the

defined clinical standards, there are eighteen scoring sections in an ROCF drawing as shown in Figure 2.5. This requires the system to first localize these sections in the drawing, however, due to the unconstrained nature of the drawn responses, localization, and segmentation of individual scoring sections becomes a highly challenging task. Authors employ fuzzy logic-based metrics based on the Gestalt theory, to localize three out of eighteen ROCF scoring sections. An accuracy of 99.3% is reported in localization of the three regions when tested on a dataset of 31 drawings made by children attending an institute of child health. Once localized, static geometric and spatial features (like size, orientation and position) are extracted from the localized regions. The features are then matched with standard baselines to determine the extent of deviation(s). Authors report an average accuracy of 75% in grading of the ROCF regions under consideration. However, the complete scoring of the ROCF drawing test is not achieved due to localization difficulties of regions of interest.

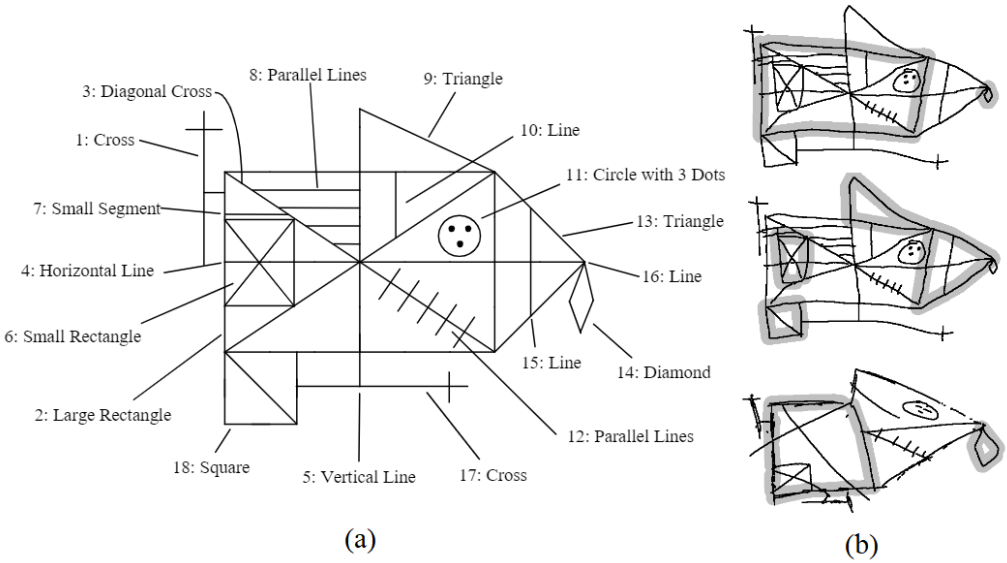


Figure 2.5: (a) ROCF scoring sections (b) Localization of scoring sections [44]

2.3.2 Complete Drawing Analysis

Recently, there have been attempts to analyze and interpret drawing samples of the Clock Draw Test (CDT), popularly employed for the screening of cognitive disorders like dementia [26]. Similar to ROCF, the CDT drawings have a complex scoring criterion. However, unlike ROCF, CDT scoring is focused on the spatial organization of its components rather than the assessment of their independent shape quality. This requires the deduction of inferences not only from the presence/absence of essential shape components but also from their organization (i.e. correct placement of the digits and clock hands). Consequently, an independent assessment of the individual components cannot aid in the complete interpretation of the CDT drawing. In an attempt to score offline CDT samples of dementia patients, authors in [47, 48] present an automated diagnostic tool. Based on a clinical presumption that dementia patients tend to write digits farther from the clock circumference and

tend to increase the size of the digits as well, authors attempt to characterize this notion. 648 CDT samples made by individuals suffering from different forms of dementia are collected from a local institute as part of their regular screening examination. To ease localization of digits, authors map a pre-defined layout on the digitized CDT drawings to divide them into eight different regions as shown in Figure 2.6. A set of 47 geometric and spatial features is then extracted from these regions, some of these are listed in Table 2.1.

Figure 2.6: Mapped layout on CDT sample to facilitate localization of components [48]

Due to the insufficiency of a heuristic-based approach, statistical analysis is employed by the authors to assess the discriminating ability of the extracted features. A network of cascaded machine learning-based classifiers (i.e. *Support Vector Machine (SVM)* with a linear kernel, *Random Forest (RF)* and *K-Nearest Neighbours (KNN)*) is trained, where the first classifier discriminates between the normal and abnormal drawing samples and the subsequent classifiers are to further classify the different stages of dementia. The authors report best accuracies achieved using RF with 500 subtrees. Mean accuracies of 77.78% and 74.38% are reported for the two stages of dementia samples employed in the experimentation.

Table 2.1: Features Extracted from CDT Drawings [47, 48]

Number	**Features**
1.	Number of digits within area 1 (outer area).
2.	Number of digits within area 2 (middle area).
3.	Number of digits within area 3 (inner area).
4.	Number of digits within quadrant 1.
5.	Number of digits within quadrant 2.
6.	Number of digits within quadrant 3.
7.	Number of digits within quadrant 4.
8.	Minimum size of the digits mm2.
9.	Ratio between the maximum digit size and minimum size.
10.	Number of digits outside the contour.
11.	Minimum angle between digits.
12.	Maximum angles between digits.
13.	Number of digits whose orientation is over 25 degree.

An alternative to post-mapping of template, as proposed in [47, 48], Harbi et al. [51, 52], suggest a pre-drawn template to restrict the subjects while drawing. The template is printed on a piece of paper that is then attached to the surface of a digitizer tablet. 65 healthy subjects (aged between 25 to 87) are instructed to draw on the templates. For the patient data, existing offline samples of 100 patients with MCI, Alzheimer's dementia and Vascular dementia are collected from the Llandough Hospital in Cardiff, UK, which originally collected the samples as part of their routine examination. To convert the offline samples into online, each sample is attached to a digitizer tablet and traced by a volunteer. Authors avoid digitizing by means of scanning to overcome the difficulties of localization. Once online samples are obtained for both groups (healthy/diseased), authors formulate an extensive ontology (Figure 2.7) to represent clinical manifestations observed while scoring a CDT sample. The ontology is then converted into fuzzy logic-based heuristics, which are then employed to assess the CDT samples. The authors report a sensitivity value of 99% and a specificity of 95.7% in their experimental study. Although a promising approach, nevertheless, the validity of the samples is questionable and thus requires further investigation in a real life scenario.

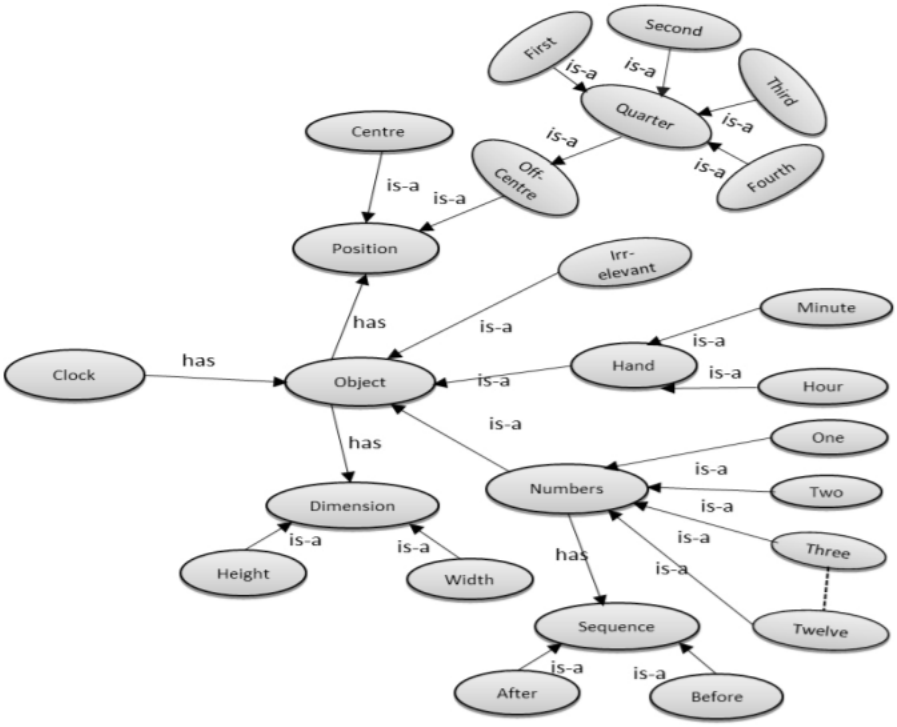


Figure 2.7: Clock ontology employed in [52]

In an attempt to characterize motor deficits like tremor from the spiral drawings of PD patients, Pereira et al. [49], employ offline samples of 55 subjects (37 PD patients and 18 healthy controls). Each participant is instructed to trace a spiral template. The offline samples are scanned and pre-processed to separate the printed template and the drawn trace as shown in Figure 2.8-a. Skeletonization and thinning is then performed and nine spatial features are extracted from each of

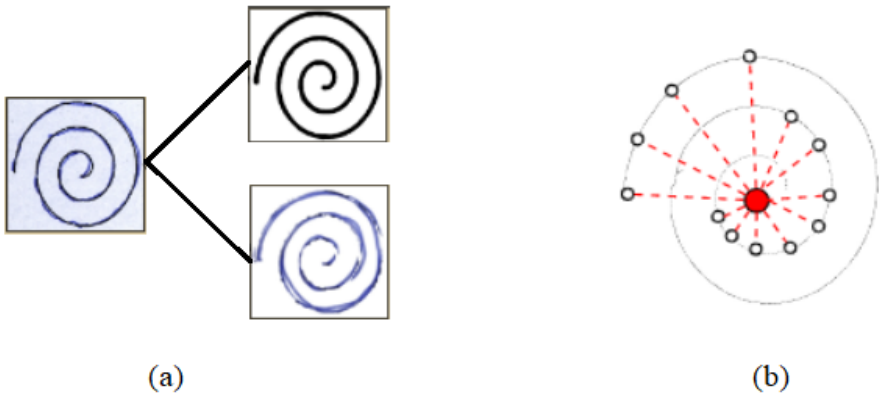


Figure 2.8: (a) Separating spiral template and drawn trace, (b) Computing Mean Relative Tremor [49]

the templates and the corresponding traces as outlined in Table 2.2. The most significant feature extracted is described as the 'Mean Relative Tremor (MRT)'. It is defined as the mean difference between the radius of a given sample and its left-nearest neighbours (Figure 2.8-b). The static features are then used to train three classifiers (i.e. *Naïve Bayes (NB)*, *Optimum-Path Forest (OPF)*, and *SVM* with a radial basis function (rbf) kernel). Authors report classification rates of 78.9%, 77.1% and 75.8% respectively, with each of the three classifiers.

Table 2.2: Features Extracted from Spiral and Meander Drawings [49, 50, 81]

Number	**Features**
1.	Root Mean Square of the difference between the template and trace radius
2.	Maximum difference between the template and trace radius
3.	Minimum difference between the template and trace radius
4.	Standard Deviation of the difference between the template and trace radius
5.	Mean Relative Tremor
6.	Maximum trace radius
7.	Minimum trace radius
8.	Standard Deviation of trace radius
9.	Number of times the difference between trace and template radius changes sign

The authors later extended their work in [50], where a new template *Meander* is introduced (Figure 2.9). More samples from PD patients were added to the original dataset 'HandPD', proposed earlier in [49], increasing its size to 92 samples (74 PD and 18 healthy). The study is exploratory in nature as it assesses the performance of spiral versus meander and reports slight improvement in results. The study also suggests that combining both tasks degrades the performance. A high sensitivity rate is achieved, however, it is attributed to the fact that 80% of the dataset comprises PD samples, resulting in a much low specificity. This imbalance is improved in the latest version of

dataset called 'NewHandPD', proposed in [105]. In [81], authors assessed the performance of same features described in Table 2.2 using a 'Cartesian Genetic Programming (CGP)' [167] approach and reported similar results.

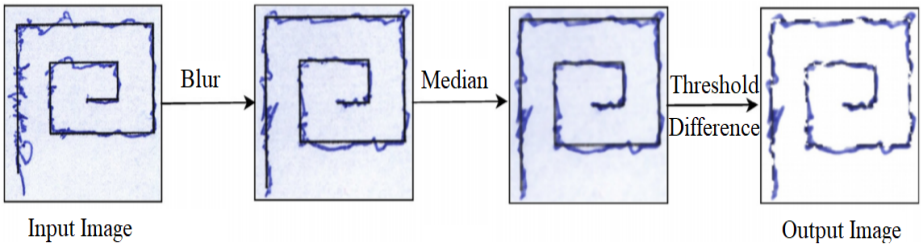


Figure 2.9: Extracting hand-drawn meander from printed template [50]

2.3.3 Limitations/Challenges in Visual Analysis Based Techniques

Prominent works that analyze a graphomotor task after its completion, discussed in the preceding section are summarized in Table 2.3. A critical analysis of the such techniques reveals that studies [43–45] that assess the quality of primitive components or analyze their spatial organization [51, 52], are characterized by the challenges of localization and segmentation. Localization of the intended segment is mostly affected due to the inherent imprecision and ambiguity of a freehand drawing. To overcome the challenges of localization, studies like [51], suggest mapping of a layout on top of the image or to restrict drawers/writers by printing a pre-drawn template. Such constraints may ease localization, however, these modifications can add further complexities to the analysis by introducing a pre-processing step to separate the drawn response from the pre-drawn trace. To address all these issues, a holistic approach of analysis is required, that can identify local deformations without the complex phase of primitive extraction.

Studies that attempt to discriminate between the samples of healthy controls and diseased subjects by employing simple geometric shapes like cubes [43], spirals [49] and meanders [50], rely on the analysis of static geometric features to assess the quality of the drawn template. Feature analysis is performed either by employing template matching or statistical approaches. While these features may prove effective for such shapes, they are insufficient to characterize complex graphomotor deformations scored in tasks like CDT, ROCF and BGT. Due to the insufficiency of static features, such studies [51, 52] have to employ extensive heuristics to estimate the extent of deviation(s). Despite being exhaustive a heuristic-based approach lacks the robustness of a statistical-based approach. Machine learned features and their statistical analysis can overcome the limitations of feature insufficiency and lack of robustness, especially in case of tasks with complex scoring criteria like BGT.

Table 2.3: Summary of Prominent Studies Employing Visual Analysis Based Techniques

Study	Tasks	Samples	Analysis	Findings
Smith and Hiller [43] (1996)	Necker's Cube	32/36 (Online)	Four static geometric features extracted from drawings of patients with VSN are compared with those of healthy controls	Two out of four features prove more successful discriminators than others
Canham et al. [45] (2005)	ROCF	31 (Offline)	Localization of 3/18 scoring regions using fuzzy-logic based heuristics followed by extraction and matching of static geometric features	Average accuracy of 75% in grading of the ROCF regions under consideration is reported
Bennasar et al. [48] (2014)	CDT	648 (Offline)	Geometric features for digit quality and spatial features for organization are classified using cascaded RFs	77.78% and 74.38% classification accuracies for different stages of dementia are achieved respectively
Pereira et al. [49] (2015)	Spiral	37/18 (Offline)	A set of nine static features including Mean Relative Tremor is computed and used to train three classifiers (NB, OPF and SVM (rbf))	78.9%, 77.1% and 75.8% accuracies are reported with NB, OPF and SVM respectively
Pereira et al. [50] (2016)	Spiral and Meander	74/18 (Offline)	A set of nine static features including Mean Relative Tremor is computed and used to train three classifiers (NB, OPF and SVM (rbf))	Meander template produces better results than spiral, however, combining both adversely affects it
Harbi et al. [52] (2017)	CDT	65/100 (Online)	Ontology-based domain knowledge representation and fuzzy logic-based heuristic analysis of geometric features extracted from clock components	99% and 95.7% classification rates are reported for identifying samples of patients and control subjects respectively
Senatore et al. [81] (2019)	Spiral and Meander	74/18 (Offline)	Same set of nine static features are extracted as in [50] and evaluated using Cartesian Genetic Programming	76.60% accuracy is reported on Meander dataset by employing the best evolved classifier

2.4 Procedural Analysis Based Techniques

Until recently, researchers were focused on extracting effective biomarkers from graphomotor samples after their completion. However, the advent of technology (i.e. digitizer tablets, electronic pens, and wearable sensors), has made it convenient to incorporate the assessment of underlying processes involved in the production of the graphomotor impressions as well. The most prominent

of these are the motor and the executive planning skills of a subject. Commonly employed devices for online sample acquisition include digitizer tablets and electronic pens. As discussed earlier, the functional attributes captured by these devices are x- and y-coordinates of the pen position alongwith their time stamps. In addition to position, pen trajectories and orientation (azimuth and altitude) and pressure, both on-surface and in-air, are also recorded. These functional attributes are then employed to compute several *kinematic* and *cognitive* features that are difficult (and in some cases impossible) to compute by visual analysis of the offline graphomotor sample. Since the inception of dynamic handwriting/drawing analysis, research in this domain shifted towards the following two directions:

- Analysis of hand movement during writing or drawing
- Analysis of executive planning involved in the completion of a graphomotor-based task

Details of each type of analysis are discussed in the subsequent sections.

2.4.1 Hand Movement Analysis

For a long time developmental psychologists believed that motor and cognitive skills are two distinct brain functionalities. However, recent theories [168] suggest a strong correlation between them due to the commonality of their underlying processes such as sequencing, monitoring, and planning. Due to this reason kinematic features are one of the most popularly employed attributes of drawing and handwriting for discriminating between the samples of healthy subjects and patients of various cognitive diseases. In a series of related studies [53–55], authors assess the performance of various temporal and kinematic features in comparison to several static features discussed in [43], in an attempt to identify drawings of patients with conditions like *VSN* and *Dyspraxia*. Simple geometric shapes like the Necker's cube [169], are used as templates to capture both static and dynamic features. Principal Component Analysis (PCA) is employed to select the most effective features outlined in Table 2.4. The studies [54, 55] suggest that a combination of both static and dynamic features further improves analysis.

Table 2.4: Dynamic Features Extracted from Necker's Cube Drawings [54]

Number	**Features**
1.	Ratio of pen-down state time w.r.t the total drawing time
2.	Ratio of deceleration time w.r.t drawing time
3.	Ratio of the number of peaks w.r.t drawing time
4.	Ratio of the time to reach maximum peak speed w.r.t drawing time
5.	Average speed during pen-up state

In another study [61], Heinik et al. propose a combination of different static and dynamic features captured from the online CDT samples of 20 healthy individuals and 20 subjects suffering from 'Major Depressive Disorder (MDD)'. The extracted features comprise pen pressure, azimuth

and a number of spatio-temporal attributes. Statistical analysis is employed to assess different combinations of the extracted features and an accuracy of 81.1% is achieved by employing pressure, segment dimensions and time. The study is exploratory in nature and is designed to assess the effectiveness of CDT drawings in identifying MDD. In a similar attempt, authors in [79] analyze the visual and procedural information obtained from the CDT drawings of healthy controls, PD patients and individuals with MCI. Visual analysis of the sample shown in Figure 2.10-b, indicates that PD patients degrade the quality of the clock components due to motor deficits, but show no signs of poor cognition. On the contrary, drawings of MCI patients (Figure 2.10-c) are quality-wise similar to healthy controls, but incorrect placement of clock hands indicates a difficulty in cognition. In case of procedural analysis based on plotting kinematic features like velocity and pressure, samples of PD and MCI patients show more peaks than samples of healthy controls, depicting irregularities due to underlying impairments. While these attributes are successful in discriminating between healthy and diseased samples, they appear to be inconclusive in differentiating between PD and MCI samples. However, the difference is more evident in the visual analysis due to the wrong placement of clock components by MCI patients.

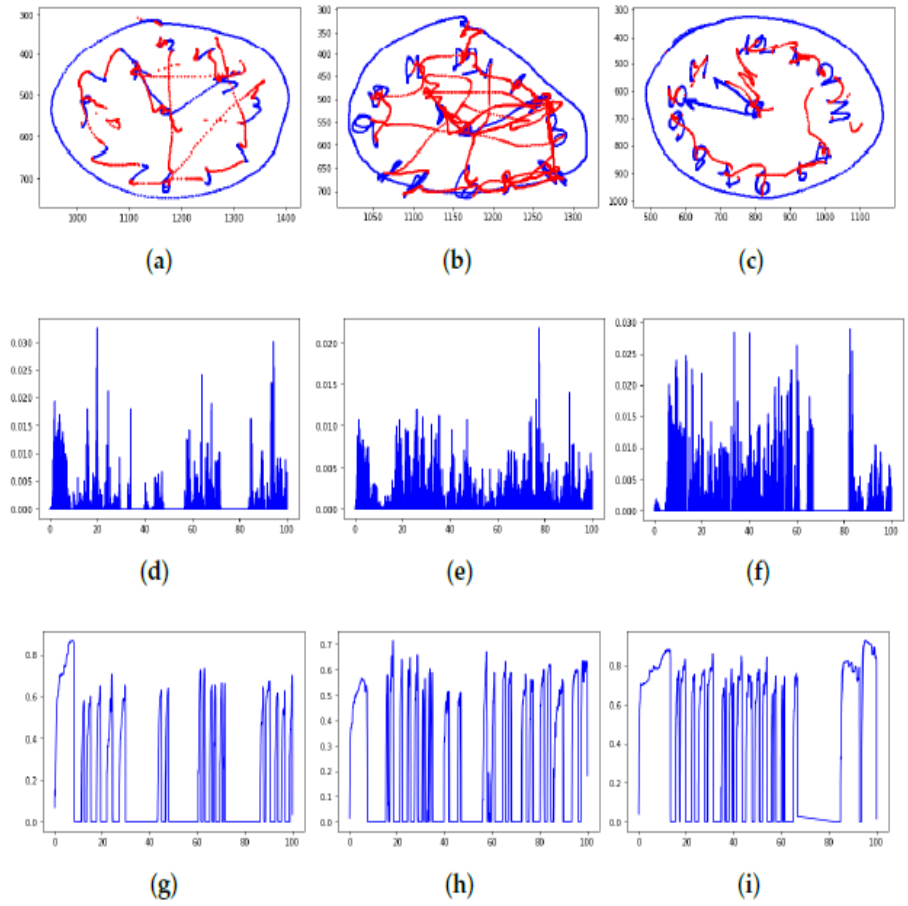


Figure 2.10: (a) Healthy CDT sample, (b) PD CDT sample (c) MCI CDT sample (d) Healthy velocity (e) PD velocity (f) MCI velocity (g) Healthy pressure (h) PD pressure (i) MCI pressure [79]

Initially, online sample acquisition aimed to facilitate localization of scoring regions [43, 51] in

various graphomotor tasks. However, soon the PR researchers realized the potential of kinematic and pressure features in the analysis of motor deficits resulting from various cognitive diseases, and shifted their focus from traditional visual analysis to kinematic analysis. Most popularly employed assessments include the handwriting and drawing samples of elderly population suffering from various neurodegenerative diseases like PD and AD. While motor deficits become obvious in later stages of AD, these are one of the earliest symptoms of PD. Graphomotor-based parkinsonian conditions include bradykinesia (slowness of movement), tremor (irregularity) and micrographia (shrinking of letters). Visual analysis-based techniques [49, 50], discussed in the previous section, primarily focused on computing tremor and micrographia from offline spiral drawings. Due to the insufficiency of static geometric features in characterization of bradykinesia and related conditions, authors in [105], capture dynamic pressure signals by means of a BiSP smart pen from online spiral drawings of PD patients and healthy controls. The signals are transformed into 2-dimensional images as shown in Figure 2.11. The images are then used to train a shallow 3-layered Convolutional Neural Network (CNN). Samples of 224 PD patients and 84 healthy controls are employed and an accuracy of 87.14% is reported. An interesting aspect revealed by the study is that the meander task which outperformed spirals while evaluating MRT, performs poorly on pressure-based discrimination. This suggests that the choice of template has an impact on the effectiveness of the extracted features.

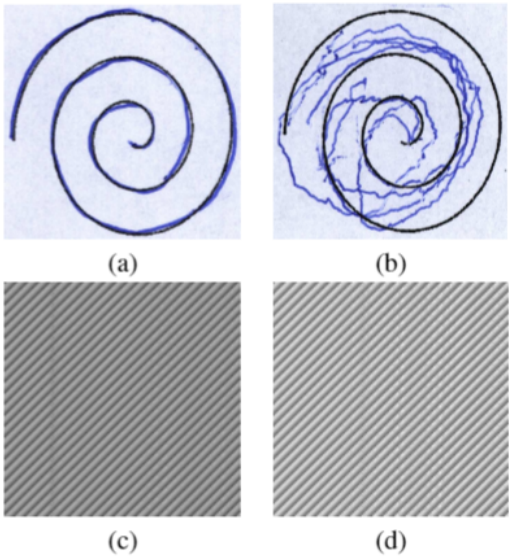


Figure 2.11: (a) Spiral drawing of a healthy subject, (b) Spiral drawing of a PD subject (c) Time series-based image of healthy subject (d) Time series-based image of PD subject [105]

In an attempt to study the impact of templates on the performance of kinematic features, authors in [69], present a template consisting of seven handwriting tasks as shown in Figure 2.12. Several kinematic features (outlined in Table 2.5) are extracted from each task. An SVM classifier with rbf kernel is employed to determine their effectiveness. Some features are single-valued while others consist of multiple-values combined in a vector. The vector values are then converted into five

single values by using different statistical measures (mean, median, 1st percentile, 99th percentile, standard deviation) to represent them.

Figure 2.12: Handwriting tasks proposed in PaHaw database [69]

Table 2.5: Dynamic Features Extracted from Handwriting Tasks [69]

Number	Features
1.	Trajectory during stroke divided by stroke duration
2.	Trajectory during handwriting divided by handwriting duration
3.	Rate at which the position of a pen changes with time
4.	Rate at which the velocity of a pen changes with time
5.	Rate at which the acceleration of a pen changes with time
6.	Velocity/acceleration/jerk in horizontal direction
7.	Velocity/acceleration/jerk in vertical direction
8.	Mean number of local extrema of velocity
9.	Mean number of local extrema of acceleration
10.	Number of changes in velocity (NCV) relative to writing duration
11.	Number of changes in acceleration (NCA) relative to writing duration
12.	Time duration (in seconds)/ length (in points) of writing
13.	Width and height of stroke

All values are further normalized before training the classifier. Authors first find the most effective features for each task (e.g. standard deviation of stroke speed for task 7) and then report the accuracies of all tasks using all features. The highest accuracy with combined features is obtained on the sentence task i.e. 78.7%. The authors also extract the same features from the popular spiral task and report an accuracy of 65.4%. An accuracy of 79.4% is reported by combining all 8 tasks.

Authors suggest that given kinematic features are more effective on handwriting tasks than on the conventional spiral task.

Drotár et al. further extended their work in [70], where they introduce a new modality called the 'In-air Movement' for the quantitative evaluation of PD samples. The original set of features presented in [69] alongwith some additional features, are now computed from both the on-surface (pen-down) and in-air (pen-up) mode. The highest classification accuracy of 80.09% is reported by using 16 in-air features extracted from combined 8 tasks. By increasing the features in [82], authors report an increased accuracy of 85.61% using a combination of both modalities. In [71], authors added new features based on energy and entropy, in the existing set of kinematic features described in [69, 70]. All features are extracted from the seven handwriting tasks shown in Figure 2.12. The authors report an accuracy of 88.1% by employing a feature set of size 162. In an attempt to find the most contributing modality of dynamic handwriting, Drotár et al. in [104] compare the performance of on-surface, in-air and pressure features. An area under the curve (AUC) of 89.09 is reported for on-surface features, while those of 74.16 and 83.83 are reported for in-air and pressure features respectively. In [75], the effectiveness of kinematic and pressure features extracted from the handwriting tasks discussed earlier, is compared. The authors report an accuracy of 82.5% using only pressure features extracted from all 8 tasks (spiral and 7 handwriting). As compared to pressure features, kinematic features produced an accuracy of 75.4%, while combining both achieved an average accuracy of 81.3%. Authors evaluated the performance of two additional classifiers (AdaBoost and K-NN) as well. Nevertheless, highest accuracies are obtained using SVM with rbf kernel.

The main contribution of Drotár et al. is the compilation of the benchmark dataset 'Parkinson's Disease Handwriting Database (PaHaW)' [69, 75], comprising several handwriting tasks alongwith the conventional Archimedean spiral. It is popularly being employed by studies that intend to assess the effectiveness of novel handwriting features for evaluating PD samples. Recently, authors in [78] presented an optimization of current modalities (on-surface/in-air) by employing 'Fractional Order Derivatives' on the spiral drawings of PaHaW database. In another study [83], Impedevo et al. evaluate various task combinations of PaHaW dataset for the best classification of PD samples. Authors report the highest accuracy of 74.76% by employing tasks 3, 6 and 8 (i.e. bigram, word and sentence).

Although much work has been performed on PaHaW dataset, yet other efforts have also been made in this direction. In [76], authors employ a smaller (and not so popular) dataset 'PDMultiMC' consisting of online samples of 32 subjects (16 PD and 16 controls), to evaluate the performance of various dynamic features described by earlier studies like [69–71, 75, 82, 84, 104]. The PDMultiMC dataset comprises 7 handwriting tasks including repetitive 'lll' like PaHaW and additional single word functional tasks (names, days etc.). By combining all features and tasks, authors report an accuracy of 90.63% using an SVM with rbf kernel. Similarly, authors in [160] employ online samples of 24 PD patients and 20 healthy controls. All subjects are instructed to draw horizontal lines. Kinematic features are extracted from the samples and are analyzed by employing Naïve Bayes (NB) classifier. The authors report an accuracy of 91%. Another study [170] employs four

sentence writing tasks of 33 PD patients and 10 healthy controls to ascertain the predictive potential of in-air modality. The authors report an accuracy of 86.05% by combining kinematic features measured from both in-air and on-surface modalities.

In addition to PD, in-air features have demonstrated efficacy in detection of other age related deformations as well [171]. Werner et al. [58] evaluated the performance of various dynamic features (including kinematic, pressure and temporal) in characterizing AD related graphomotor difficulties. Discriminant analysis is performed to determine the best predictor of AD. In-air temporal measures for AD patients were higher as compared to healthy controls, while the mean pressure was lower. The in-air temporal values for AD increased as the complexity of the task increased. Velocity-based features based on the kinematic theory of rapid human movement [172] (Sigma-Lognormal model [173]) have also been adopted [174]. With a Bagging CART (classification and regression tree) approach, authors were able to achieve an equal error rate of 3%. Other than handwriting-based tasks, authors in [77] employ drawing tasks including two- and three-dimensional figures and CDT, to extract kinematic and pressure features. Discriminant analysis is used in order to classify participants into healthy controls, subjects with MCI and those with mild AD. The authors report high dependency on the choice of task employed as functional tasks outperformed motor ones. Similar observations are reported in [161, 162].

Review of the relevant literature suggests that the focus of most hand movement analysis based techniques is to explore effective discriminators between the samples of healthy controls and diseased individuals. In order to extract meaningful attributes from graphomotor-based tasks various conventional and non-conventional templates have been employed. Both drawing and handwriting tasks have been evaluated and it has been observed that the effectiveness of a procedural attribute is highly dependant on the nature of the task. Studies that extracted such features from drawing-based templates are listed in Table 2.6. While those that extracted features from handwriting are summarized in Table 2.7. To highlight the impact of the task involved, a number of studies have compared the performance of same features extracted from both drawing and handwriting samples. Table 2.8 summarizes the findings of these studies.

2.4.2 Drawing Strategy Analysis

Another trend in the computerized analysis of graphomotor tasks involves observation of the executive planning adopted by the subject while drawing/writing. For instance, some subjects trace the contours of the figure to be copied, others put points first and then connect them with segments, and so on. Although the idea is not so recent, it has gained renewed popularity due to the availability of online data capturing systems. Such an analysis focuses on the behavior and preferred drawing/writing strategy of the subject while producing a response, rather than analyzing the end product or the hand movement. This is also termed as the 'grammar of action' [62]. Based on the same assumption, Remi et al. assess the hand-drawn samples of children with learning difficulties in [56]. A template consisting of handwritten sentences and a set of geometrical shapes is used. The constituent parts of each drawn stimulus are localized using shape-based clustering and a collection of drawing sequences are extracted to determine the signs of learning and writing

Table 2.6: Summary of Prominent Studies Employing Procedural Analysis Based Techniques Using Drawing-Based Tasks

Study	Tasks	Samples	Analysis	Findings
Garbi et al. [55] (1999)	Necker's Cube	32/36 (Online)	Various temporal and kinematic features are combined with static geometric features extracted from drawings of patients with VSN and are compared with those of healthy controls	Discriminating potential increases by combining all features
Heinik et al. [61] (2010)	CDT	20/20 (Online)	A combination of kinematic, pressure and spatio-temporal features are extracted from drawings of healthy controls and MDD patients	81.1% accuracy is achieved in classifying samples of MDD patients
Periera et al. [105] (2016)	Spiral and Meander	224/84 (Online)	Pressure signals captured by BiSP are converted into 2D images and are classified using a 3-layered CNN model	87.14% accuracy is achieved in classifying samples of PD patients
kotsavasiloglou et al. [160] (2017)	Horizontal lines	24/20 (Online)	Kinematic features are extracted and analyzed using NB	91% accuracy is achieved in classifying samples of PD patients
Müller et al. [161] (2017)	3D house drawing	20/20 (Online)	In-air, on-surface and total time-based features are assessed using linear regression	92.5% accuracy is achieved in classifying samples of AD patients
Müller et al. [162] (2017)	CDT	20/20 (Online)	In-air, on-surface and total time-based features are assessed using linear regression	87.2% accuracy is achieved in classifying samples of AD patients
Mucha et al. [78] (2018)	Spiral	37/38 (Online)	Fractional order derivatives of kinematic features are used to train RF and SVM	72.38% accuracy is achieved in classifying samples of PD patients

difficulties. In a similar attempt, authors in [57] employ syntax analysis on the online samples of children's drawings. A combination of selective patterns from different drawing-based tests is used as the set of templates to determine the gender and handwriting ability of each participant.

In an attempt to model the acceleration sequences captured by a digitizer tablet while performing the 'Cube Drawing Test' (Figure 2.13), multiple classifiers and a graph-based genetic programming technique is employed in [60]. A multi-class AUC score of 0.70 on both the training and the test data is obtained by the best-evolved classifier. A total of 120 drawings from 40 subjects (multiple drawings from each subject) are used to conduct the analysis.

A more comprehensive technique of modeling the sketching gestures of subjects is proposed

Table 2.7: Summary of Prominent Studies Employing Procedural Analysis Based Techniques Using Handwriting-Based Tasks

Study	Tasks	Samples	Analysis	Findings
Werner et al. [58] (2006)	5 functional writing tasks	22/41 (Online)	Spatio-temporal, kinematic and pressure features extracted from tasks are assessed using discriminant analysis	72% accuracy is achieved in classifying samples of AD patients using pressure and temporal features
Drotár et al. [71] (2014)	7 writing tasks	37/38 (Online)	Energy and entropy features extracted from tasks are combined with kinematic and temporal features and classified using SVM (rbf)	88.1% accuracy is achieved in classifying samples of PD patients
Pirlo et al. [174] (2015)	Signature	29/30 (Online)	Velocity-based features modeled on Kinematic theory of rapid hand movement are assessed using a bagging cart approach	Equal error rate (EER) of 3% is achieved in identifying samples of PD patients
Taleb et al. [76] (2017)	7 handwriting tasks	16/16 (Online)	A combination of several kinematic, temporal, pressure, energy and entropy features extracted from all tasks are classified using SVM (rbf)	96.88% accuracy is achieved in classifying samples of PD patients using combined decision of 3 tasks
Jerkovic et al. [170] (2019)	4 handwriting tasks	33/10 (Online)	A combination of in-air and on-surface kinematic features from all tasks are classified using LDA	86.05% accuracy is achieved in classifying samples of PD patients

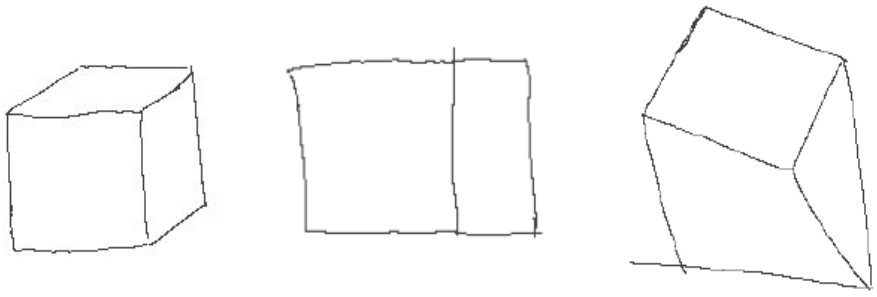


Figure 2.13: Samples of Cube Drawing Test drawn by subjects described in [60]

in [64]. The drawing order of the shape primitives is observed in addition to their organizational relation with the neighboring primitives (Figure 2.14). Based on predefined rules (like specific angle, sequence and the number of primitives required for each shape class etc.), the quality of

Table 2.8: Summary of Prominent Studies Employing Procedural Analysis Based Techniques Using Drawing and Handwriting-Based Tasks

Study	Tasks	Samples	Analysis	Findings
Drotár et al. [69] (2013)	Spiral and 7 writing tasks	37/38 (Online)	On-surface kinematic and temporal features from all tasks are combined and classified using SVM (rbf)	79.4% accuracy is achieved in classifying samples of PD patients
Drotár et al. [70] (2013)	Spiral and 7 writing tasks	37/38 (Online)	In-air features extracted from tasks are combined and classified using SVM (rbf)	80.09% accuracy is achieved in classifying samples of PD patients
Drotár et al. [82] (2014)	Spiral and 7 writing tasks	37/38 (Online)	On-surface and in-air features extracted from all tasks are combined and classified using SVM (rbf)	85.61% accuracy is achieved in classifying samples of PD patients
Drotár et al. [75] (2016)	Spiral and 7 writing tasks	37/38 (Online)	Pressure and kinematic features are extracted from all tasks and classified using SVM (rbf), K-NN and AdaBoost	81.3% accuracy is achieved in classifying samples of PD patients using SVM classifiers
Garre-Olmo et al. [77] (2018)	2D and 3D Cubes and 2 handwriting tasks	23/17 (Online)	A combination of kinematic, temporal, pressure, energy and entropy features are analyzed using discriminant analysis	63.5% to 100% accuracy is reported on different tasks while classifying MCI samples
Impedevo et al. [83] (2018)	Spiral and 7 handwriting tasks	37/38 (Online)	Task-wise performance is evaluated for a combination of kinematic, temporal, pressure, energy and entropy features using RF, SVM, K-NN, NB, LDA and AdaBoost	74.76% accuracy is achieved in classifying samples of PD patients using combined decision of 3 tasks

the produced sketch is evaluated. The system is tested on both hand-drawn samples and synthetic shapes from HHreco [175] dataset. The drawing gestures of patients with disabilities are also analyzed in a series of related studies [63, 65–67], using an optoelectronic system. In another series of related works [68, 74], authors attempt to determine the correlation between the preferred polygon drawing strategy of school children with their handwriting skill development. Drawing sequences of 178 school children (ages 6-7 years), captured using a digitizer tablet are employed. A 63.48% success rate is reported in correlating the estimated drawing strategy and the predicted handwriting performance by using an SVM classifier. A similar approach is previously reported in [62].

Table 2.9 enlists the popular studies that attempt to assess the executive planning and constructional sequencing of the drawer/writer to determine any indicators of dysfunction.

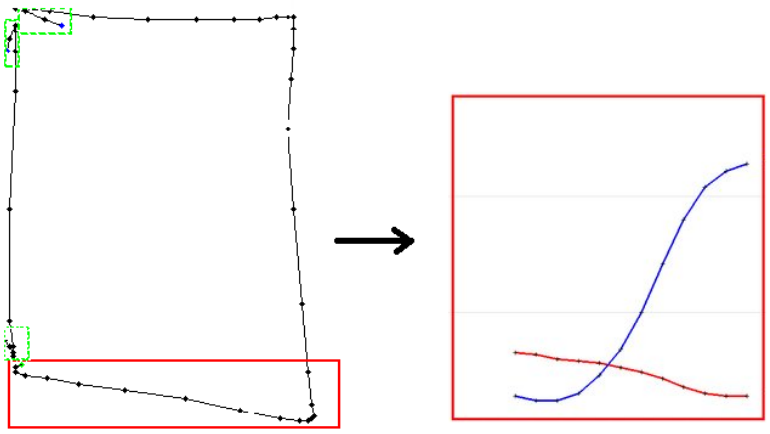


Figure 2.14: A free-hand drawn square with corresponding spatio-graphs [64]

Table 2.9: Summary of Prominent Studies Analyzing Drawing Strategy

Study	Tasks	Samples	Analysis	Findings
Remi et al. [56] (2002)	geometric shapes and writing	10/153 (Online)	Drawing sequences of children with 5 different levels of learning difficulties are assessed using LDA	76.92%, 80%, 67.57%, 61.11% and 70.27% accuracies are reported for each level respectively
Chindaro et al. [57] (2004)	Square and Cross templates	120 (Online)	A set of static and dynamic features is extracted to model stroke sequences which are then analyzed using HMMs	55.9% and 69.5% classification rates are reported for square and cross respectively
Smith and Lones. [60] (2009)	Cube drawings	120 (Online)	Acceleration sequences are analyzed using graph-based genetic programming	An AUC of 0.70 is reported in identifying samples of individuals with VSN
Khalid et al. [62] (2010)	Simple geometric shapes	631 (Online)	Drawing sequences of children with learning difficulties are compared with those of regular children	The study is exploratory and suggests further investigation
Renau et al. [64] (2011)	Simple geometric shapes	1800 (Synthetic)	Based on sketching gestures analysis, accordance between human and synthetic shapes is determined	91.1% accordance is observed
Tabatabaey et al. [74] (2015)	Polygon	178 (Online)	Drawing sequences are classified using SVM to find correlation between preferred strategy and expected outcome	63.48% accuracy is achieved in classifying correlation between drawing strategy and drawn response

2.4.3 Limitations/Challenges in Procedural Analysis Based Techniques

An extensive review of the procedural analysis based techniques has been presented in this section. Two types of procedural analysis have been employed in the literature, those that analyze the hand movement and others that assess the preferred drawing strategy. Both techniques do not depend on the visual feedback of the final response, but rather rely on the features that represent the motor and cognitive functionalities employed during the process. Although, additional information provided by procedural analysis can give better insight regarding disease associated pathophysiological changes, nonetheless, the research in this direction requires more interdisciplinary correspondence to establish norms. For instance, studies like [69, 83, 161, 162] that extract various hand movement-based features, have shown that the nature of task employed for the capture of these novel features has a significant impact on their predictive performance. Same features when extracted from two different templates can result in a different outcome regarding the same disease. Due to this reason, authors in [75], suggest employing non-conventional handwriting based tasks instead of the conventional spiral drawing, for discriminating between samples of PD patients and healthy controls. However, in [105], spiral drawings outperform a non-conventional meander task, when pressure signals are employed. This prompts a careful consideration of feature-task relation to avoid feature validity issues. Modification of conventional test templates can lead to resistance from the target users i.e. clinical practitioners. To bridge the gap between clinical practices and AI-based solutions, the proposed technique must firstly improve results on conventional templates and subsequently suggest modifications for the practitioner to decide.

Dynamic handwriting analysis provides an opportunity to assess various attributes including kinematic, temporal and pressure measures. Nevertheless, almost all studies reviewed in the literature, indicate the insufficiency of these features when employed independently. As a consequence, a combined high dimensional feature vector approach is adopted, as shown in [71, 76, 77, 83, 170]. In some cases [55, 60, 61], these additional features are combined with the static visual features to enhance their performance. An early fusion-based approach can perform adversely due to the different nature of features being combined. Consequently, feature selection has to be employed to select the most effective feature combinations [55]. Due to this reason, a hand-crafted feature designing attempt can increase human effort to much extent. On the contrary, machine learning-based feature extraction can significantly reduce this effort and provide an efficient solution. Therefore, efforts must be made in this direction.

Despite realizing the impact of template selection on feature performance, most of these studies combine the features extracted from multiple templates due to the scarcity of data. This can lead to a degradation in overall performance as suggested in [83]. Instead of combining features extracted from all tasks before training a single classifier, a late fusion-based approach like majority voting should be adopted where decision of every task is taken into consideration separately.

Studies like [70, 82, 161, 162, 170] advocate the effectiveness of a novel modality called ‘in-air’, where kinematic and temporal features can be computed while the pen is not touching the surface. Authors attribute the success of this modality to the fact that the extent of pause

between writing/drawing strokes increases due to underlying motor (in case of PD) or cognitive (in case of AD) dysfunction [100, 101]. Although the rationale is sound and the performance of the modality in discriminating between healthy and diseased samples is promising, nonetheless, it is inconclusive whether same modality can distinguish between samples of diseases with overlapping conditions. This concern has been discussed in [79], where visual analysis-based discrimination of CDT samples drawn by a PD and an AD patient proved more successful than the procedural analysis. A major contributing reason for this is that hand movement based procedural analysis techniques are not designed to characterize visuo-spatial and visuo-perceptual deformations, that are necessary assessments in most cognition-based dysfunctions. Further exploration is required in this direction, until then a strong visual analysis based technique can prove more effective in differential diagnosis.

Techniques that assess the quality of a graphomotor response by analyzing the preferred drawing/writing strategy, present a promising approach as well [56, 57, 60, 62, 64, 74]. Nevertheless, models for correct strategy need to be developed before such techniques can mature. While these techniques are more suitable for rehabilitation purposes, their success in deformation characterization is not yet conclusive and calls for deeper research investigations.

2.5 Deformation Representation and Estimation for Graphomotor Analysis

An in depth review of the techniques presented in both Section 2.3 and Section 2.4 highlights several open issues that require attention from the relevant community. However, the most significant one is the domain knowledge representation. In an attempt to effectively represent the clinical manifestations and graphomotor deformations being assessed, a number of computable feature classes have been proposed in the literature. These include various static and dynamic features specific to the objective of the study. Popularly employed features have been categorized and outlined in Table 2.10.

2.5.1 Spatial and Geometric Features

Most visual analysis based techniques [43–45, 47–53, 60] discussed in Section 2.3 rely on spatial and geometric features. These features are either extracted at component-level or globally. Spatial and geometric features have also been employed in a number of procedural analysis based techniques [58, 76, 81] to enhance the performance of dynamic features. Commonly employed spatial and geometric features include height, width, tangential angle, size, dimension, orientation etc. Spatial features are useful in characterization of visuo-spatial or visuo-perceptual deformations, however, due to limitation of definition and unconstrained nature of the free-hand responses, traditional hand-crafted features prove insufficient for a complete analysis.

Table 2.10: Summary of Deformation Representation Methods Employed In Related Studies

Feature Class	Study	Features
Spatial and Geometric Features	[43–45, 47–53, 58, 60, 76, 81]	Number of strokes/components, stroke/component size, stroke height/width, angles, orientation, radius, MRT
Statistical Features	[69, 70, 78, 83, 104, 104]	Mean, mode, median, 1st percentile, 99th percentile, 99th–1st percentile, standard deviation, first and second derivatives, fractional order derivatives, ratio
Kinematic Features	[55, 58, 61, 69, 76–78, 83, 104, 160, 170]	Horizontal/vertical velocity, horizontal/vertical acceleration, horizontal/vertical jerk, horizontal/vertical displacement, velocity, acceleration, jerk, displacement, stroke speed, number of changes in the velocity per stroke (NCV), number of changes in acceleration per stroke, relative NCV/NCA
Pen and Pressure Features	[58, 61, 75–77, 83, 104, 105]	Position, azimuth, tilt, pressure signals, number of changes in velocity pressure per stroke (NCP), correlation between stroke pressure and velocity/acceleration, relative NCP
Temporal Features	[55, 58, 61, 76, 77, 83, 104, 161, 162]	In-air time, on-surface time, total time, stroke duration, stroke speed/displacement
Non-linear Dynamic Features	[71, 76, 77, 82, 83, 104]	Shannon entropy of x and y position, Renyi entropy second and third order of x and y position, conventional energy of x and y position, first order Teager-Kaiser energy of x and y position, signal to noise ratio calculated from the conventional energy, signal to noise ratio calculated from the teager-kaiser energy, empirical mode decomposition
Neuromotor Features	[174, 176, 177]	sigma-lognormal, delta-lognormal, number of lognormal components

2.5.2 Statistical Features

Handwriting and pressure signals captured by a digitizer tablet or a smart pen have been employed to provide additional information regarding the fine motor skills of an individual. Raw signals are converted into dynamic features by applying certain statistical functions. As a result some single valued features (e.g. number of strokes, speed and time) and some feature vectors (velocity,

acceleration, jerk) are obtained. Several statistical features are then computed from such feature vectors that include median, mean, percentiles, moment and standard deviation etc. [69, 70, 83, 104, 104]. All features are normalized before classification. It is worth noting that the main purpose of these statistical features is to map a time-series based sequential information into a single value. However, vital information may be lost by condensing a sequence into a single parameter. Nonetheless, several representations have been proposed in the literature, most recent being the fractional derivatives of dynamic features [78].

2.5.3 Kinematic Features

The Cartesian coordinates of the pen position captured by the device while completing a grapho-motor task, can be employed to compute the pen trajectories in horizontal, vertical and tangential directions with respect to pen-up and down states. A single connected trace of the handwritten or drawn pattern on the surface is considered as a *stroke*. In some studies [70, 82, 83], in-air strokes have also been segmented based on pen trajectories captured during pen-up state. The cartesian distance between the two consecutive strokes is called displacement and is used to compute a number of kinematic features [55, 58, 61, 69, 76–78, 83, 104, 160, 170]. Due to the high sampling frequency of these devices, a good approximation of the rate of change of position with respect to time (velocity) can be calculated. Velocity can then be used to estimate acceleration and jerk. As discussed in the preceeding sections, the resultant features are vectors of variable lengths depending upon the time taken by the subject to complete a task. In monst cases, these vector values cannot be used directly to train a machine learning classifier and therefore, several statistical features are computed from these, such as the number of changes of velocity or acceleration (NCV/NCA). These kinematic features are used to characterize the fine motor skills and handwriting fluency of an individual, that can be affected by an underlying impairement like PD. Kinematic features are highly dependent on the task being employed and can result in poor performance if extracted from a non-relevant task [69, 104].

2.5.4 Pen and Pressure Features

In an attempt to evaluate different modalities, pressure signals have also been employed in several studies [58, 61, 75–77, 83, 104, 105]. Two types of pressure units are mostly evaluated i.e. pen pressure exerted on the surface while writing [104] and the grip of the fingers on the writing instrument (usually a smart pen) [105]. Statistics like average pressure over on-surface strokes, number of changes of pressure (NCP) and relative NCP, are computed and used in combination with other dynamic features to enhance the performance of the system. In a similar attempt, pressure signals captured by a smart pen have been transformed into 2D images to extract visual features from them [105]. Other pen-based features include the angle and the altitude between the pen tip and the writing surface. These are recorded as *azimuth* and *tilt* values by means of a digitizer tablet with a smart pen [61]. Although pen and pressure based features demonstrate

potential in a combined approach, their effectiveness in isolation is still debatable and requires further investigation.

2.5.5 Temporal Features

Time-based assessments have been employed in several studies [55, 58, 61, 76, 77, 83, 104, 161, 162], to indicate several cognitive impairments. Popular temporal features include on-surface stroke duration, in-air pauses and total time required to complete a task. Amongst the given temporal measures, in-air time has shown promising results especially in identification of cognitive impairments resulting from neurodegenerative diseases like PD and AD. Despite the promising results presented in studies like [161, 162], temporal features may not prove effective for differential diagnosis of diseases with overlapping conditions. For instance, time delay in PD can be due to motor impairments while in AD and MCI, this can be attributed to deterioration of memory or executive planning.

2.5.6 Non-linear Dynamic Features

Signal-based features like *entropy* and *energy* can capture randomness and irregularities in a given signal. In a handwriting signal, these irregularities can highlight impairements of fine movements, which are otherwise difficult to analyze using only the kinematic features. Based on this assumption, studies like [71, 76, 77, 82, 83, 104] have extracted several non-linear dynamic features from handwriting/drawing signals captured by a smart pen. These include Shannon and Rényi entropy [178], signal-to-noise ratio (SNR) and empirical mode decomposition (EMD). Enery-based features are also computed to observe fluctuations in the continuous time series based handwriting. A popular example is the Teager-Kaiser energy [179]. As with all other dynamic feature vectors, some statistical functions of the feature vectors are computed before classifying.

2.5.7 Neuromotor Features

Stroke velocity profiles have also be employed to extract the intrinsic properties of the neuromuscular system and the control strategy of an individual performing a graphomotor task. Inspired by the Kinematic Theory of Rapid Human Movement [172], different parameters that characterize a stroke velocity profile, are computed. A popular approach is called the *Sigma-Lognormal model* [180], that decomposes the stroke velocity signals with lognormal shape. Another technique employed is the *Delta-lognormal model* [181]. These features have been adopted by studies like [174, 176, 177] to discriminate between normal and abnormal age related changes in handwriting. Although lognormal models of neuromotor abilities of an individual have extensively been evaluated in the domain of online signature analysis, nonetheless, their potential to characterize graphomotor deformations for neurological/neuropsychological assessment is still in its infancy. Further analysis is required to establish its effectiveness.

To assess the effectiveness of the proposed deformation representations, several approaches have been employed in the literature. These include template matching [43, 76], heuristics [44, 45,

51, 52, 81], statistical analysis [82, 104] and machine learning [48–50, 55, 56, 58, 61, 69, 71, 76, 78, 82, 83, 104, 160–162, 170]. Table 2.11 summarizes the popular estimation techniques discussed in the literature.

Table 2.11: Summary of Deformation Estimation Methods Employed In Related Studies

Estimation Method	Study	Techniques
Template Matching	[43, 76]	Pixel-wise distance
Heuristics	[44, 45, 51, 52, 81]	Ontology-based, fuzzy logic-based, genetic programming-based
Statistical Analysis	[82, 104]	Mann-Whitney U test, t- test, Pearson's correlation, Spearman's correlation
Machine Learning	[48–50, 55, 56, 58, 61, 69, 71, 76, 78, 82, 83, 104, 160–162, 170]	SVM with (rbf/linear) kernel, NB, K-NN, LDA, RF, linear regression, OPF, AdaBoost, XGBoost, CART, DT

Template matching techniques prove beneficial in estimating primitive component-level deformations. Pixel-wise distance is computed between the drawn component and the expected stimulus. Although some techniques [43, 76] in literature have employed a template matching based approach, nevertheless, such approaches have limited applicability in this domain due to the unconstrained nature of the free hand drawings. Heuristic based techniques have also been employed in the literature [44, 45, 51, 52, 81]. Their prime objective is to provide explainable solutions for the target users i.e. the clinical practitioners. Due to the inherent superiority of fuzzy logic over a strict rule-based approach, it is a preferred design strategy in this domain. However, despite being advantageous in modeling clinical manifestations, heuristics are difficult to design and lack scalablity.

Due to the limitations of template matching and heuristic-based approaches, recent works in the literature are attempting to employ statistical [82, 104] and machine learning algorithms [48–50, 55, 56, 58, 61, 69, 71, 76, 78, 82, 83, 104, 160–162, 170, 174] to model deformations. Statistical methods are mostly applied to determine the strength of association between the computed features and the clinical scores. These methods determine the effectiveness of a feature to represent the target deformation. One of the popularly employed methods is the Pearson's correlation test [182]. Since different features are combined in a high-dimensional feature set, to reduce issues of cardinality some form of statistical ranking is performed. Spearman's correlation test [183] is employed in most cases to select the most significant features for the final model training. Depending upon the distribution of the feature set, methods like t-tests [184] and Mann-Whitney U tests [185] are also employed.

The most popular approach in the literature for assessing the predictive potential of the proposed feature representations, is by mean of machine learning algorithms. Graphomotor-based analysis can be addressed as a *regression* [161, 162, 174] or a *classification* [48–50, 55, 56, 58, 61, 69, 71, 76, 78, 82, 83, 104, 160, 170] problem. Popularly employed classifiers include Support Vector Machines (SVM) [186, 187] with linear/rbf kernels, K-Nearest Neighbour (K-NN) [188],

Decision Trees (DT) [189], Random Forests (RF) [190], Optimum-Path Forest (OPF) [191], Linear Discriminant Analysis (LDA) [192]. Ensemble approaches like Boosting (AdaBoost [193] and XGBoost [194]) and Bagging (Classification and Regression Tree (CART)) [195] have also been employed. An interesting aspect revealed from the literature suggests that no single classifier alone performs best in all scenarios. Due to this reason, most studies [75, 78, 83] evaluate the performance of their proposed features on several classifiers before selecting the best.

2.6 Critical Analysis and Research Gaps

A critical review of the literature presented in the previous sections reveals that defining representations for graphomotor deformations is a highly challenging task from the perspective of computerized analysis. The difficulty of finding an effective representation for deformations reflects on the performance of the estimation method and ultimately, on the outcome of the analysis. Both static and dynamic features can be extracted to represent visuo-motor and visuo-perceptual deformations resulting due to some cognitive disorder. In an attempt to do so, several features have been proposed from both offline (scanned images of paper-based responses) and online (acquired by digitizer tablet) samples. A summary of limitations and challenges in these techniques is outlined in the following.

- Online (dynamic) features which are employed in procedural analysis carry rich information however, as mentioned previously, they not only require specialized devices but also result in modification of the standard test protocols. This, in some cases, may not be acceptable by the practitioners. There is a need to propose solutions that employ conventional templates and standard acquisition modalities so that the gap between clinical practices and intelligent solutions can be bridged.

- Visual analysis based techniques that employ primitive components of a drawing for analysis are marked by challenges of localization and segmentation. It is common to address these challenges by restricting the subject to provide data on pre-printed templates which can limit the expressiveness of the individual in free-hand drawing. This advocates the development of holistic approaches which can trace local deformations without segmentation into components.

- Deformation analysis with template matching or statistical features is characterized by a large set of heuristics. Since the deviation of the graphomotor impression from expected prototype can exhibit large variation, robust deformation representations must be investigated which are able to generalize well to unseen examples.

- The problem of deformation modeling is different from the conventional classification framework where solutions are sought which minimize the intra-shape variations and maximize the inter-shape variations. In the problem at hand, the same deformation is required to be

measured across multiple shape classes and there is a need to develop methods which are able to learn the deviation rather than the shape.

The insufficiency of conventional features - deformation estimation advocate the investigation of machine learned features for this problem. Machine learned and more specifically deep learned features have emerged as a robust alternative to classic hand-crafted approaches. However, their applicability in the domain of graphomotor deformation representation and estimation has not been thoroughly investigated. Such features, though not always intuitive to correlate/or interpret, they may provide useful information regarding the intrinsic properties of the writer/drawer.

2.7 Summary

This chapter presented a detailed literature review of the various trends and approaches employed for the analysis of graphomotor tasks for the purposes of deformation estimation and disease diagnosis. A contextual categorization of the relevant works based on the mode of analysis is presented. Popularly employed features and estimation methods have also been discussed. Critical gaps in the state-of-the-art are outlined, that this thesis intends to address.

Chapter 3

Deformation Representation and Estimation Using Convolutional Neural Networks

3.1 Introduction

Typical steps in computerized analysis of neuropsychological graphomotor tasks include *Data Acquisition*, *Pre-processing*, *Feature Extraction* and *Analysis*. Based on the method of acquisition employed, both offline and online samples can be obtained. The offline samples are digitized and may require pre-processing tasks like noise removal, binarization or segmentation. On the contrary, online samples are mostly represented by functional measures like Cartesian coordinates and time stamps, and require plotting or normalization before application. Once the data is prepared, various features (*static* or *dynamic*) can be extracted to characterize deformations. These features are then assessed to estimate the extent of deviation(s) from a target response. Based on the outcome of assessment, a decision regarding analysis is made. From the literature review it is evident that the most critical step in the computerized analysis of neuropsychological graphomotor tasks is the identification of those discriminating features that can effectively characterize clinical deformations and thus enhance learning. A number of hand-crafted features extracted from offline [43–45, 49, 50, 105] and online [58, 60, 69–71, 75, 78, 81, 104, 161] samples are discussed in the literature, which are then used to train a machine learning-based classifier (e.g. SVM, LDA, DT etc.). However, the potential of automatically learned features is relatively less investigated in this domain.

Deep learning-based methods [196], have recently gained immense popularity in the domain of feature representation and classification. Convolutional Neural Networks (CNNs) [197–199], a branch of deep learning, has urged researchers to revisit many popular computer vision and pattern recognition problems including 'Writer identification' [200] and 'Sketch recognition' [201, 202]. In this chapter, we propose a novel formulation of the problem of neuropsychological graphomotor-based deformation representation by employing Convolutional Neural Networks

(CNNs) as feature extractors. The basic motivation behind the use of CNNs is that they can extract discriminating features from various regions of the complete drawing without localization or segmentation. Convolutional filters employed on different layers of a CNN produce activations for specific discriminating patterns. CNNs can generalize features across a wide variety of deviations and thus the need to explicitly define heuristics for each deviation is no longer necessary. In this regard, CNN-based features can overcome the insufficiency of hand-crafted features and also provide robustness.

CNNs can be trained to extract both intrinsic and extrinsic information from a drawn/handwritten sample, thus providing a mean to effectively analyze various aspects (motor or perceptual) of the same sample. Finally, CNNs provide a strong visual analysis based alternative that can be utilized to assess existing offline samples collected by the clinical experts. In this way, the proposed approach can resolve the need for any modifications in the original test conduction and sample acquisition protocol.

Nonetheless, despite the apparent advantages of deep learning-based methods for feature representation, their direct application in the domain of neuropsychological graphomotor task analysis, is challenging due to several reasons. One of the key challenges is the scarcity of training data that is a common scenario in this domain (commonly around 20-60 samples). Furthermore, in a conventional setting most CNNs are designed as shape recognizers (as shown in Figure 3.1) and achieve this by minimizing the intra-shape class variations and by enhancing the inter-shape class variations. However, we expect the CNN model to learn deformation-specific intra-shape variations (Figure 3.2), as well as generalize deformation-specific similarities across various shapes (i.e. same deformation across different tasks).

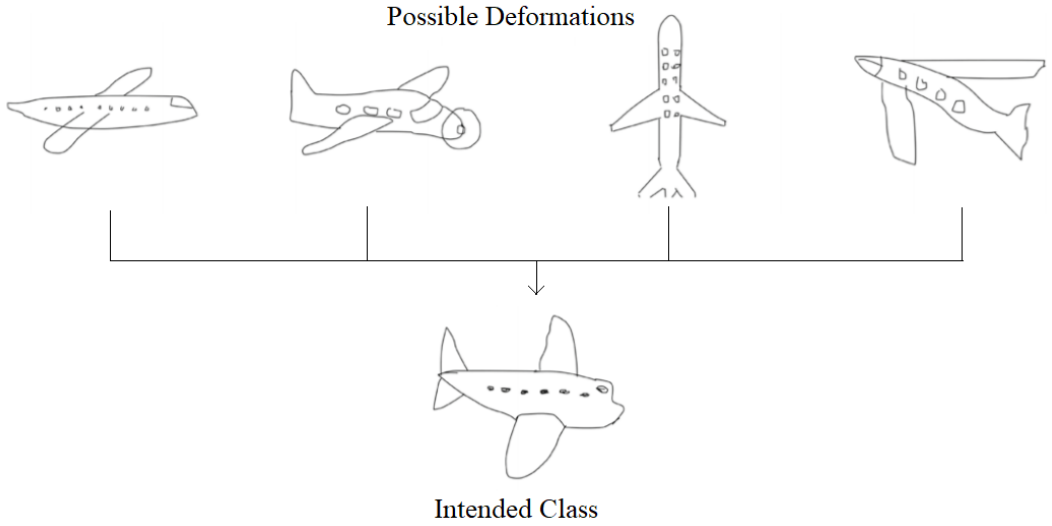


Figure 3.1: Shape recognizers intend to diminish intra-shape variations [201]

In this thesis, we have addressed these key challenges and consequently employed CNNs to characterize various motor and perceptual deformations from both offline, as well as online

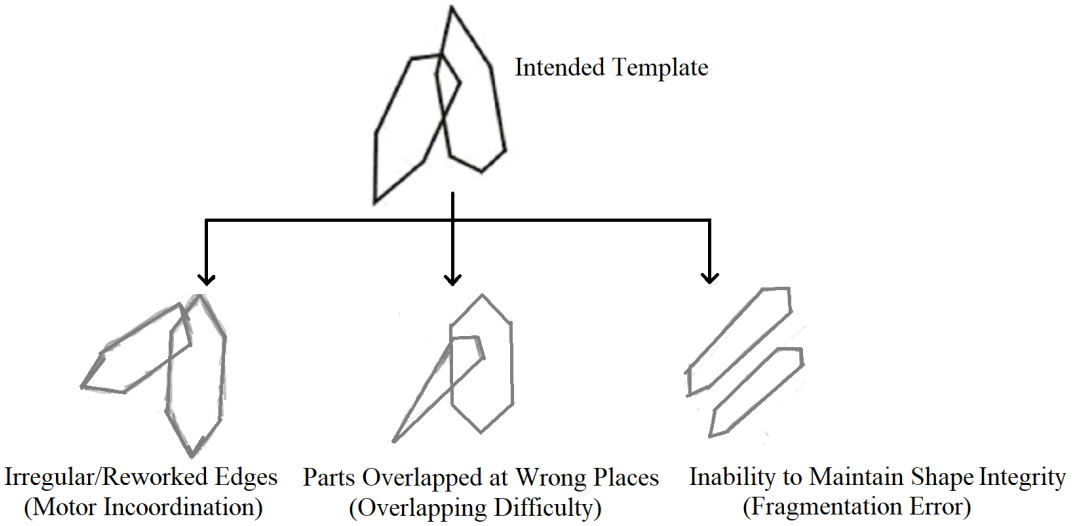


Figure 3.2: Our proposed methodology intends to identify deformation-specific intra-shape variations (three deformations scored during manual assessment of a standard BGT test)

graphomotor samples. In the forthcoming sections of this chapter, we discuss the proposed approach in detail. Section 3.2 provides the relevant theoretical background. Section 3.3 describes the main steps of the proposed graphomotor-based deformation modeling and classification scheme. Lastly, Section 3.4 summarizes the chapter.

3.2 Theoretical Background

A Convolutional Neural Network (popularly known as ConvNet) is inspired by the natural visual perception mechanism of the living creatures and can obtain effective representations of image directly from raw pixels with little-to-none preprocessing. It is a non-linear model capable of learning non-linear features [197]. A typical CNN architecture primarily comprises a series of convolutional layers followed by some intermediate layers (known as the convolutional base) [199], as shown in Figure 3.3. The connectivity between a pair of consecutive layers in a convolutional base is designed to facilitate the detection of distinctive local patterns in an input image. These patterns can then be employed for various visual classification tasks including the one being addressed in this thesis. To effectively employ CNNs for the problem under consideration, we must first understand the processing of the fundamental constituent layers.

3.2.1 Convolutional Base

The convolutional base consists of the three basic components of a CNN model i.e. convolutional layer, activation function and pooling layer. The first basic constituent i.e. the convolutional layer, is designed to take a tensor as an input and transform it into feature maps. A feature map is obtained by applying convolution with a matrix of weights (or kernel), followed by the addition of bias. Each convolutional layer has a number of kernels that replicate the process to generate a set of feature

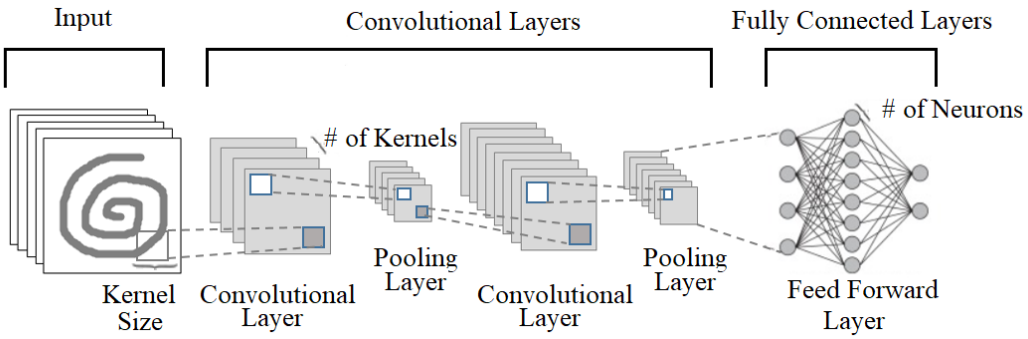


Figure 3.3: Typical structure of a sequential ConvNet

maps as an output. The processing of a particular convolutional layer l in a deep CNN consisting of L layers can be generalized by Equation 3.1, where $l = 1, ..., L$.

$$X_l^{(m)} = f\left(\sum_{n=1}^{N} W_l^{(n,m)} * X_{l-1}^{(n)} + B_l^{(m)}\right) \tag{3.1}$$

$X_{l-1}^{(n)}$ represents one of the N input feature maps from the previous layer, while $X_l^{(m)}$ represents an output feature map of the current layer, where $m = 1, ..., M$. Each input feature map is convolved with the kernels $W_l^{(n,m)}$ of the layer l, resulting in a sum of N two-dimensional convolutions denoted by $\sum_{n=1}^{N} W_l^{(n,m)} * X_{l-1}^{(n)}$. Each feature map has its own bias which is presented by $B_l^{(m)}$.

A subsequent feature map is obtained by applying convolution on the input with a learned kernel followed by the application of an element-wise nonlinear activation function. The activation function introduces the desired non-linearity to a multi-layered CNN network. Three basic activation functions commonly employed in a CNN model are *sigmoid*, *tanh* and *ReLU*. In most cases, the *Rectified Linear Unit (ReLU)* [203], is the preferred activation function. ReLU activation preserves all the positive values from the resultant feature map and eliminates the negative ones by converting them to zeros. This enhances the mapping capability of a ConvNet for the desired attributes. Studies like [204] suggest that ReLU allows efficient training of deep networks as compared to sigmoid or tanh. Let $f(.)$ represent an activation function that is applied on the result x, then ReLU can be expressed as Equation 3.2.

$$f(x) = max(x, 0) \tag{3.2}$$

A convolutional layer is paramterized by a large number of features. In order to reduce the resolution of the resultant feature maps, a sub-sampling (or pooling) layer is commonly introduced before the next convolutional layers. The feature maps of a pooling layer are connected to the corresponding feature maps of the preceding convolutional layer. The pooling layer applies a function (typically *Average-pool* or *Max-pool*) on the input maps to select a subset of the values from the local sub-regions. It also allows shift-invariance, furthermore, by stacking several convolutional and pooling layers, higher-level feature representations can be extracted.

The convolutional, ReLU and pooling layers together serve as a feature extraction convolutional

base. Deep ConvNets exploit the combined benefits of the regional convolutions and a layered hierarchy to learn effective representations for specific visual recognition tasks. The depth of a CNN architecture depends on the level of abstraction required. The intermediate convolutional layers are designed to extract distributed representations from the input data. After several convolutional and pooling layers, there exist one or more fully-connected layers similar to a conventional 'Multi-layer Perceptron (MLP)'. A fully connected layer employs all the features extracted by the previous layers in order to amplify the selected distinctive patterns. The features extracted from a fully connected layer are then employed for classification.

3.2.2 CNN Architectures

As the popularity of ConvNets increased, several architectures were proposed. These include series models like AlexNet [199] and VGGNet (16 and 19) [205], Inception models like GoogLeNet [206] and InceptionV3 [207], and residual models like ResNet (50 and 101) [208]. AlexNet, VGG16 and VGG19 architectures consist of a series of alternating blocks of convolutional (with ReLU activation) and pooling layers. Each of these has three fully connected layers, out of which, the last layer is used for the classification purposes. The Inception architectures introduced the concept of *Inception blocks*, in which different convolutional filters and pooling layers are concatenated to enhance learning. ResNet architectures (ResNet50 and ResNet101) represent the residual networks, where the layers contain direct, additive connections referred to as *Skip connections* to the next layers.

Several variations of the existing architectures have also been proposed. For instance, SqueezeNet [209], is a CNN micro-architecture with compressed filter sizes. It attempts to optimize learning without compromising on performance and reports an accuracy comparable to that of AlexNet. The architecture employs *Fire modules* comprising of a *squeezed* convolutional layer (with 1x1 filters) feeding into an *expanded* layer (with a combination of 1x1 and 3x3 filters). The fire modules allow network learning with a comparatively lesser number of parameters than the AlexNet architecture. Similarly, the idea of skip connections has been extended to connect the subsequent blocks of densely connected layers in DenseNet architectures. The dense blocks alternately comprise of 1x1 convolutional filters and max-pooling layers in order to reduce the number of tunable parameters. Contrary to the skip connections in the residual networks, the output of the dense blocks are not added but instead concatenated.

A typical trend that can be seen from the evolution of the ConvNet architectures is the increase in depth. For instance, ResNet is approximately 20x deeper than AlexNet and 8x deeper than VGGNet. The basic notion behind the increased depth is to increase non-linearity to enhance approximation of the target function and thus, achieve improved feature representations. Nonetheless, increasing depth can also increase the complexity of the network making it more difficult to optimize and prone to overfitting. Therefore, selecting a ConvNet architecture for feature representation is a crucial decision.

3.2.3 Transfer Learning

A forward run in a CNN model is employed for the prediction of probabilities, nevertheless, before a CNN model is ready to predict, it requires an extensive training. The training process involves running the CNN network in both directions i.e. forward (*Feed Forward*) and backward (*Back Propagation*) [197], as shown in Figure 3.4. Let us consider a forward run first, where N number

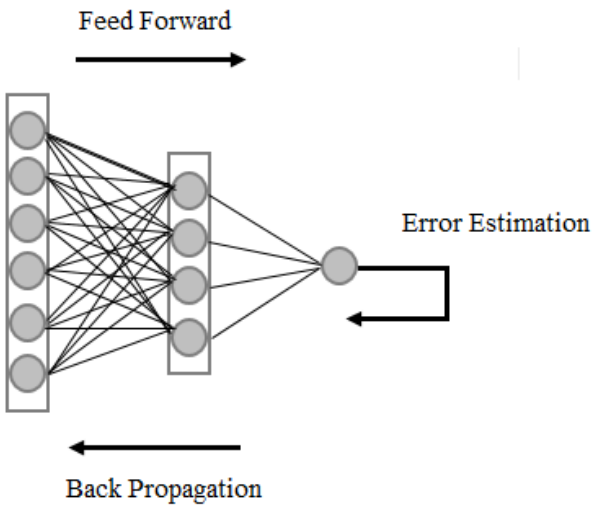


Figure 3.4: Network learning in forward and backward direction

of training samples are employed. For a given training sample (x_n, y_n), where $n \in [1, ..., N]$, (x_n denotes the input and y_n denotes its corresponding label), the network produces an output o_n. This o_n is then used to estimate the error and compute the loss z, i.e. the difference between the output o_n and target ground-truth y_n. The loss is then used to update the network parameters by employing back propagation. The objective is to optimize parameter learning by minimizing the loss. To achieve this objective the training process is repeated several times until a minimum value for loss is achieved.

Training a CNN model relies on the tuning of a large number of parameters, which is only possible in the case of sufficiently large amount of annotated data. As discussed earlier, the lack of sufficient training samples has been one of the major limiting factors in the use of CNNs for neuropsychological drawing analysis. Nevertheless, recent advances in machine learning have proposed several solutions to overcome this limitation. Transfer learning [210] is one such alternative, which allows a CNN architecture to transfer learned weights across different source and target datasets.

Leveraging on the idea that the convolutional base of a CNN model learns hierarchical feature representations, it is common to train a CNN on a different task with a larger dataset (e.g. ImageNet [86] etc.) and exploit the learned features to characterize a smaller dataset. A common practice in transfer learning is to employ a pre-trained model and continue back propagation with the target dataset either on all or a subset of layers, concept known as *fine-tuning*. Since the initial layers of a CNN are known to learn low level, task-independent features, it is common to freeze

their weights and tune the subsequent convolutional and fully connected layers on the task-specific dataset. Another popular approach is to employ a pre-trained ConvNet as a *feature extractor* only where the convolutional base of a trained model is used to convert the samples under study into robust feature representations (learned on another lager dataset). The extracted features can then be fed to train a separate classifier to learn to discriminate between the classes under study. Exploiting pre-trained CNNs not only reduces the time and computational complexities, it also allows use of deep CNNs for problems with smaller datasets making it an ideal solution for scenarios like ours.

3.3 Proposed Deformation Modeling Using Pre-Trained ConvNets

Based on the theoretical support of employing a pre-trained ConvNet for deformation-specific feature representation, we have designed a framework to empirically assess its effectiveness in the domain of neuropsychological graphomotor task analysis, where specific intra-class variances and inter-class similarities need to be modeled to determine the presence of a cognitive impairment. Figure 3.5 presents the schematic pipeline of the proposed method for deformation modeling and classification. Each step is explained in the subsequent subsections.

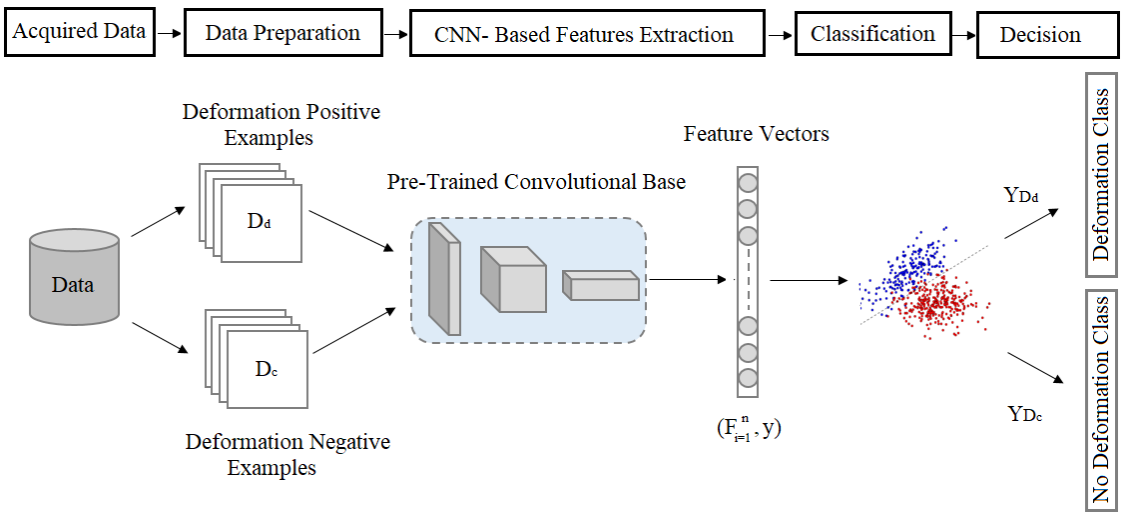


Figure 3.5: Overview of proposed methodology for deformation modeling and classification

.

3.3.1 Data Preparation

Data preparation is an essential preliminary step in any pattern recognition and computer vision task. Depending on the mode of sample acquisition (offline or online), the data is first converted into an appropriate representation for the convolutional base. CNNs generally take image data in the form of a tensor (most commonly 4D array) containing the size of the image (in terms of height, width and depth) and input batch size. Depth corresponds to the type of the image (i.e. grayscale or RGB). Most ConvNets process three channel images, therefore, input data, if binary or

grayscale, is first extended to three channels. Batch size can vary based on the size of the training data, however, image size (height and width) must correspond to the size of the input layer of the model being employed. This is usually achieved by means of scaling or resizing, nonetheless, care must be taken to mitigate the loss of important information. An important assumption in our proposed methodology is that a single task is being analyzed at a time. However, in most real life cases, a graphomotor test comprises multiple tasks on a single sheet of paper (e.g. BGT), therefore, segmentation of individual tasks is required prior to input, in such scenarios.

3.3.2 Feature Extraction Using Pre-Trained ConvNets

As discussed in the previous section, training a deep CNN model from scratch requires considerably large amount of training examples which is not possible in the current scenario. Consequently, a transfer learning based approach is adopted, where ConvNets pre-trained on ImageNet are employed as feature extractors. ImageNet [86] is a very large dataset consisting of approximately 1.2 million annotated images belonging to 1000 categories. Although, the target images (handwritten characters and hand drawn shapes), in the problem under consideration are different from the source dataset, nonetheless, ConvNets pre-trained on such huge collections of images are able to extract robust feature representations that could prove effective in characterization of relatively less complex hand drawn shapes or handwriting. Furthermore, in our case the size of the target dataset is relatively very small as compared to feature dimension, hence fine-tuning of ConvNets may result in overfitting. Consequently, we opt to freeze the entire convolutional base of the pre-trained models and employ them only as feature extractors. Samples under study are passed through the trained models and the output of the convolutional base is extracted as the feature representation of the input sample (as shown in Figure 3.6). These features are subsequently fed to a classifier to learn the discrimination between classes of interest.

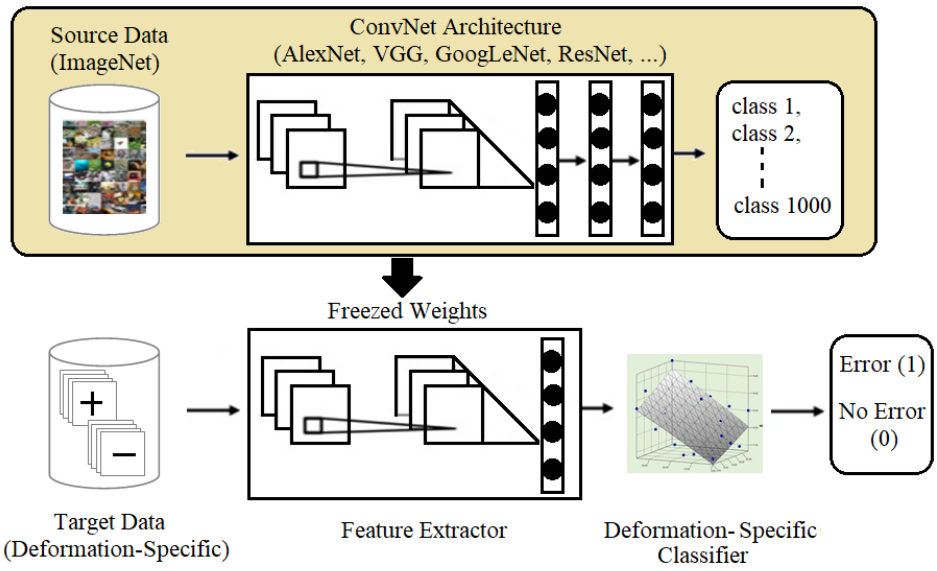


Figure 3.6: Transfer learning from source data to target data

3.3.3 Feature Representation for Deformed and Non-Deformed Data

While extracting features, we have considered deformed and non-deformed data as two separate classes, same formulation is considered for diseased and healthy samples. For instance, let D_d be a dataset consisting of samples from a specific deformation class Y_{D_d}, while D_c is a dataset consisting of all samples without the particular deformation Y_{D_c}. The objective is to project graphomotor samples X onto a representation space $\phi(X)$ where deformed and non-deformed samples are better separated. The intuition is that in order to classify between erroneous and non-erroneous samples, the ConvNet will extract visual cues that are particular to each class. Formally, we consider a training set composed of representatives from both classes. To extract representations, we perform feedforward propagation with the training tuples (x, y) where x is the input image, and y is the class label $(y \in Y_{D_d} \vee y \in Y_{D_c})$, until the last layer before softmax. The activations at that layer are used as a n-dimensional feature vector $(F_{i=1}^{n}, y)$ for the given training image x. Similarly, labeled feature vectors for all images are obtained.

In a real application scenario, deformed samples are comparatively harder to get than non-deformed ones. To address this issue, we employ augmentation, whenever required. Several techniques have been proposed in the literature to augment images [211]. Some techniques (e.g. Geometric transformations) augment images in the data space, while others (e.g. Synthetic Minority Over-Sampling Technique (SMOTE) [212]) augment data in the feature space. Recently, Generative Adversarial Networks (GANs) [213] have also gained popularity for generating augmented data. In a domain-specific problem like clinical deformation modeling, the preservation of class labels after augmentation is a critical concern. In most cases, the augmented data also requires to be labeled by the domain-expert. This indicates a need for human intervention and control while generating near-realistic examples especially for a specific deformation class. Based on this notion, we have employed various image processing based *generic* and *deformation-specific* procedures on the existing samples to generate relevant examples. Post-augmentation cleaning is then employed to ensure class label preservation. As mentioned earlier, the main objective of data augmentation in the present scenario is to ensure adequate representation of each class by complementing the original samples rather than producing extensive synthetic data that might not be able to capture properties of actual real life examples.

3.3.4 Deformation Classification

Feature vectors extracted by the pre-trained ConvNet model from both deformed and non-deformed classes are then used to train a deformation-specific binary classifier. A training set consisting of N tuples of form $(F_{i=1}^{n}, y)$ is created, where N_p tuples with $y \in Y_{D_d}$ are considered as positive examples and N_N tuples with $y \in Y_{D_c}$ are considered as negative samples. For each type of deformation, a separate classifier is trained. During the testing phase, a disjoint set of positive and negative samples of the particular deformation is employed that is not used for training. Features extracted from a test image are fed to the specific classifier to determine the presence or absence of that particular error (deformation). The classifier then outputs the probabilities for the input feature vector to belong

to either of the two classes i.e. Y_{D_d} or Y_{D_c}. This approach has several benefits, some of which are outlined below.

- Same formulation can be used to discriminate between samples of healthy and diseased subjects.
- Same deformation can be identified across multiple tasks, especially if task-wise examples of that deformation are scarce. In such a case, deformation-specific generic features extracted from other tasks can be employed to facilitate classification.
- In case multiple deformations exist in a single image, then each deformation can be identified by testing the same image by separate deformation-specific classifiers. This is a significant contribution of this thesis, as it provides a robust alternative to a heuristic based analysis, while providing similar explainability to the end users.

3.3.5 Application to Visual-Motor and Visual-Perceptual Deformation Identification

In order to assess the efficacy of our proposed method in identifying different graphomotor-based deformations, we have considered two predominantly analyzed deformations during drawing and handwriting analysis. These include the *visual-motor* and *visual-perceptual* deformations. Visual-motor deformations are mostly associated with Parkinsonism, while visual-perceptual deformations can indicate a wide range of underlying dysfunctions including dementia, brain injuries and developmental issues. Literature review suggests two scenarios in which these deformations are most frequently assessed and thus are considered as our case-studies as well. These two scenarios are:

- Identification of visual-motor deformations in the graphomotor samples of Parkinson's patients
- Identification of visual-perceptual deformations in the graphomotor samples of school children (with or without learning difficulties)

Visual-motor deformations like tremor [36], micrographia [35, 93], and bradykinesia [97, 98] are commonly observed in graphomotor tasks performed by individuals with neurological diseases like Parkinson's disease [103, 214, 215]. Characterization of visual-motor deformations in graphomotor samples of patients suffering from (or at-risk of onset of) such diseases can facilitate early detection and disease progression monitoring [91]. It can also be employed to assess the efficacy of the treatment. As discussed in the introductory chapters, due to the insufficiency of existing visual analysis based techniques to characterize visual-motor deformations, researchers have shifted their focus to procedural analysis based methods that require modifications in conventional test conduction and analysis protocols [71, 75]. If by applying the proposed visual analysis based method, we can produce comparable results, it would be a worthwhile contribution. To identify visual-motor deformations (more specifically *tremor* and *micrographia*) in the graphomotor samples

of PD patients, we have employed some conventional and non-conventional templates comprising of simple drawings (Archimedean spiral) and handwriting tasks (repetitive letters, bigrams, long words and sentences). Task-wise analysis is also performed to assess the impact of tasks on the effectiveness of the proposed formulation. Details of the application of our proposed approach in the identification of visual-motor deformations in graphomotor samples of PD patients are discussed in Chapter 4.

The second scenario in which we have applied our proposed formulation is the identification of visual-perceptual deformations (like rotation, perseveration, simplification etc.) in drawing samples of children. As discussed earlier, children's drawings can give useful insight regarding their emotional state and perceptual maturity [165]. Both these factors can impact the social skills and the academic performance of children. Poor performance in drawing tasks can indicate underlying developmental issues resulting in learning difficulties [166]. By identifying the extent of visual-perceptual deformations in the drawing samples of children, timely intervention can be provided. Nevertheless, despite its advantages, computational modeling of human perception is a challenging task and usually requires an extensive heuristic-based approach to overcome the insufficiency of existing visual features. As mentioned earlier, CNNs are inspired by the human visual perception mechanism and are known to outperform other computational models in characterizing perceptual data [216]. Consequently, our hypothesis is that CNN-based features can effectively represent perceptual deformations. To assess our hypothesis, we have employed the samples of a popular neuropsychological test called the *Bender Gestalt Test (BGT)* [22], that comprises nine geometric shapes. Each shape is carefully selected to assess eleven visual-perceptual deformations based on the Lacks' scoring system [31]. Previously, such a large number of visual-perceptual deformations have not been analysed in the literature. Details of the application of our proposed approach in the identification of Lacks' visual-perceptual deformations in BGT samples of children are discussed in Chapter 5.

3.4 Summary

In this chapter, we introduced our research hypothesis that features extracted from pre-trained ConvNets can provide an effective and robust representation of graphomotor-based deformations, despite the scarcity of data. This is in contrast to the conventional exploitation of CNNs for classification tasks where features enhancing the inter-class variations are sought. We presented the theoretical support to our hypothesis and described the proposed formulation in detail. We then discussed the application of the proposed scheme in two scenarios i.e. to characterize visual-motor and visual-perceptual deformations, to assess its performance. Due to the different nature of deformations, both scenarios require specific methodologies, nonetheless, the basic pipeline remains generic. The forthcoming chapters describe the application methodologies and analyze the findings.

Chapter 4

Identification of Visual-Motor Deformations - An Application to Detection of Parkinson's Disease

4.1 Introduction

In this chapter, we present the detailed application of our proposed deformation modeling scheme for the detection of Parkinson's disease (PD) from the graphomotor samples of patients. Computerized handwriting analysis for the screening of PD is a well established research problem and has attracted the attention of a significant proportion of the PR research community, as evident from the literature review. For instance, studies including [49, 50, 61, 69, 70, 75, 78, 79, 81, 82, 104, 105, 160] have targeted the identification of effective visual or procedural attributes to discriminate between samples of PD patients and healthy subjects. In an attempt to promote further exploration in this direction, some of these studies have introduced publicly available datasets (like PaHaW [75] and HandPD [49]) as well, a practice that is usually scarce in this domain. Motivated by the interest of the relevant research community in this area, we too have selected it as a case-study to assess the effectiveness of our proposed approach. Furthermore, the availability of benchmark datasets has provided us an opportunity to compare the performance of our proposed scheme with the state-of-the-art, which is otherwise difficult in this domain.

We present the technical details of the proposed methodology for the identification of visual-motor deformations in the graphomotor samples of PD patients in Section 4.2. Details of the dataset employed for algorithmic development and evaluation are also presented in the same section. The experimental protocol is described in Section 4.3, while the results along with their detailed analysis are provided in Section 4.4. Finally, Section 4.5 summarizes the chapter.

4.2 Proposed Methodology

As discussed in the previous chapter, due to the inherent differences in the nature of deformations and the samples being employed in both case-studies (PD and BGT), specific methodologies are required for each, nonetheless, the basic pipeline remains generic. For the application of our proposed deformation estimation and classification approach in the current scenario, we have primarily selected the benchmark database, 'Parkinson's Disease Handwriting Database (PaHaW)' [75]. PaHaW is an established database and has been employed in a large number of studies [69–71, 75, 78, 81–84, 104] (discussed in the literature review), targeting the identification of PD. This allows us to compare the performance of our proposed scheme with the state-of-the-art. Another reason for selecting the PaHaW database is that it comprises a number of graphomotor samples (drawing, as well as handwriting) of PD patients, thus providing means to carry out in depth analysis of the impact of different tasks on the performance of CNN-based features. This is a vital analysis in this domain as studies have shown varying impact of tasks on the effectiveness of the proposed features.

Since our proposed scheme is visually analyzing the completed response, we expect it to discriminate between the samples of PD and healthy subjects based on the identification of two visual-motor deformations i.e. tremors and micrographia. As discussed earlier, tremors are irregularities resulting from involuntary muscular movements. As a consequence, PD samples are characterized by non-smooth, irregular formations as shown in Figure 4.1-b. Similarly, micrographia or abnormal reduction in size, is mainly attributed to impaired wrist control. As a result, a reduction in horizontal/vertical strokes is usually observed. In a spiral drawing, it often results in a failure to maintain the angle versus radial distance relation, thus small overlapping spirals with tightly bunched turns are observed in samples of PD patients (also visible in Figure 4.1-b). Visual feature extraction has always been a challenging aspect in graphomotor analysis due to limited representation. Due to this reason, we have hypothesized the use of CNN-based features for the characterization of discriminating visual attributes between healthy and diseased samples. Nonetheless, before explaining the proposed methodology, we will first elaborate on the database (PaHaW), in the subsequent section.

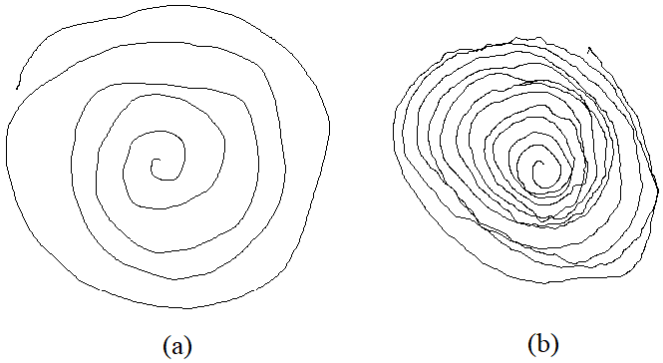


Figure 4.1: Spiral drawings from PaHaW database of (a) Healthy subject (b) PD patient

.

4.2.1 Parkinson's Disease Handwriting Database (PaHaW)

PaHaW dataset comprises samples collected from 75 subjects (37 PD patients and 38 (age and gender comparable) healthy controls (HC)). All participants are Czech citizens and were enrolled at the First Department of Neurology, Masaryk University and at the St. Anne's University Hospital, in Brno, Czech Republic, at the time of data acquisition. The subjects are predominantly right-handed and literates of Czech language. Prior to sample acquisition, PD patients were evaluated by a neurologist using UPDRS-Part V, (Modified Hoehn and Yahr staging score [217]), for the inclusion purposes. On the contrary, the HC group was examined to ascertain that there was no movement disorder that could affect handwriting. It is important to mention that the PD group attempted the test in their 'ON-state' (dopaminergic medication (L-Dopa)). Demographics and clinical details of test participants are summarized in Table 4.1. It contains the number of subjects in each group, gender, mean and standard deviations of ages, disease durations and levedopa equivalent dosages (LED). The table also contains an average UPDRS-Part V score of disease severity for the PD group. Three subjects (1 PD, 2 HC) did not complete all tasks, hence their samples were excluded during our evaluation, reducing the total number of samples to 72.

Table 4.1: Participants' Demographics and Pre-test Clinical Diagnosis

Gender	**Number**	**Age** (Years)	**Duration** (Years)	**UPDRS-V Score**	**L-Dopa Dosage** (mg/day)
Parkinson's Disease Patients					
Females	18	71.76±10.93	9.88±5.27	2.18±0.86	1146.03±543.89
Males	18	66.50±13.44	7.44±4.04	2.31±0.75	1673.38±616.66
All	36	69.21±11.10	8.70±4.82	2.24±0.80	1401.72±630.71
Healthy Controls					
Females	17	61.59±10.17	-	-	-
Males	19	63.32±13.14	-	-	-
All	36	62.50±11.70	-	-	-

All subjects were instructed to attempt 8 graphomotor tasks as shown in Figure 4.2. The template includes the conventional Archimedean spiral drawing alongwith seven handwriting based tasks, all written in Czech language. The handwriting tasks include simple cursive letters like repetitive *l*, a bigram *le*, and a trigram *les*. In addition to these, long words like *lektorka*, *porovnat*, and *nepopadnout* are also employed. The attribute of a simple orthography is the contributing reason for the selection of these words. All these tasks are without breaks and require long continuous strokes on the writing surface, that are more suitable to capture effects like tremor and micrographia. Lastly, a whole sentence task is also included (i.e. *Tramvaj dnes už nepo-jede*). By employing a long sentence comprising of multiple words, analysis of pen-up measurements during the transitions (breaks) between the words of the sentence, is made possible.

An Intuos 4M digitizing tablet (Wacom technology) was used to record signals of handwriting samples. The tablet was overlaid with a white template paper and a conventional ink pen was used.

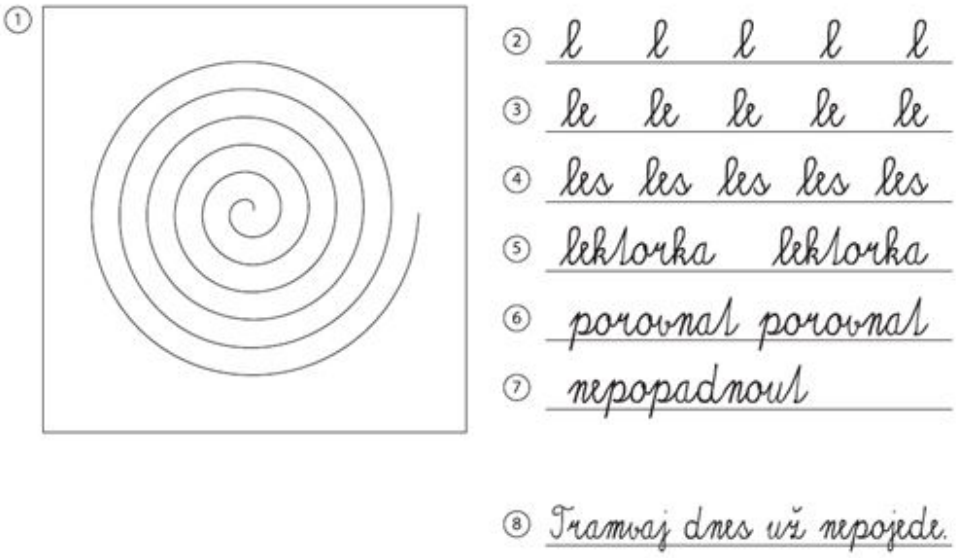


Figure 4.2: Graphomotor tasks employed in PaWaH database [75]

.

The raw data captured by the device includes the x- and y-coordinates of the pen position and their time stamps at a sampling rate of 150 Hz. The device also recorded the units of pen pressure exerted over the writing surface and angles (i.e., azimuth and altitude), representing the pen inclination. Another value recorded by the device is the button status, which is a binary variable depicting the pen-up and pen-down states. The status is 1 when the pen is on the surface, and 0 when the pen is not touching the surface (which allows the acquisition of 'in-air movements', if required). Figure 4.3 shows the recorded values of a sample task (i.e. Archimedean Spiral).

Y-coordinate	X-coordinate	time-stamp	button-state	azimuth	altitude	pressure
4161	2474	625877	1	3447	633	78
4164	2472	625884	1	3457	624	124
4166	2472	625892	1	3457	624	162
4167	2471	625899	1	3457	624	212
4168	2471	625907	1	3457	624	266
4169	2470	625914	1	3457	624	312
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Figure 4.3: Signal values captured by the device while performing Task 1 (i.e. Archimedean Spiral).

4.2.2 Data Preparation

The original PaHaW database does not include images, but separate files containing signal-based measurements (mentioned in the previous section) for each task performed by a subject (PD/HC). Since, we intend to exploit the visual features from these samples, the given information is used to reconstruct static images of the completed responses for each task. Reconstruction of offline samples from online signals is a common practice with datasets belonging to forensic writings [218–221] and same can be employed in case of health related writing samples. To reconstruct a visual image of the sample produced by the subject, x- and y-coordinates corresponding to all positions where the pen is touching the writing surface are rendered. The coordinates are then normalized, the x-coordinate is normalized to 0 (by subtracting the minimum value from every coordinate) while the y-coordinate is normalized by subtracting the mean from each value, and plotted. Since the sampling frequency is sufficiently high (i.e. 150 Hz), therefore connecting the plots of coordinates of pen trajectories produces near realistic drawing traces (D_r), as shown in Figure 4.4.

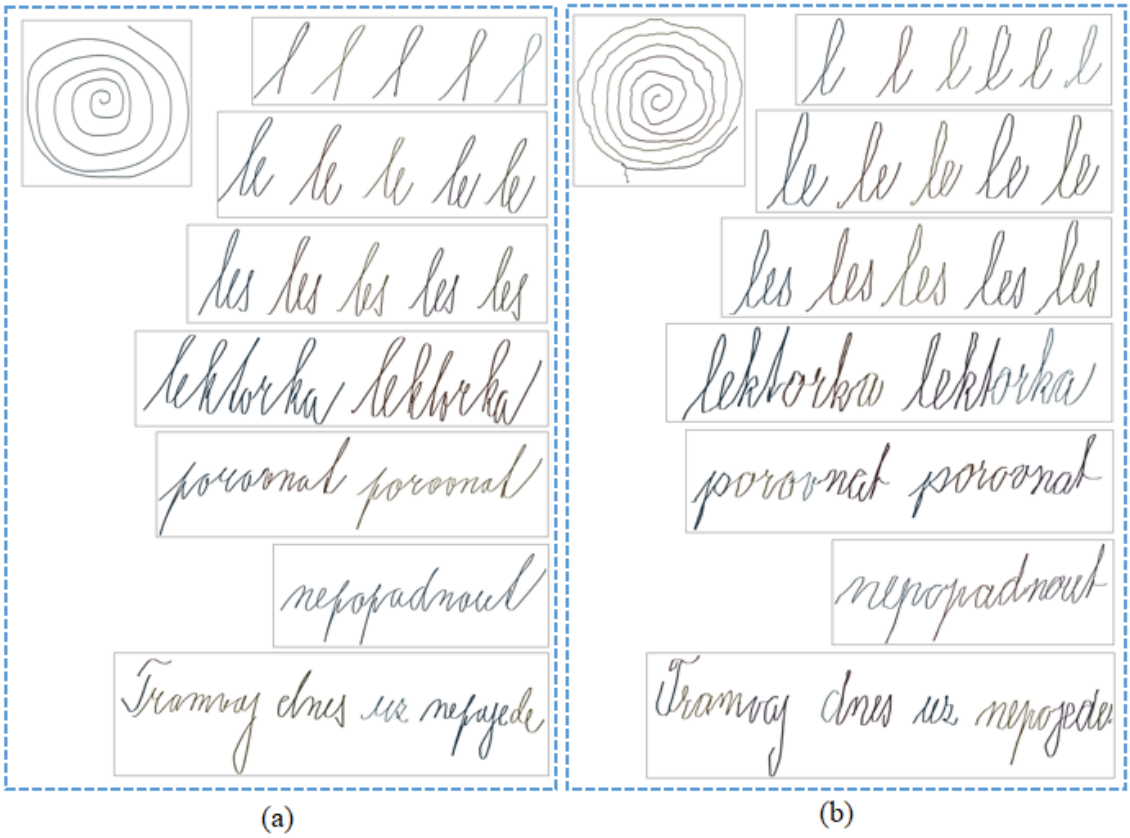


Figure 4.4: Reconstructed images of templates produced by (a) a Healthy subject and (b) a PD patient

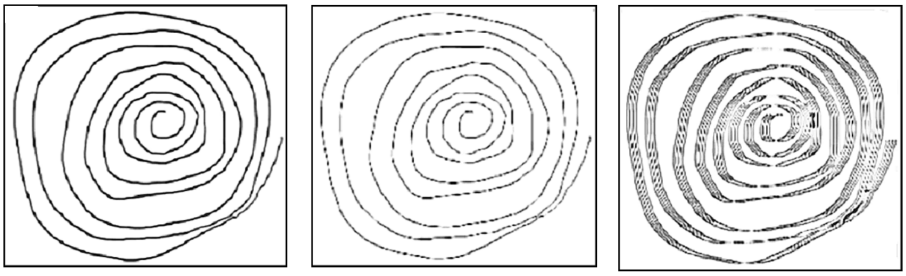
.

Once the offline images are generated, we employ these to generate additional representations. The prime objective of generating multiple representations of the raw input data is to enrich the

resultant feature set. Multiple representations of raw images have several benefits in our proposed technique. Some of these are listed below:

- Multiple representations enable us to highlight imperfections resulting from subtle muscular dysfunctions, like tiny jerks and other irregularities, captured by the device while the tasks are being performed.
- As discussed earlier, supervised machine learning methods require a large number training examples which are usually not available in this domain. Multiple representations can provide extra information regarding the input data that can be employed to enhance learning.
- Transforming raw data by applying different non-linear transformations can enrich representation of the data and consequently, enhances learning of rich features.

In light of the aforementioned benefits, we employ two representations in addition to raw data i.e. 'Median Filter Residual' and 'Edge Detection Filter Resultant', as illustrated in Figure 4.5. Selection of effective representation is vital for better classification. Both these representations are selected keeping the prime objective in view, i.e. to capture better attributes for modeling visual-motor deformations like tremor and micrographia. Brief details of the two representations are discussed below.



(a) (b) (c)

Figure 4.5: Multiple representations of input data (a) Raw generated image (b) Median filter Residual (Pixel values inverted for better visualization) (c) Edge detection filter resultant image (Pixel values inverted for better visualization))

- **Median filter residual (D_m):** To compute the median filter residual, we apply a 3x3 median filter on the raw image and then subtracted the raw image from the resultant filtered image. The idea is to preserve high frequency imperfections that can model tremor-like irregularities.
- **Edge detection filter resultant (D_e):** Edges are known to contain useful information in most cases. By applying linear convolutional filters in both vertical and horizontal directions, the magnitude of the gradient is computed in a non-linear way. As a result, we obtain emphasized edge information of the shapes.

Similarly, for each of the 8 tasks performed by every subject, corresponding three representations (D_r, D_m, D_e) are generated. It is worth mentioning that since signal values for each task are

provided in separate files and thus after reconstruction, separate images for each task are available. Therefore, segmentation is not required in this scenario.

4.2.3 Feature Extraction Using Pre-Trained ConvNets

As discussed in the previous chapter, feature extraction for deformation modeling is performed by employing pre-trained ConvNets. Leveraging on the ability of a CNN to extract highly discriminating features, we assume that CNN-based attributes will be able to characterize intra-shape deformations (tremor and micrographia), required for PD diagnosis. Although the internal representations of CNN layers are hard to decode, an intuitive guess can be made by visualizing the output of various layers of the network. Figure 4.6, shows outputs of a random activation channel after convolutional layer 3 on two input images of a spiral drawn by a healthy subject and a PD patient. It is clearly seen that neurons of the same activation channel react differently to smoothly drawn spiral edges and to irregular ones. This gives some indication that the network is capable of learning discriminating features required for this problem.

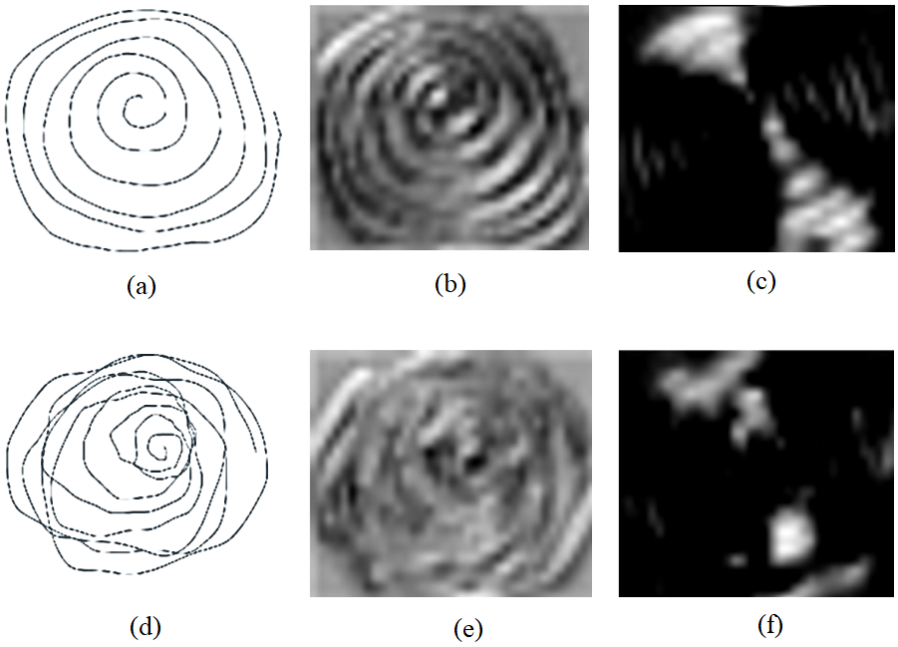


Figure 4.6: Visualization of CNN based features (a) Input image drawn by a healthy subject (b) Output of convolutional layer 3 (c) Corresponding activation channel showing neurons with maximum activity (d) Input image drawn by a PD patient (b) Output of convolutional layer 3 (c) Same corresponding activation channel showing neurons with very little activity

Figure 4.7 shows the proposed feature extraction and enhancement methodology. CNN-based features are extracted from each of the three representations (D_r, D_m, D_e) of the input data, independently. For this purpose, raw images and their corresponding multiple representations are fed to three independent pre-trained CNN models (e.g. AlexNet). The basic notion behind the independent extraction is to reduce combined feature noise, since all three representations are inherently different. The original images are binary, while CNN models usually expect a three

channel image as input. Therefore, we replicate the same image into all three channels before feeding it to the ConvNet. The features extracted by the three independent pre-trained models are then fused into a high dimensional combined feature vector. This is a kind of early fusion technique. It is worth mentioning that in all our experiments, features extracted from the last layer before softmax are employed. Therefore, the size of the feature vectors depends on the number of neurons of the last fully connected layer (e.g. 4096 for *fc7* layer of AlexNet). Same is repeated for each of the 8 tasks performed by all subjects belonging to both groups (PD/HC).

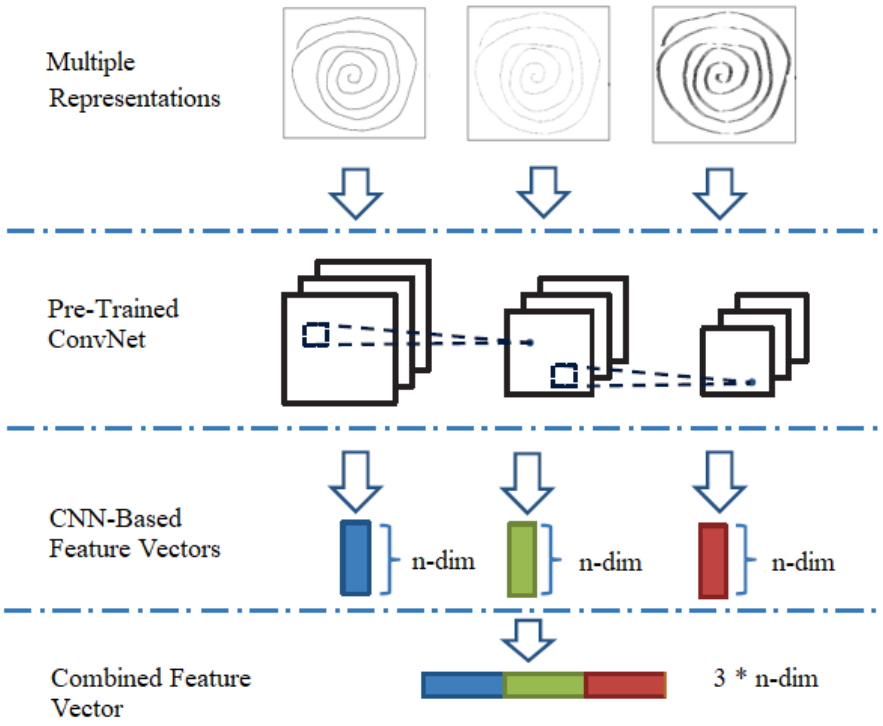


Figure 4.7: Proposed feature extraction and enhancement methodology

4.2.4 Classification and Disease Prediction

As discussed earlier, due to the scarcity of training examples for both (PD/HC) groups, we did not train the CNN models over these data, but rather used them only as feature extractors. The combined feature vector of each task is therefore, used to train a dedicated machine learning classifier. As discussed earlier, one of the incentives of selecting PaHaW database is that it contains multiple graphomotor samples of an individual. This gives us an opportunity to train task-specific classifiers. For instance, the resultant feature vectors of all spiral tasks of PD subjects and healthy controls are employed to train a spiral classifier. Similarly, seven handwriting task-specific classifiers are also trained. The proposed approach is different from the state-of-the-art, where features extracted from all tasks are combined to train a classifier. The proposed approach has two major benefits:

- Firstly, it allows us to study the impact of each graphomotor task on the classification performance.
- Secondly, it enables our system to make the final decision based on majority voting.

Each classifier performs a binary classification (since there are two output classes (PD/HC)). It is vital to mention that, despite training independent classifiers for each task, the entire sample of a subject (PD/HC) is employed for training, as well as, testing. For instance, if spiral drawing of a subject is employed to train a spiral classifier then the handwriting samples of the same subject are employed to train the handwriting classifiers as well. Same is considered for a test sample. This is necessary for the final decision regarding a subject. The objective at hand is to take task-level decisions from multiple samples of the same subject. The outputs of the 8 classifiers in our system form the decision vector d defined as $d = [d_1, d_2, d_3, ..., d_8]^T$, where $d_i \in \{c_1, c_2\}$ and c_i denotes label of either of the class (i.e. PD/HC). In the next step, we apply voting based late fusion. This strategy is suitable for a multiple classifier system, where each classifier gives a single class label as an output, as considered in our proposed system. By varying parameters, we can adjust the weightage given to number of tasks used for final diagnosis. The complete working of the system is illustrated in Figure 4.8.

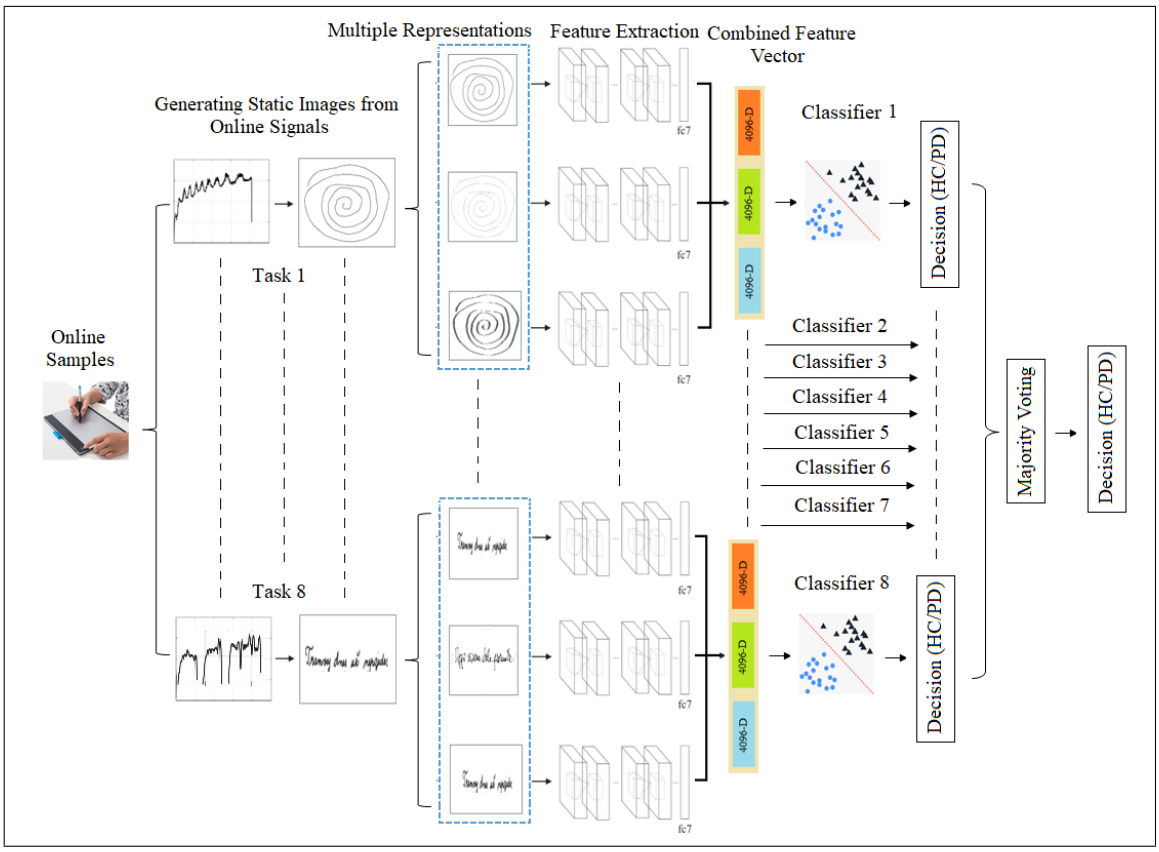


Figure 4.8: Complete schematic flow of the proposed methodology

4.3 Experimental Protocol

In this section, we describe the experimental protocol designed to assess the effectiveness of our proposed scheme for the identification of PD-related deformations from the graphomotor tasks of a subject. The prime objectives intended from the validation are as follows:

- To determine the impact of multiple representations of the input data on the classification performance
- To validate the effectiveness of our combined feature vector-based early fusion approach
- To observe the impact of individual graphomotor tasks on the proposed methodology
- To establish the effectiveness of our voting-based late fusion approach
- To compare the performance of CNN-based visual features with procedural features outlined in state-of-the-art
- To compare the predictive performance of our proposed scheme in comparison with the state-of-the-art

To achieve each of the aforementioned objectives, we employ 10-fold cross validation. This is a common practice employed with small datasets, (such as PaHaW), to estimate model performance and to determine the impact of overfitting. Distribution for each task considered in our study is presented in Table 4.2. Stratified sampling is employed to split the data to ensure equal representation

Table 4.2: Task-Wise Distribution of Samples for Each Class (PD/HC) in PaHaW

Task	Description	PD	HC	Total
1	Archimedean Spiral	36	36	72
2	Letter *l*	36	36	72
3	Bigram *le*	36	36	72
4	Trigram *les*	36	36	72
5	Word *lektorka*	36	36	72
6	Word *porovnat*	36	36	72
7	Word *nepopadnout*	36	36	72
8	Sentence *Tramvaj dnes už nepo-jede*	36	36	72

of each class in every fold. All experiments are conducted on a CUDA-enabled NVIDIA GPU.

4.3.1 ConvNet Model Architecture

It is important to establish that our prime concern is not to assess the impact of different pre-trained models but to assess the applicability of CNN-based features to characterize PD associated visual-motor deformations. Consequently, we have employed the convolutional base of AlexNet [199] model as feature extractor in our experiments. AlexNet is one the earliest and popularly employed

ConvNet model. AlexNet architecture consists of 5 convolution layers and 3 fully connected layers. The input layer accepts an image of size (227 x 227 x 3). The first convolutional layer applies convolution on the input images with 96 kernels of size (11 x 11 x 3). A stride size of 4 is employed during convolution. The subsequent convolutional layer accepts the response-normalized and pooled output of the previous layer as an input and applies convolution with 256 kernels of size (5 x 5 x 48). Maxpooling is then applied on the output of the second layer and fed to the third convolutional layer. This layer has 384 kernels of filter size (3 x 3 x 256). The output of this layer is then fed to the fourth and subsequently fifth convolutional layer without an intermediate pooling layer. The number of kernels in convolutional layers four and five are 384 and 256, respectively. Fully connected layers *fc6* and *fc7* contain 4096 neurons each. The detailed architecture of AlexNet model (till last layer before softmax) is described in Table 4.3.

Table 4.3: Detailed AlexNet Architecture Employed in The Experiment

Layer	**Type**	**Input**	**Filter Size & Number**	**Stride**	**Output**
Data	Input Data	227 x 227 x 3	-	-	3 x 227 x 227
Conv1	Convolution	227 x 227 x 3	11 x 11 and 96	4	55 x 55 x 96
pool1	Max Pooling	55 x 55 x 96	3 x 3	2	96 x 27 x 27
Conv2	Convolution	27 x 27 x 96	5 x 5 and 256	2	27 x 27 x 256
pool2	Max Pooling	27 x 27 x 256	3 x 3	2	256 x 13 x 13
Conv3	Convolution	13 x 13 x 256	3 x 3 and 384	1	13 x 13 x 384
Conv4	Convolution	13 x 13 x 384	3 x 3 and 384	1	13 x 13 x 384
Conv5	Convolution	13 x 13 x 384	3 x 3 and 256	1	13 x 13 x 256
pool5	Max Pooling	13 x 13 x 256	3 x 3	2	256 x 6 x 6
Fc6	Fully Connected	6 x 6 x 256	-	-	4096 x 1
Fc7	Fully Connected	4096 x 1	-	-	4096 x 1

As mentioned earlier, AlexNet is originally trained on 1.2 Million images (with 1000 different classes) of the ImageNet database. The network constructs a hierarchical representation of input images. Deeper layers contain higher-level features, constructed using the lower-level features of earlier layers. Together, the convolutional and down sampling layers serve as feature extractors while the fully connected layers represent a trainable classifier similar to a standard multi-layer neural network. For feature extraction, the softmax (classification) layer is removed and the output of the last layer before the softmax layer (i.e. *fc7*) is extracted. This produces a 4096-dimensional feature vector. Independently extracted 4096-D feature vectors from each of the three representations (D_r, D_m, D_e) are then concatenated to form a 3x4096-D fused feature vector C_F for each task performed by a subject. C_F is formulated by Equation 4.1, where C_{f_r} is the feature vector extracted from raw image D_r, C_{f_m} is extracted from median residual image D_m and C_{f_e} is the feature vector extracted from the edge image D_e. $\bigoplus$ is the concatenation symbol.

$$C_F = \bigoplus [C_{f_r}, C_{f_m}, C_{f_e}] \tag{4.1}$$

4.3.2 Classifier Employed

The resultant feature vectors $C_{Fi=1}^{N}$ extracted from N training samples of a task are then fed to the task-specific machine learning classifier. In our experimentation, we use Support Vector Machine (SVM) [186] for the purpose of classification. In general SVMs, work by constructing a separating hyperplane between the two classes so that the minimal distance from the closest data points of either class (PD/HC) is the largest. Test examples are predicted to belong to a class based on which side of the hyperplane they fall. Since we are already employing high-dimensional non-linear data, it is practical to use a linear kernel instead of a radial basis function (rbf) kernel, while training an SVM. Radial basis function (rbf) is commonly employed to map features to a high dimensional space to enhance learning, however, we believe that this may cause overfitting in our scenario. Consequently, a linear kernel is employed in all our experiments.

4.3.3 Performance Metrics

The effectiveness of the proposed scheme is evaluated by computing the system accuracy for each of the tasks separately and against each representation. Accuracy is also reported by combining the feature vectors of the three representations (early fusion) as well as by combining the predictions of the eight modalities through majority vote (late fusion). Furthermore, class-wise precision, specificity and sensitivity values are also reported. Equations 4.2, 4.3, 4.4, and 4.5 define the respective performance metrics in terms of True Positives (t_p), False Positives (f_p), True Negatives (t_n) and False Negatives (f_n).

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{4.2}$$

$$Sensitivity = \frac{t_p}{t_p + f_n} \tag{4.3}$$

$$Specificity = \frac{t_n}{t_n + f_p} \tag{4.4}$$

$$Precision = \frac{t_p}{t_p + f_p} \tag{4.5}$$

4.4 Results and Analysis

In this section we present and analyze the empirical results obtained from our experimentation. As discussed earlier, the main objectives of our experimentation are: (i) to determine the impact of multiple representations independently and in fused form, (ii) to determine the impact of templates (tasks) on the performance of the extracted features, (iii) to ascertain the effectiveness of the voting-based decision method in our scenario, (iv) to assess the performance of CNN-based features as compared to procedural features, and lastly, (v) to compare the performance of our complete proposed scheme for PD identification with the state-of-the-art.

4.4.1 Impact of Multiple Representations and Early Fusion Based Approach

Our first set of experiments is conducted to investigate the impact of each of the three input data representations on the classification accuracy of a task-specific classifier. Table 4.4 reports the accuracy of the system against individual representations for each of the 8 tasks. The values reported represent average accuracy of ten runs with standard deviations. Independent comparison of the performances of the three image representations under consideration suggests that median filter residual images (D_m) and edge detector filter resultant images (D_e) outperform the raw generated image representation (D_r), in all 8 tasks. This suggests that our hypothesis that both transformations can characterize fine irregularities caused due to movement impairments is correct. However, it is not conclusive which representation is better than the other. In some tasks median filter residual outperforms edge detector resultant while in others edge detector resultant performs better. For instance, in the spiral task both representations produce relatively same accuracies (i.e. approximately 65%). In tasks *l* and *le*, (D_e) outperforms (D_m). In all the remaining tasks i.e. trigram (*les*), words and sentence, (D_m) performs better than (D_e).

Table 4.4: Task-wise System Accuracies for Different Data Representations

	Data Representation		
Task	D_r	D_m	D_e
Archimedean Spiral	0.57 ± 0.05	0.65 ± 0.06	0.65 ± 0.09
Letter 'l'	0.53 ± 0.09	0.55 ± 0.10	0.57 ± 0.09
Bigram 'le'	0.48 ± 0.09	0.51 ± 0.09	0.54 ± 0.08
Trigram 'les'	0.50 ± 0.11	0.57 ± 0.09	0.55 ± 0.07
Word 'lektorka'	0.49 ± 0.10	0.58 ± 0.07	0.52 ± 0.11
Word 'porovnat'	0.46 ± 0.08	0.49 ± 0.09	0.48 ± 0.08
Word 'nepopadnout'	0.54 ± 0.07	0.64 ± 0.07	0.60 ± 0.05
Sentence	0.48 ± 0.08	0.49 ± 0.09	0.48 ± 0.09

After determining the impact of all representations (D_r, D_m and D_e) independently, we now present the results of the experiments carried out to validate the proposed early fusion-based technique to enrich feature sets for training purposes. Different combinations of the three representations are evaluated to select the best combination for the fusion vector C_F. Four possible combinations that can be formulated with the given three representations are expressed in Equations 4.6, Equations 4.7, Equations 4.8, and Equations 4.9, respectively. Experiments were carried out by employing all these combinations and the results are summarized in Table 4.5. All experiments were conducted following the 10-fold cross validation protocol.

$$C_{F_{r,m}} = \bigoplus [C_{f_r}, C_{f_m}] \tag{4.6}$$

$$C_{F_{r,e}} = \bigoplus [C_{f_r}, C_{f_e}] \tag{4.7}$$

$$C_{F_{m,e}} = \bigoplus [C_{f_m}, C_{f_e}] \tag{4.8}$$

$$C_{F_{r,m,e}} = \bigoplus [C_{f_r}, C_{f_m}, , C_{f_e}] \tag{4.9}$$

Table 4.5: Task-wise System Accuracies for Different Combinations of Data Representations

	Data Representation			
Task	$D_r + D_m$	$D_r + D_e$	$D_m + D_e$	$D_r + D_m + D_e$
Archimedean Spiral	0.67 ± 0.08	0.70 ± 0.05	0.65 ± 0.08	0.76 ± 0.08
Letter 'l'	0.55 ± 0.12	0.52 ± 0.08	0.50 ± 0.09	0.62 ± 0.08
Bigram 'le'	0.51 ± 0.09	0.52 ± 0.11	0.55 ± 0.07	0.57 ± 0.09
Trigram 'les'	0.54 ± 0.07	0.52 ± 0.08	0.57 ± 0.05	0.60 ± 0.08
Word 'lektorka'	0.54 ± 0.08	0.52 ± 0.11	0.51 ± 0.09	0.60 ± 0.07
Word 'porovnat'	0.50 ± 0.09	0.49 ± 0.09	0.47 ± 0.06	0.51 ± 0.09
Word 'nepopadnout'	0.65 ± 0.06	0.59 ± 0.06	0.63 ± 0.08	0.68 ± 0.07
Sentence	0.50 ± 0.09	0.49 ± 0.10	0.49 ± 0.06	0.51 ± 0.08

We first discuss the results of the combination of D_r and D_m on all tasks. In comparison to the results presented in Table 4.4, it is evident that the combined feature vectors $C_{F_{r,m}}$ extracted from all 8 tasks outperform individual D_r-only results. This reconfirms our prime hypothesis that multiple representations provide additional information as compared to raw data, in this scenario. On the contrary, the comparison of the results with D_m-only and D_e-only experiments show a mixed trend. An overall improvement in classification results is observed in tasks like the Archimedean spiral, words (*porovnat* and *nepopadnout*), and the sentence, when compared with both D_m-only and D_e-only outcomes. However, in case of the trigram *les* task, $C_{F_{r,m}}$ observes a drop in the accuracy as compared to the ones obtained by the D_m-only and D_e-only experiments. The results of the given combination in letter *l* and bigram *le* remain approximately the same as D_m-only, however, when compared with the outcomes of D_e-only, a decrease in accuracy is observed. For the word *lektorka*, the combination outperforms D_e-only but observes a decrease as compared to D_m-only outcomes.

For the second combination ($D_r + D_e$) represented by $C_{F_{r,e}}$, it is observed that accuracies in all tasks improve when compared with D_r-only, except for letter *l*. While comparing with the results of D_m-only, we observe a decrease in accuracies of tasks letter *l*, trigram *les*, and words (*lektorka* and *nepopadnout*) and an increase in the Archimedean spiral and bigram *le*. No effect is observed on the outcomes of word *porovnat* and the sentence task. Finally, we compare the outcomes of $C_{F_{r,e}}$ with D_e-only results and observe an increase in the Archimedean spiral, word *porovnat* and the sentence task. A degradation in performance is observed in all other tasks except for the word *lektorka* which shows no significant difference. While comparing the results of the two combinations $C_{F_{r,m}}$ and $C_{F_{r,e}}$, we notice that $C_{F_{r,m}}$ outperforms $C_{F_{r,e}}$ in all tasks except the Archimedean spiral and the bigram *le* task.

We now evaluate the performance of the third combination ($D_m + D_e$) represented by $C_{F_{m,e}}$ on all tasks. As expected, the accuracies of the $C_{F_{m,e}}$ combination increase as compared to D_r-only results (except in letter *l* task). However, while comparing with D_m-only outcomes, a decrease in performance is observed in the letter and word tasks. An improvement in bigram task is observed, while no significant changes are seen in the spiral, trigram and sentence tasks. When compared with the D_e-only outcomes, there is an improvement the performance of tasks like bigram, trigram, word *nepopadnout* and the sentence task, while the spiral shows no difference. Classification accuracies

decrease in task of letter and the two words *lektorka* and *porovnat*. When the results of $C_{F_{m,e}}$ are compared with $C_{F_{r,m}}$, it is noticed that the performance decreases in all tasks except the bigram and the trigram. A similar trend is observed while comparing $C_{F_{m,e}}$ and $C_{F_{r,e}}$, where only the results of bigram, trigram and word *nepopadnout* improve and the rest decrease. The sentence task shows no effect in this comparison.

Finally, we compare the performance of the last combination $C_{F_{r,m,e}}$, where the feature vectors from all three representations (D_r, D_m, D_e) are combined. Contrary to the other combinations discussed in this analysis, we observe a steady improvement in all tasks as compared to the performance of the independent representations. Similar trends are observed while comparing the results of the three combinations $C_{F_{r,m}}$, $C_{F_{r,e}}$, and $C_{F_{m,e}}$ with the outcomes of $C_{F_{r,m,e}}$. This validates the effectiveness of our proposed early fusion-based approach where we hypothesized that CNN-features extracted from all three representations of the input data can be combined to achieve enriched feature sets, that can enhance the learning of each task-specific classifier.

To establish the statistical significance of each of the three representations of input data and their combined approach, we employed some statistical tests on our empirical results. We first employed the non-parameteric 'Friedman pre-test' [222] to analyze the performance of each representation independently and then in the combined form. The test ranks different scenarios based on their performance, for instance the best performance is ranked 1 and so on. In case of a tie, the two scenarios are assigned an average rank. The Friedman statistic is then computed as a function of the average rank of all scenarios which is computed by means of the Equation 4.10.

$$X_F^2 = \frac{12N}{k(k+1)}\left(\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right) \tag{4.10}$$

Where R_j is the average rank of scenario j, N is the number of experiments (i.e. $N = 10$ for 10-cross validation in our case) and k is the number of scenarios (i.e. $k = 4$ for three independent representations and one combined state). An improved version of the Friedman statistic is generally employed [223], which is given by Equation 4.11.

$$F_F = \frac{(N-1)X_F^2}{N(k-1) - X_F^2} \tag{4.11}$$

According to Friedman pre-test, our Null hypothesis stated that all scenarios have an identical impact on the performance of the classifier. The alternate hypothesis suggested that at least one scenario has a different impact on the classification than at least one other scenario. The results of the Friedman pre-test rejected the null hypothesis that all scenarios are equally effective. As a consequence, Nemenyi post-hoc test [224] was performed to determine the difference between the impacts of the three representations independently and in a combined state. The test computes a Critical Difference (CD) that is used to determine whether the distance between the average ranks of a pair of scenarios is statistically significant. For instance, if the difference between the mean ranks of two scenarios is greater than CD, there exists a significant performance difference between the two. The results of the Nemenyi test are presented in Figure 4.9.

Representation	Combined	Dm	De	Dr	Mean Rank
Combined	0	1	1	1	1.1
Dm	-1	0	0	1	2.5
De	-1	0	0	1	2.7
Dr	-1	-1	-1	0	3.6

0	Row representation is same as column representation
-1	Row representation is worse than column representation
1	Row representation is better than column representation

Figure 4.9: Performance comparison of individual representations (D_r: Raw Image, D_m: Median Residual Image, D_e: Edge Image) & combined $C_{F_{r,m,e}}$ approach using Nemenyi pairwise statistical test

We first discuss the results of independent representations. It is observed that by employing the proposed non-linear transformations on the raw input data, significant improvement of classification results is obtained. This further supports our original hypothesis that the proposed representations can capture rich deformation-specific features across all tasks, as compared to raw-only data. However, the test shows that the performance difference between the two non-linear transformations is not significant. The statistical significance of the proposed early fusion-based technique is also established by means of the Friedman pre-test and the Nemenyi post-hoc test in the same manner. It can clearly be observed in Figure 4.9 that the combined approach achieves a statistically significant improvement in the classification of all tasks as compared to individual approaches. Thus, validating the effectiveness of the proposed early fusion technique.

4.4.2 Impact of Graphomotor Tasks and Ensemble Approach

Task-wise deformation classification has been an important consideration in our proposed methodology. As discussed in the literature review, the selection of task/template during feature extraction, plays a vital role in the performance of the particular feature set during classification. Task-wise performances achieved by features extracted from the three representations independently and in the fused state are presented in Table 4.6.

Table 4.6: Task-Wise Performance using Individual & Combined Representations

	Data Representations			
Task	D_r	D_m	D_e	$D_r + D_m + D_e$
Archimedean Spiral	0.57 ± 0.05	0.65 ± 0.06	0.65 ± 0.09	0.76 ± 0.08
Letter 'l'	0.53 ± 0.09	0.55 ± 0.10	0.57 ± 0.09	0.62 ± 0.08
Bigram 'le'	0.48 ± 0.09	0.51 ± 0.09	0.54 ± 0.08	0.57 ± 0.09
Trigram 'les'	0.50 ± 0.11	0.57 ± 0.09	0.55 ± 0.07	0.60 ± 0.08
Word 'lektorka'	0.49 ± 0.10	0.58 ± 0.07	0.52 ± 0.11	0.60 ± 0.07
Word 'porovnat'	0.46 ± 0.08	0.49 ± 0.09	0.48 ± 0.08	0.51 ± 0.09
Word 'nepopadnout'	0.54 ± 0.07	0.64 ± 0.07	0.60 ± 0.05	0.68 ± 0.07
Sentence	0.48 ± 0.08	0.49 ± 0.09	0.48 ± 0.09	0.51 ± 0.08

It can be observed that Archimedean spiral reports highest accuracies across all three data representations (i.e. approximately 57% on raw images and 65% using median residual and edge resultant images). Same trend can be observed with the fused feature approach as well, where features extracted from all three representations of spiral task are combined to yield a classification accuracy of approximately 76%. This shows that visual features extracted from samples of Archimedean Spiral drawings provide better results than those extracted from handwriting tasks. Effectiveness of the Archimedean spiral task in visual analysis can be attributed to the fact that it is specifically designed to capture motor dysfunctions like tremor and micrographia. Amongst the remaining 7 handwriting tasks, it is observed that visual features extracted from the word *nepopadnout*, produce best classification accuracies. The length of the on-surface continuous stroke/movement can attribute to the effectiveness of the long word task in capturing the relevant features. Handwriting tasks with upward strokes or rising edges like *l*, *le*, *les* and *lektorka* produce relatively similar results. However, the word *porovnat* produced least effective results like the sentence task.

As expected from the literature review, varying impact of the tasks is observed in our proposed method as well. To overcome this limitation, we proposed a voting-based ensemble approach where decisions of each task-wise classifier contribute to the final prediction. As expected the classification results in all scenarios (i.e. individual and combined input representations) improved by employing the decisions of all tasks. This is contrary to the approach proposed in the state-of-the-art, where features from all tasks are combined to train a single classifier. The combined feature approach for all tasks can result in a detrimental performance due to the varying impact of graphomotor tasks on the effectiveness of same features. Therefore, we believe that instead of combining the features of all tasks, task-wise decisions may be employed to produce the final outcome. The ensemble of all task-wise decisions using majority voting results in improved accuracies in all scenarios (i.e. individual representations and combined), as shown in Table 4.7. We also carried out

Table 4.7: Performance Results of Voting Based Ensemble Approach

	Data Representations			
Metric	D_r	D_m	D_e	Combined
Accuracy	0.58 ± 0.07	0.68 ± 0.07	0.66 ± 0.07	0.83 ± 0.09
Precision	0.64 ± 0.13	0.67 ± 0.05	0.75 ± 0.19	0.89 ± 0.12
Sensitivity	0.55 ± 0.13	0.69 ± 0.14	0.72 ± 0.14	0.84 ± 0.14
Specificity	0.64 ± 0.07	0.65 ± 0.13	0.63 ± 0.24	0.82 ± 0.15

statistical investigations to evaluate the performance of our scheme on various tasks. The statistics computed by the Friedman pre-test state that performance of our proposed scheme on individual tasks is significantly different. As a result Nemenyi post-hoc test was performed. The results of Nemenyi post-hoc test are summarized in Figure 4.10. It is clearly seen that fusing decisions of all tasks based on majority voting significantly outperforms individual task-wise classification (with an exception of Task 1). While considering effectiveness of individual tasks, it is seen that classification performance of Task 1 (Archmidean Spiral) and Task 7 (word *nepopadnout*) is statistically better than the rest of the handwriting tasks. Same observations were made while

conducting the experiments suggesting that static visual features can capture effective information from templates which support on-surface continuity. Task 8 and Task 6 have the least impact on the performance of the system. Although Tasks (2, 4 & 5) perform significantly better than Task 8 and Task 6, there is no significant difference between their own performances.

Task	All Tasks	Task1	Task7	Task2	Task4	Task5	Task3	Task8	Task6	Mean Rank
All Tasks	0	0	1	1	1	1	1	1	1	1.4
Task1	0	0	1	1	1	1	1	1	1	2.3
Task7	-1	-1	0	1	1	1	1	1	1	3.8
Task2	-1	-1	-1	0	0	0	0	1	1	5.2
Task4	-1	-1	-1	0	0	0	0	1	1	5.5
Task5	-1	-1	-1	0	0	0	0	1	1	5.8
Task3	-1	-1	-1	0	0	0	0	0	0	6.3
Task8	-1	-1	-1	-1	-1	-1	0	0	0	7.2
Task6	-1	-1	-1	-1	-1	-1	0	0	0	7.3

0	Row task is same as column task
-1	Row task is worse than column task
1	Row task is better than column task

Figure 4.10: Performance comparison of tasks using Nemenyi pairwise statistical test

Since we are applying our proposed deformation estimation methodology to discriminate between samples of healthy controls and PD patients, it is important to determine the sensitivity and the specificity values in addition to overall classification accuracies. As discussed in the earlier sections, sensitivity measures the ability of the proposed methodology to correctly identify PD samples, while specificity measures the ability to correctly identify healthy controls. Both are important parameters in any medical diagnostic system. Table 4.7 shows an 84% sensitivity rate of the proposed scheme with early fusion of features from all three representations and a late fusion of decisions from all tasks. Similarly, a specificity rate of 82% is achieved by the proposed scheme. Since the dataset is balanced and has equal representations for each class in the training and test groups, the results reveal equally effective performance for the identification of both classes. Our proposed methodology also achieves an 89% precise identification results.

4.4.3 Comparative Analysis

The effectiveness of a proposed methodology is best established by a comparison with the state-of-the-art. Due to this reason, we have also performed a performance comparison of our scheme with the existing literature. Since we have employed the PaHaW database for the assessment of our proposed deformation identification methodology, therefore a meaningful comparison can be made with studies that employ the PaHaW database under similar experimental protocol (i.e. tasks, cross validation and classifier). Table 4.8 enlists the prominent studies discussed in the literature review that have also worked on the same dataset.

Table 4.8: Performance Comparison with Studies Employing PaHaW Database

Study	Features	Analysis	Accuracy
Drotar et al. (2013) [69]	On-surface kinematic and temporal features from all tasks are combined	Procedural	79.4%
Drotar et al. (2013) [70]	In-air features from all tasks are combined	Procedural	80.09%
Drotar et al. (2014) [82]	Both On-surface and In-air features from all tasks are combined	Procedural	85.61%
Drotar et al. (2015) [84]	On-surface non-linear dynamic features from all tasks are combined	Procedural	88.13%
Drotar et al. (2015) [104]	Energy, entropy, NCP, writing length, duration, writing length, duration, stroke (height/width)	Procedural	AUC (89.09%)
Drotar et al. (2016) [75]	Pressure and kinematic features from all tasks are combined	Procedural	81.3%
Mucha et al. (2018) [78]	Fractional derivative based kinematic features from Archimedean spiral	Procedural	70.55%
Impedevo et. al (2018) [83]	Kinematic, spatio-temporal and non-linear dynamic features from all tasks are combined	Procedural	71.33%
Proposed Technique	**CNN-based visual features from multiple representations of each task are combined**	**Visual**	**83%**

We first compare the performance of the proposed scheme against the common practice of fusing features from multiple tasks. As discussed earlier, combining features from all tasks can result in lower accuracies due to the varying impact of the tasks being employed. As a consequence, we proposed an ensemble of combining decisions of different tasks rather than combining features. It can be observed that with few exceptions ([82, 84, 104]) the proposed approach outperforms most of the techniques that rely on fusion of features.

Another interesting observation is that in comparison to popular procedural analysis based techniques, our proposed CNN-based enriched features provide comparable results. This is a significant contribution, since the insufficiency of existing visual features was the reason that led the researchers to explore the potential of new modalities. Although the reported performance is some cases is higher than that of ours, these procedural features rely on online information in the tasks which requires specialized hardware for sample acquisition. We, on the other hand, propose visual features for this problem which can be extracted directly from images of various tasks.

Comparative review of the outlined studies also reveals some other interesting trends. For instance, in almost all the studies that employ dynamic features (kinematic, pressure, non-linear dynamics and neuromotor), the least effective classification results are reported with the spiral task. Drotár et al. in [75] report that no statistically significant kinematic features could be extracted from the Archimedean spiral. On the contrary, pressure features reported in [75] showed potential, nonetheless, the performance was least effective as compared to other tasks. This is contrary to the results reported in our experiments, where the spiral task outperformed the rest. As discussed

earlier, Archimedean spirals are designed to capture visual-motor deformations like micrographia and tremor which are better analyzed using visual analysis based techniques as compared to procedural analysis based techniques. Another interesting observation is that in almost all these studies, the sentence task produced the best results. This is again contradicting with the outcomes of our experiments, where the sentence task produced the least effective results. The sentence task does not represent continuous on-surface movements but rather measures intermediate pauses between words. This is a procedural analysis-specific task where in-air movements produce the most effective analysis. On the contrary, the relatively less effective performance of the CNN-based features in the sentence task can be attributed to the fact that handwriting tasks comprising of varying length characters or intermediate pauses between words are not suited to capture the effects of micrographia. Due to this reason least effective accuracies are observed in the sentence task in our experiments.

While most of the work done in PD identification comprises procedural analysis based techniques, nonetheless, there are some attempts in the literature that address the problem from a visual analysis based perspective. These include the works of Pereira et al. [49, 50, 105] and others like [81] on the HandPD [49] dataset. The results reported in these studies are outlined in Table 4.9. Though primarily all our experiments are conducted on the PaHaW dataset, for comparison purposes, we also evaluated the performance of CNN-based visual features on the samples in the HandPD dataset. It can be observed from Table 4.9 that our proposed technique outperforms those employing hand-crafted visual features. The technique in [105] employs CNNs but the authors do not analyze drawn tasks. Instead, the pen-based signals captured using a smart pen are plotted and are fed to a 3-layered CNN model. Again, our proposed method reports better performance as opposed to that of [105] validating the ideas put forward in this study.

Table 4.9: Performance Comparison with Visual Analysis Based Techniques (HandPD Dataset)

Study	**Subjects**	**Task**	**Technique**	**Result**
Pereira et al. (2015) [49]	55	Archimedean spiral	Nine static features including Mean Relative Tremor are computed and analyzed using SVM	75.8%
Pereira et al. (2016) [105]	35	Archimedean spiral	Pen-based signals are converted into 2D images and used to train an 3-layered CNN	87.14%
Senatore et al. (2019) [81]	92	Archimedean Spiral	Same set of nine static features are extracted as in [49] and analyzed using CGP	76.6%
Proposed Technique	92	Archimedean Spiral	CNN-based visual features from multiple representation of each task are combined and analyzed using ensemble SVM	92.47%

4.5 Summary

This chapter investigated the effectiveness of visual attributes in identification of visual-motor dysfunctions associated with Parkinson's Disease (PD). While the existing literature primarily targets kinematic, pressure, spatio-temporal, non-linear dynamic and neuromotor features, we exploit the visual attributes of handwriting. The idea is not to deny the effectiveness of the rich online features but to manifest the fact that visual information in handwriting can still be effectively employed for this problem. Due to this reason, we employed a popular database PaHaW that is commonly used to assess the performance of various online features. To apply a visual analysis based technique on the online samples, we first converted them into static images by ploting on-surface pen positions captured by a digitizer tablet. Multiple representations of the raw generated images are then produced by employing non-linear transformations sensitive to minute imperfections. The feature set is enriched by combining resultant feature vectors of multiple representations into a single vector. Enriched feature vectors are then used to train a dedicated binary classifier (SVM in our case), which classifies the sample as healthy or PD. Multiple graphomotor samples of a subject are employed and majority voting-based late fusion approach is used to determine the final outcome. The proposed ensemble approach is different from the existing all task feature fusion-based approach that can effect the performance of the combined features due to the varying impact of different templates.

We designed an experimental protocol to first ascertain the effectiveness of multiple representations and the combined early fusion-based approach. We then determined the impact of various graphomotor tasks on the performance of our proposed features and validated the use of a voting-based late fusion approach, where decisions of task-specific classifiers are combined to predict the final outcome. The empirical results are also validated by means of statistical analysis.

Finally, we compared the performance of our proposed methodology with the state-of-the-art. The comparative analysis not only validated the effectiveness of our proposed methodology but also provided some interesting insights. It is evident that visual features can be enriched by the use of multiple representations of raw data. Furthermore, various templates have significantly different impact on the predictive performance of our methodology. Therefore, the choice of an appropriate template is highly correlated with the type of features (offline/online) being extracted. An interesting observation in this regard is the fact that our proposed technique performed best on the standard template (Archimedean spiral) used by the practitioners as compared to the non-conventional templates employed in most studies. This supports our initial hypothesis that our proposed technique can reduce the gap between conventional practices and modern technologies.

Chapter 5

Identification of Visual-Perceptual Deformations - An Application to Scoring of Bender Gestalt Test (BGT)

5.1 Introduction

In this chapter, we present the detailed application of our proposed deformation modeling scheme for the identification of visual-perceptual deformations. Both handwriting and drawings can be employed as psychometric tools for measuring the perceptual orientation of an individual. However, studies [225, 226], suggest that drawing being a multi-componential process can be affected by a wide variety of brain lesions. Furthermore, drawing-based tests are preferred amongst pre-school and primary children. Unlike handwriting, drawings are not affected by linguistic barriers and do not rely on literacy. Consequently, as discussed previously, we employ the analysis of a popular drawing-based test, the Bender Gestalt Test (BGT) [22], as our case study to identify the visual-perceptual deformations. Preliminary details on BGT test and the corresponding scoring have already been introduced in Chapter 1. In the current chapter, we provide the details of the proposed methodology and the samples being employed in Section 5.2. The experimental protocol is described in Section 5.3 and the results alongwith their detailed analysis are presented in Section 5.4. Finally, Section 5.5 summarizes the chapter.

5.2 Proposed Methodology

Challenges in computerized scoring of BGT samples (outlined in Chapter 1 reveal that hand-crafted features and a heuristic-based analysis will prove to be insufficient in modeling the wide variety of errors across multiple shapes. This observation was also validated by one of our pilot studies [27] where we strived to model the deformations using shape-specific geometric features. A heuristic-based technique to analyze the clinical deformations led us to the conclusion that such an approach is too tailored and despite being exhaustive, it will not be sufficient to model all

possible deformations that are observed by clinical practitioners (Preliminary findings summarized in Appendix C). Consequently, a deep learning-based approach is adopted for the analysis of BGT responses, an overall schematic flow of the methodology being presented in Figure 5.1.

As described earlier, BGT samples comprise multiple shapes drawn on a single sheet of paper. However, clinical analysis is performed on individual shapes. Due to this reason, our proposed methodology also treats a single sample as comprising of nine independent tasks, (similar assumption was made in the previous case-study). However, there are certain differences in both scenarios due to the nature of the samples (online/offline) and the analysis criteria under consideration. In the current study, offline BGT samples are being employed and hence, it requires pre-analysis segmentation of individual shapes. Furthermore, since Lacks' deformations are shape-specific, therefore, shape recognition is also an important pre-processing step. Once recognized, each BGT shape is then analyzed for a particular deformation, if applicable. All nine shapes are assessed for multiple deformations, thus we train deformation-specific classifiers capable of analyzing multiple shapes for the existence of a specific deformation. A decision vector is maintained for each shape of a sample containing the outcomes (error (1)/no-error (0)) of each deformation-specific classifier.

An interesting aspect of clinical BGT scoring is that the existence of a particular deformation is counted only once despite of it being present in all nine shapes. In other words, the frequency of the error is not important at sample level. The rationale behind this is the widespread use of BGT across multiple age groups, and hence, drawing ability of subjects may vary due to extraneous factors. Some shapes might be difficult for one subject to draw, while easy for another, simply due to the difference of age. By including several templates designed to capture the same deformation, chances of missing an indicator or counting redundant deformations is avoided. Based on the same scoring criteria, our proposed scheme combines the decisions of all nine shapes of a sample by employing *logical OR*, (i.e. even if the error exists in multiple shapes, it will still be considered as one for the sample). The final score is computed by counting the number of deformations present in the sample out of all eleven indicators. Generally, a cut-off value is employed by the practitioner to determine the severity of a disorder. For instance, for a value of 5, if the number of deformations is ≤ 5, the sample is considered 'normal', else otherwise. The threshold value can vary depending upon the target group. To facilitate practitioners, we compute a score but leave the prognosis for the expert. Consequently, we design our experimental protocol to evaluate the deformation classification performance of our methodology, rather than disease diagnosis.

5.2.1 Sample Acquisition and Ground Truth Labeling

To the best of our knowledge, there is no publicly available dataset of BGT drawings. Due to this reason, scored BGT drawing samples of 60 children (30 male / 30 age-matched females) are collected from the Department of Professional Psychology at Bahria University Islamabad, Pakistan. All participants were originally enrolled in a research study regarding learning disabilities in children, being conducted by the Department of Professional Psychology, in collaboration with a

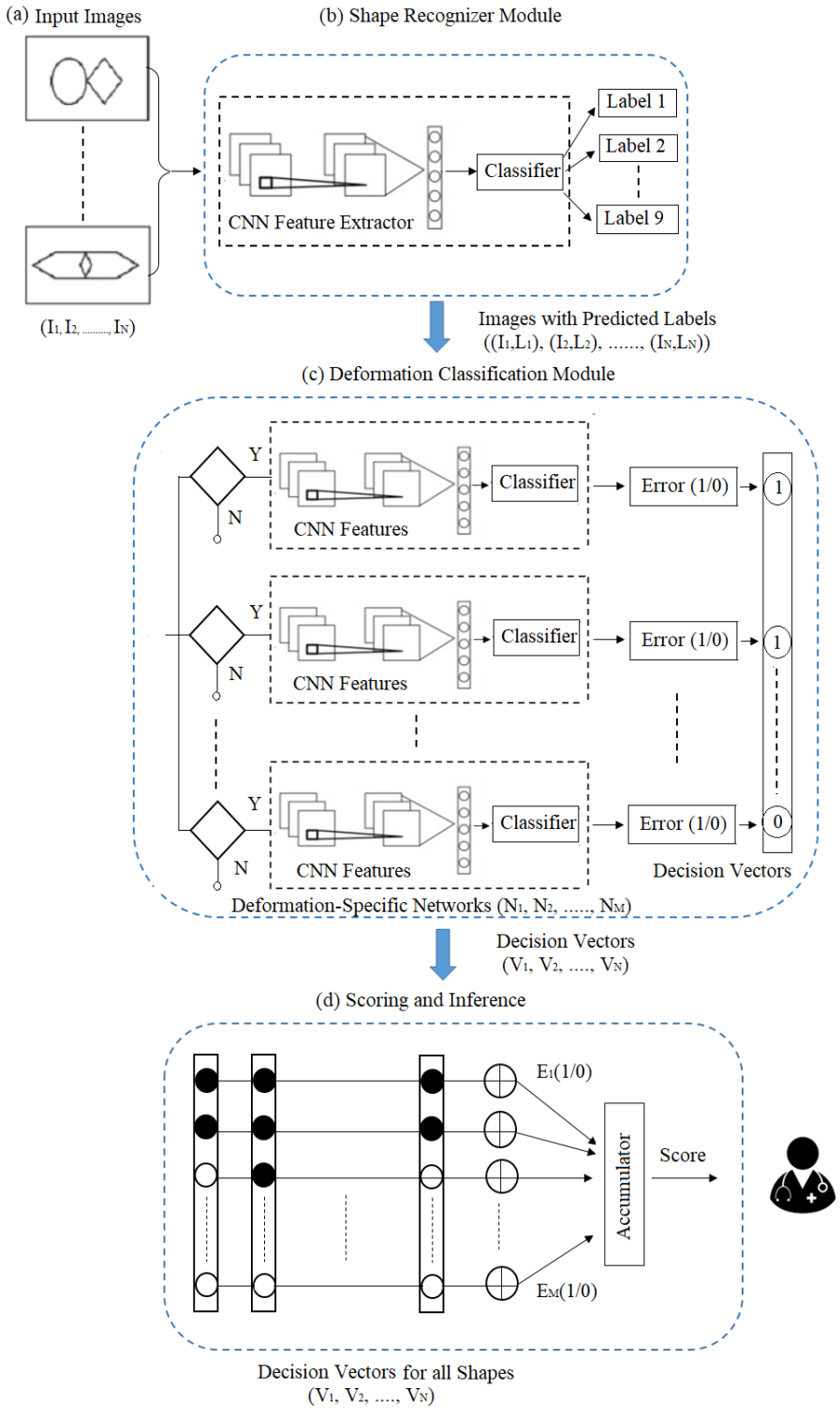


Figure 5.1: Proposed system architecture for deformation modeling and classification of BGT shapes: (a) Individual segmented shapes from each BGT sample are given as input, (b) Features extracted from each shape are fed to a classifier to determine the shape class, (c) Recognized image is then fed to each deformation network to determine the presence of the corresponding deformation, (d) Decision vectors from each sample are used to generate the final score.

local school. All participants were screened for learning disabilities using a test battery consisting of three tests, i.e. *Bangor Dyslexia Test (BDT)* [227], *Wide Range Achievement Test-4 (WRAT-4)* [228], and *Bender Gestalt Test (BGT)*. Since in our study, we are only focusing on BGT test, therefore, only the samples and the scores of BGT test are acquired. According to Lacks' scoring criteria, the maximum attainable score on a BGT test is eleven, while the minimum score is zero. An accumulated score of 5 is commonly considered as a cut-off, while a higher score is an indication of a possible dysfunction. The same criteria is considered for the inclusion of the subject population into the patient group (i.e. BGT score >5). All participants with BGT scores ≤ 5 are included in the control group.

As mentioned earlier, the objective of our study is not to diagnose learning disabilities or to validate the effectiveness of the BGT test for its diagnosis. Instead, we are targeting the identification of visual-perceptual deformations in the BGT samples of both healthy/diseased subjects. As evident from the score ranges of both groups (i.e. range of the scores obtained in the control group is [0-5], while that in the patient group is [6-11]), an overall healthy score (≤ 5) also contains some deformations (unless BGT score = 0). Similarly, the samples of children with disabilities can also contain some drawings without any perceptual deformation. Therefore, to achieve an overall score we must first be able to identify the type of perceptual deformations across all the shapes of a sample belonging to either group. Samples of children with learning disabilities were collected only to increase the probability of obtaining maximum training examples for each type of deformation. Table 5.1 outlines the demographic and education level data of the participant groups along with their mean BGT test performance scores.

Table 5.1: Demographic, Education and BGT Performance Levels of the Participants

Gender	**Number**	**Mean Age [Years]**	**Education Level [Grade]**	**Mean BGT Score**
Patient Group				
Males	15	13.2 ± 2.0	6-10	8.46 ± 1.98
Females	15	12.7 ± 2.2	6-10	8.20 ± 2.01
All	30	12.9 ± 2.1	6-10	8.33 ± 1.99
Control Group				
Males	15	13.6 ± 2.1	6-10	2.26 ± 1.48
Females	15	12.9 ± 2.5	6-10	2.02 ± 1.40
All	30	13.2 ± 2.3	6-10	2.14 ± 1.44

The samples are digitized by means of a scanner, while the scores are recorded as well. Both shape-specific deformation scores and the final sample-wise cumulative scores are provided. Nevertheless, our prime concern is the former.

5.2.2 Data Preparation

In order to keep the original test conduction protocol, our proposed system is designed to take offline scanned images of the test samples produced by the subjects (healthy/patient) as raw input. An important consideration at this stage was to decide whether to give the complete image containing multiple templates as an input or to provide each template response separately. As

discussed in the previous section, deformations can be shape specific and therefore, analysis must be performed shape-wise. Due to this reason, pre-segmented shapes of a sample are given as input to the deformation-specific classifier. To achieve this, we attempted two automatic segmentation techniques; one based on image processing and gestalt theory, while the other is based on deep learning. A brief discussion of both is provided in the following sections.

5.2.2.1 Image Processing Based Segmentation

The proposed technique groups the original nine gestalt shapes into three groups based on their perceptual similarities, as shown in Figure 5.2.

- **Group A:** This group consists of the BGT shapes A, 7, and 8. All three are closed shapes as shown in Figure 5.2-a.
- **Group B:** This group consists of the BGT shapes 4 and 6. These shapes are formed by continuous strokes of maximum length, as shown in Figure 5.2-b.
- **Group C:** This group consists of the BGT shapes 1, 2, 3, and 5. The primitive components of these shapes are disconnected dots or small circles drawn close to each other, as shown in Figure 5.2-c.

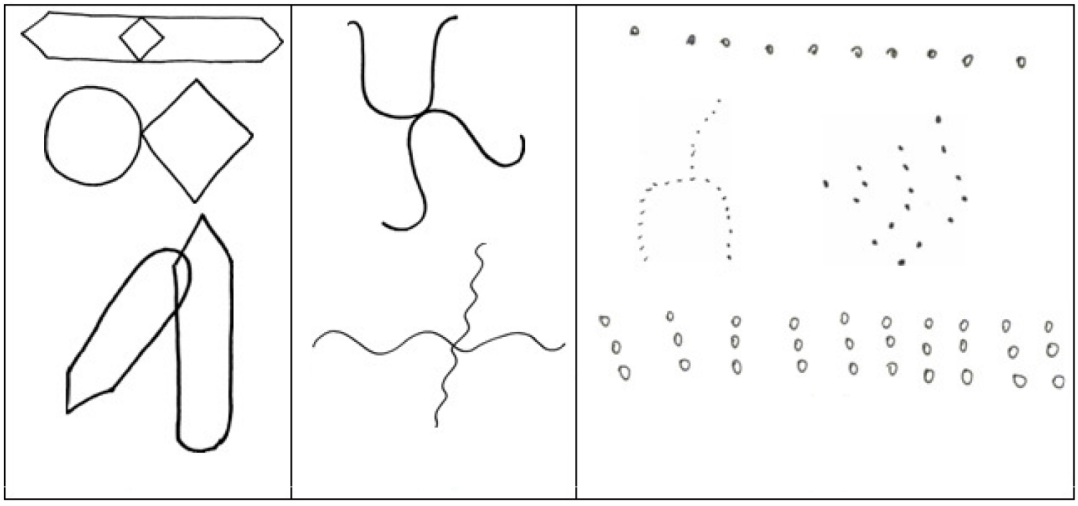


Figure 5.2: (a): Group A - Enclosed shapes (b): Group B - Shapes formed by solid lines (c): Group C - Shapes formed by dots or small circles/lines [229]

Due to the inherent attributes of each group, separate segmentation approach is required. The proposed technique has several benefits, firstly it allows us to segment each group separately thus reducing the impact of cluttering and noise, secondly it enhances localization by removing easier groups from the original sample in a hierarchical manner and leaving behind the most complex group i.e. Group C *(shapes formed by disconnected dots or small circles/lines)* for the last. Figure 5.3 shows the schematic order of the proposed segmentation technique. It can been seen in

Figure 5.2-a that Group A consists of three BGT shapes (A, 7, 8), all having a common attribute of enclosure. The area enclosed by these shapes is relatively greater than that enclosed by small circles of Group C. Leveraging the difference of enclosed area, we first isolate the shapes of Group A by applying various morphological procedures. After localization, we map the coordinates of the bounding boxes onto the original image and extract the required shapes of Group A. These shapes are then removed from the original image and we proceed to the extraction of the shapes in the next two groups.

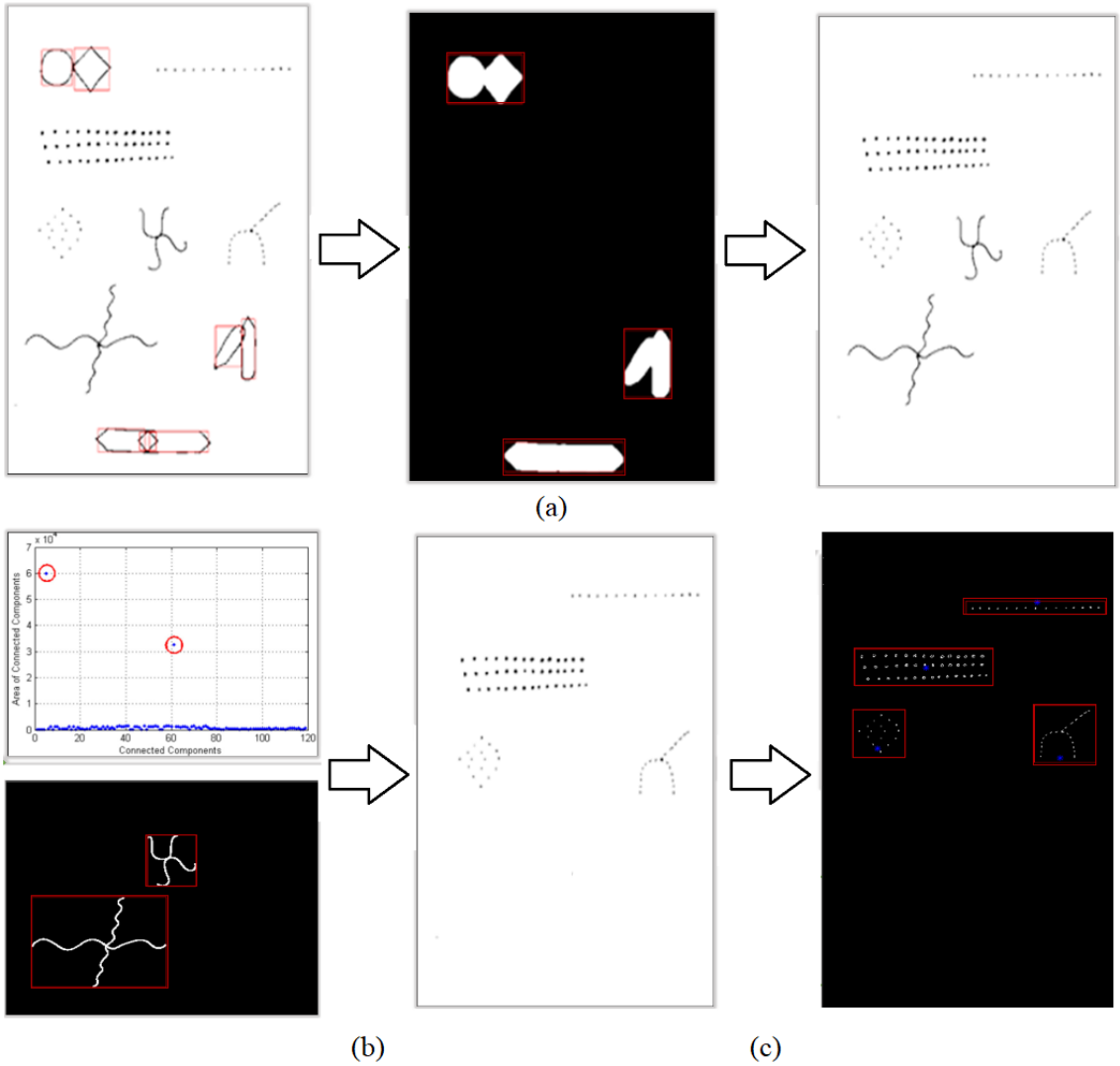


Figure 5.3: (a) Detection and segmentation of Group A shapes using morphological operations from original sample (b) Detection and segmentation of Group B shapes using connected component area (c) Detection and segmentation of Group C shapes using K-mean clustering

Group B consists of shapes which are formed by solid lines but are not enclosed, as shown in Figure 5.2-b. After removal of Group A, these two shapes are characterized by a common distinguishing attribute i.e. relatively much larger size than the remaining components. Therefore, in order to segment this group, we first determine a threshold area so that all components greater

than the threshold are localized while smaller ones are removed. For this purpose, areas of all the connected components are computed and the median area is determined. It is observed that desired (Group B) components are much larger in size than the rest. As a consequence, the threshold area A^T is determined by $\alpha * A^m$, where α is a constant value and A^m is the median of all computed areas. Once all components with area greater than the threshold are localized, we extract the shapes from the original sample.

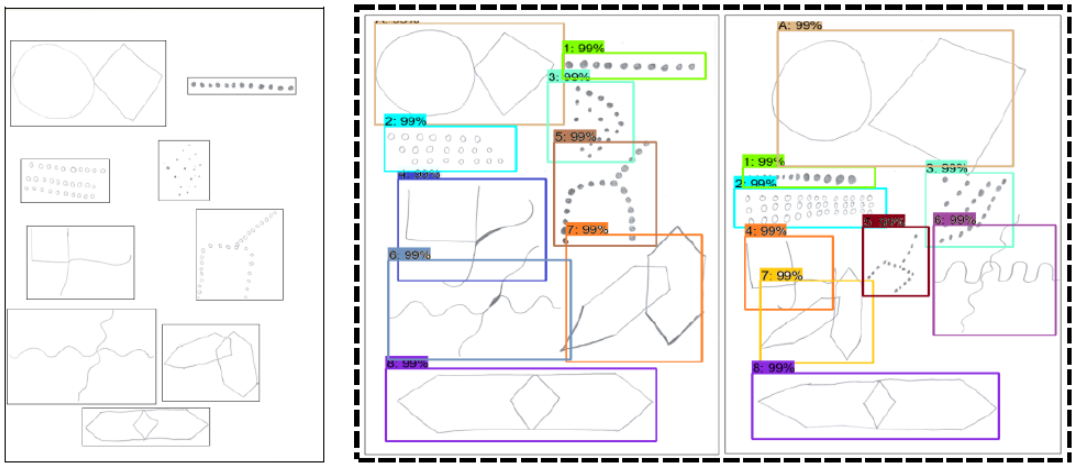
After removing the first two groups, only the shapes of Group C are left in the sample. Each shape consists of a large number of small sized separate components. As a result, detection and segmentation of this group is considerably more challenging than the previous two groups. It is observed that in most cases a subject simplifies the shapes consisting of circles by replacing them with dots or dashes and vice versa. Consequently, primitive shape features cannot be used to detect any of these shapes. Moreover, counting the number of connected components in each shape is not beneficial either, as different shapes may have approximately the same number of components and subjects may not always draw the shapes correctly. An appropriate feature for the detection of these shapes is the distance between its primitive components. Ideally, the connected components within a particular shape must be closer to one another as compared to the components of other shapes. Based on this hypothesis, we carry out k-means clustering on the spatial coordinates of the center of gravity of the components with k = 4. In most cases, acceptably good clustering is observed. The resulting clusters are then treated as shapes and are segmented from the image.

The details of the proposed technique are published in [229]. The technique proved effective in segmenting Group A and Group B shapes, however did not perform well on Group C due to the degree of deformations introduced by the subjects. The results of the experiments are reported in Appendix D for reference.

5.2.2.2 Deep Learning Based Segmentation

Recently, convolutional object detectors like 'Faster Region-based Convolutional Neural Networks (Faster R-CNNs)' [230], 'Single Shot Multibox Detector (SSD)' [231] and 'Region Based Fully Convolutional Networks (R-FCNs)' [232], have gained much popularity in automatic object localization and detection. Nevertheless, these meta-architectures have not yet been evaluated for the localization of hand drawn shapes and sketches, especially in the domain of neuropsychological drawings. To improve automatic segmentation of all BGT shapes, we evaluated the performance of various convolutional object detectors on the BGT samples. Figure 5.4 shows the detection of BGT shapes using convolutional object detectors. The details of the experimentation are reported in [233], while results are presented in Appendix D. As expected, the convolutional object detectors are able to detect target shapes (even with significant deformations) in a cluttered sample. Furthermore, an apt combination of ConvNet detector and feature extractor is capable of detecting even the most challenging shapes (i.e. shapes of Group C) which are otherwise difficult to localize as a whole object.

Despite the success of either of the two segmentation techniques evaluated during our study, it is important to avoid any mis-classification introduced due to segmentation errors. Therefore, all



(a) (b)

Figure 5.4: (a) Example of Multi-Object Sketch Detection Using Convolutional Object Detectors (a) BGT Training Sample with Ground Truth Bounding Boxes (b) BGT Test Samples with Cluttering and Shape Deformations

automatically segmented shapes are first manually inspected. In case of incorrect segmentation, the sample or the particular shape is then manually cropped to ensure correct segmentation. Consequently, the results of automatic segmentation of BGT shapes are not considered in overall performance evaluation of the proposed deformation modeling and estimation methodology. Hence, the input to the system are all the pre-segmented constituent shapes $I_1, I_2, ..., I_N$ where $N = 9$, of an offline BGT sample.

5.2.3 Shape Recognition

Shape recognition is a vital step in the analysis of a multi-object test like BGT, as deformations are highly shape specific (as discussed in Lacks' scoring system). Although the BGT test comprises of only nine shape classes, nevertheless, since the test is designed to screen visual-perceptual dysfunctions, therefore, shapes drawn by subjects can range from mildly to highly deformed. This makes automatic shape recognition a challenging task. Initially, in our pilot study [229], we explored the potential of 'Shape Context Descriptors' [234] for shape recognition (Figure 5.5). The detailed results are provided in Appendix E.

Later, in another pilot study [235], we evaluated the performance of CNN-based shape recognizers for the classification of the nine BGT shape classes. As expected, CNN-based features proved to be more robust than the high-level shape descriptors, therefore, we employ the same in this study as well. As discussed earlier, due to limited training samples, we use a pre-trained convolutional base as a feature extractor for deformation modeling. Same is considered for shape recognition purposes also. Each input image $I_1, I_2, ..., I_N$ of a given sample is fed to the feature extractor after resizing to match the input layer of the respective ConvNet employed. The extracted features are then used to train a machine learning classifier (e.g. SVM, LDA, etc.), which then predicts the shape class label $(L_1, L_2, ..., L_N)$, where $N = 9$. We conduct an in-depth empirical analysis on the performance

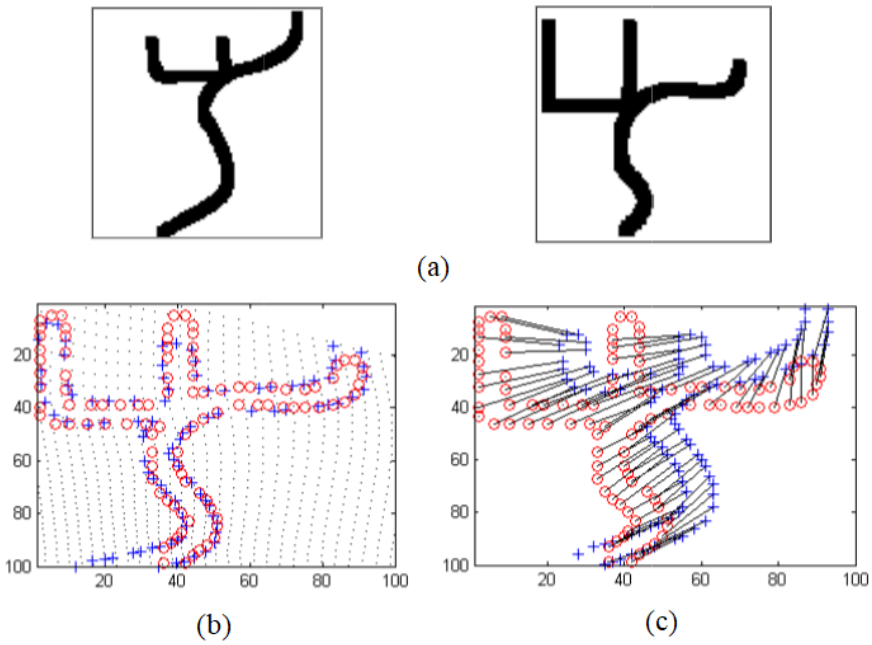


Figure 5.5: Matching of shapes (a) Original shapes (b) Sampling points (c) Correspondences [229]

of various pre-trained ConvNets in combination with a number of classifiers, to determine the best CNN-classifier combination for the BGT shape recognition task. The architectural details of the pre-trained ConvNets and the hyper-parameter specifications of the classifiers employed are presented in the experimental protocol section.

5.2.4 Deformation Modeling and Classification

The conceptual model of our proposed deformation classification module is illustrated in Figure 5.1-c. In order to model deformations, we again employ CNNs to extract deformation-specific features which are then used to train a classifier. The outcome of the classifier is a binary decision regarding the presence/absence of a particular deformation. The main design issue in this module was to decide whether:

- To train a generic model for each deformation across all shapes, for instance, one network to model the rotation error across all BGT shapes or,
- To train individual shape-specific deformation models (i.e. separate rotation models for each of the nine BGT shapes)

However, to avoid similar customization that is criticized in tailored rule-based approaches, a generic model for each deformation across all the shapes is designed to provide generality. Another reason to avoid a shape-specific deformation modeling approach is the scarcity of training data for a particular deformation across all templates. To overcome this issue and to further enhance representations, we employ augmented data customized specifically to represent a certain

deformation. The details of our deformation-specific augmentation techniques are discussed in the subsequent sections. Once all the deformation models are trained, we feed the recognized shapes with shape labels (from the shape-recognizer module) $((I_1,L_1),(I_2,L_2),...,(I_N,L_N))$ to our proposed deformation classification module. The module consists of all the deformation-specific networks $(N_1,N_2,...,N_M)$, where $M = 11$. Each input shape is assessed independently by all of the applicable deformation networks. The decision from each deformation model is then stored in a decision vector. Each decision vector consists of the results (error (1) / no-error (0)) of all the deformation models applicable to the particular BGT shape.

5.2.5 Deformation-Specific Data Augmentation

Data augmentation is a common practice to overcome the data scarcity, over-fitting and class imbalance issues. In our proposed methodology, deformation-specific augmentation is primarily being performed to provide missing examples of some deformations across each shape class. It is mentioned in the previous sections that the availability of shape-wise samples for each deformation is not feasible in a real-life scenario. Due to this reason, some deformation classes do not have considerable representation for each shape. Although shape-wise deformation modeling is not being performed, nonetheless, to generate some samples of the missing shape-wise deformations, we employ deformation-specific transformations on the non-erroneous shape samples. Our deformation-specific transformations can be categorized as *Generic* and *Shape-Specific*.

5.2.5.1 Generic Augmentation

All deformations except simplification, retrogression and perseveration, have common characteristics across all shapes on which they are applicable. Due to this reason, augmentation techniques for these deformations are relatively generic. A brief description of the transformations applied for a generic deformation across all BGT shapes is presented below:

- To generate data with rotation error, shapes from the original training samples, are rotated 2° apart to achieve rotated copies of the shape between 80° to 180° or mirror image (Figure 5.6-a). Rotation is also performed on the already erroneous data with a caution to ensure that no shape with original rotation error, becomes error-free. Also, for some BGT shapes, mirror image produces the same shape as the original (e.g. BGT shape 1 and 8, etc.). For such shapes, 180° rotation is not considered. Like rotation, the generation of angulation examples is achieved by rotation of the original images of template 2 and 3, at the angles between 45° and 80°.

- Samples for the spatial deformations like overlapping difficulty, collision and closure difficulty are generated by employing a controlled translation of the constituent parts of a shape. For instance, the overlapping difficulty error for BGT shapes 6 and 7, is generated by translating and merging the individual hand-drawn samples of their constituent parts in a way to produce an incorrect or missing overlap. Similarly, translation is also applied to

the separately drawn constituent parts of BGT shape A, 4 and 7, to join them at the wrong points, to represent the closure difficulty as shown in Figure 5.6-f. In the case of collision, different BGT shape templates are translated within proximity of one another. For some scenarios, two shapes are translated as shown in Figure 5.6-e, while for others, three or more BGT shapes are translated close to one another.

- Besides rotation and translation, other meaningful representations of the raw data are also evaluated for our selective augmentation technique. For instance, the median residual of the hand-drawn samples of PD patients has been employed in the previous case study, to detect tremors by highlighting the fine irregularities present in the shape contours. A similar technique is employed in our study to represent motor incoordination. After selecting shapes representing the motor incoordination error, we generate their median residuals. Both representations (i.e. raw and median residual) of the examples are used to generate relevant features as training data. For illustration purposes, Figure 5.6-g demonstrates the inverted image of the median residual of a sample of BGT shape A.

- As discussed earlier fragmentation error is represented by incomplete shapes or disconnected strokes. To produce fragmentation data, a $0.25r \times 0.25c$ sized window is randomly placed on the original shape image of size $r \times c$. If the window contains foreground pixels, they are converted into background and the image is saved. Consequently, several copies of the original shape image with the missing details are created. Figure 5.6-c shows an example of automatically introduced fragmentation error in BGT shape A.

- Size imbalance of the constituent parts of a shape represents cohesion error. Such shapes whose parts are already separated or can easily be separated by applying morphological operations, are used to generate the examples for cohesion. Some of the separated constituent parts are scaled up while others are scaled down (Figure 5.6-h) and merged to generate a shape with disproportionately sized components.

5.2.5.2 Shape-Specific Augmentation

As mentioned earlier, errors like simplification, retrogression and perseveration are marked by different characteristics across different shapes. Due to this reason, their data generation is highly shape-specific, as discussed below:

- To generate the perseveration examples, extra row(s) or column(s) of dots or circles in BGT shapes 1, 2 and 3 are added by replicating and merging the constituent parts of the shapes. In case of replacing circles with dots in BGT shapes 3 and 5, we apply morphological hole filling followed by erosion with a disk-shaped structuring element of an appropriate size. Both techniques represent perseveration data.

- Retrogression error is scored when a constituent part of a BGT shape is replaced by a primitive shape i.e dots with dashes, circles with loops, triangle, square or rectangle for a

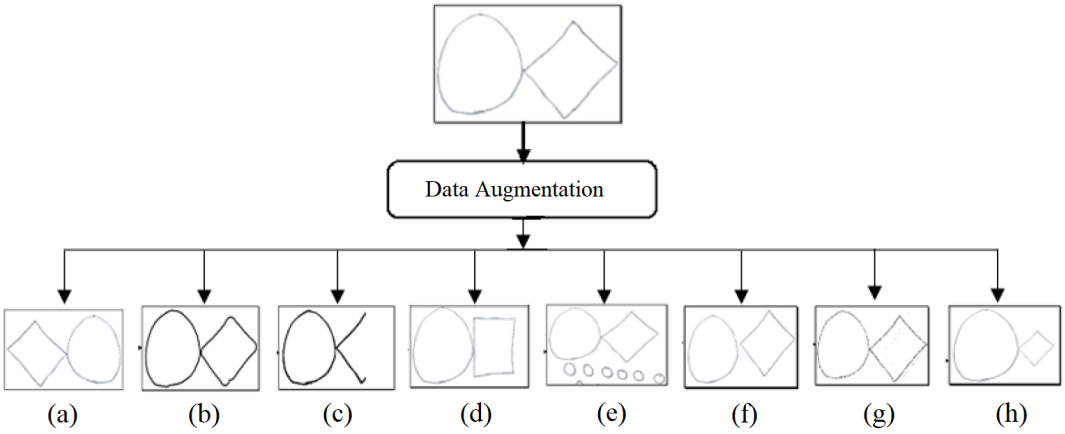


Figure 5.6: Example of deformation-specific augmentation results for BGT Shape A (a) Mirror image produced by rotation (b) Simplification of sharp angles of diamond into curves using morphological operations (c) Fragmentation introduced by converting part of foreground image into background (d) Replacement of constituent diamond with square to produce retrogression example (e) Translation of BGT Shape A and 2 to produce collision (f) Significant separation of circle and diamond for closure difficulty (g) Inverted median residual of original BGT Shape A for motor incoordination (h) Resizing of diamond to produce cohesion

diamond or hexagon. One method to generate this error is to create synthetic geometric shapes and merge them accordingly. However, synthetic shapes may not represent the imperfections of a hand-drawn shape. Therefore, to generate a near-realistic data, we asked some subjects to draw primitive geometric shapes like triangles, rectangles, squares, and circles, etc., on separate sheets of paper. The individual shapes are then segmented and merged as the deformed BGT shapes (e.g. Figure 5.6-d). Same augmentation technique is applied to all other errors where another shape replaces the original constituent part.

- Generally, simplification is marked whenever the fine details of the drawing are distorted. For instance, angles are curved, overlapping is missed by a great distance and fine dots are replaced by tiny circles. For overlapping shapes, we applied the same technique which was used for the overlapping difficulty i.e. the constituent parts are translated in a manner to allow maximum separation between them. BGT shapes A, 4, 6 and 8 are simplified by applying various morphological operations to convert their sharp angles into smooth curves (Figure 5.6-b). For BGT shape 1, where simplification means the replacement of dots with circles, we applied dilation with caution to avoid joining dots together. Once dilated and enhanced, a boundary extraction technique is applied to form circles from blobs. We also used hand-drawn samples of small circles to generate a row at different distances.

Figure 5.6 shows examples of deformation-specific augmented data generated for BGT template 'A' by employing some of the proposed techniques. The samples of the augmented data were verified by domain experts from the Department of Professional Psychology. A ground truth labeling tool was developed for similar purposes, (as shown in Appendix C).

5.2.6 Scoring and Inference

As discussed in the previous sections, the decision vector for each shape template contains the results of all the deformations applicable to it. These decision vectors $(V_1, V_2, ..., V_M)$ provide useful information to the practitioner regarding the sample drawn by a particular subject. A practitioner can use the decision vectors to validate his/her decision or can apply them to draw various statistics regarding the frequency of a particular deformation as well. Nevertheless, for this case study, we apply inference rules outlined in the Lacks' scoring manual.

According to Lacks' scoring, the occurrence of an error is more important than the frequency of the error. As a result, for all the instances of a particular type of error across all BGT shapes, a score of one is generated. For instance, if the rotation error exists in all of the nine BGT shapes, the score for the rotation error will be considered as one, irrespective of its frequency of occurrence. The same will be considered even if only one of the BGT shapes is rotated between 80° and 180°. Similarly, scores for other deformations are also generated. The final score of the sample will be the sum of all the deformations which are independently scored. The practitioner then decides a threshold value to determine whether the score lies above or below it. Any score above the threshold indicates a sign of brain dysfunction. The flexibility of the threshold value is due to the demographics of the subjects taking the test. To automate the generation of a final score, in the same way, we apply 'logical OR' on each of the error decision present across all the decision vectors and feed the results to an accumulator which then generates a final score of the test.

5.3 Experimental Protocol

In this section, we describe the experimental protocol employed in our study. As discussed earlier, the prime objective of our study is to assess whether the CNN-based visual features can represent visual-perceptual clinician deformations sufficiently to be employed for a high-level domain knowledge representation problem like the computerized analysis of neuropsychological drawings. To assess that, we evaluate the predictive potential of CNN-based features in following two scenarios:

- To recognize the nine BGT shapes even with high degree of deformations and a limited amount of training data
- To identify intra-class variations and inter-class similarities for deformation classification, under similar constraints as above

In order to assess the performance of our proposed *shape recognition* and *deformation classification* methodologies, we carried out separate experiments. All experiments are conducted on a CUDA-enabled NVIDIA GPU.

5.3.1 Data Distribution for Shape Recognition

For shape recognition, the pre-segmented drawing samples of all 60 subjects are employed, where each sample consists of the nine BGT shape classes, resulting in a total of 540 shape samples (60 samples for each BGT shape class). 5-fold cross validation is employed and in each fold, the data is divided into training and test sets accordingly, with each set containing an equal representation of samples from the two subject groups (healthy/patient) under study. For fair evaluation, caution is taken while dividing the data for training and testing. All the segmented shapes selected for training belong to the samples originally selected for training and the same is done for the testing samples.

5.3.2 Data Distribution for Deformation Classification

Data distribution for deformation training and testing is not so straightforward. Despite an equal number of samples of children from both groups, the individual deformation examples are highly unbalanced. One reason for this is that not all deformations are applicable on all 9 BGT shapes. For instance, *angulation* is scored only in BGT shape 2 and 3, thus, out of all 540 shape samples *angulation* can only be scored on 120 samples. Same is the case with *overlapping difficulty* that is only scored across BGT shape 6 and 7. The second reason of imbalance is due to the fact that, despite being applicable on multiple shapes, examples of a particular deformation may not exist with the same frequency across all. This can attribute to an unbalanced representation of same deformation across multiple shapes. Due to this reason, the impact of different templates is difficult to assess in this scenario, and therefore, task-wise deformation classification is not employed for BGT. Instead, a deformation-specific classification approach is adopted, where features extracted from all available samples of the deformation are employed to train a binary classifier. This enables classification of all deformation classes independently from each other, thus reducing the impact of class imbalance. After separating the classification for each deformation, the remaining data scarcity issue for some of the classes is overcome by the data augmentation techniques already discussed in the previous sections. After the data for each deformation class is sorted out into equal number of deformation-positive and -negative samples, stratified sampling based on 5-fold cross validation is employed for the experiments. Currently, our original dataset comprising of 540 individual shape samples drawn by all 60 (healthy/patient) subjects, contains a total of 299 deformation examples. The distribution of each deformation class is presented in Table 5.2. It is worth mentioning that since multiple deformations can co-exist, therefore, amongst the 299 deformation samples, same examples have been employed to represent more than one deformation classe.

5.3.3 Pre-Trained ConvNet Architectures Employed

As described earlier, the overall methodology consists of two steps i.e. *shape recognition* and *deformation classification*. As explained in the introductory chapters, CNNs are conventionally employed as a shape recognizers, where they extract shape-based features to enhance inter-shape class variances. At the same time, shape recognizers attempt to diminish intra-shape class variances

Table 5.2: Distribution of Deformation-Wise Samples in the Dataset

Deformation	**Applicable on BGT Shape**	**# of Samples**
Simplification	All nine BGT shapes	275
Cohesion	All nine BGT shapes	257
Fragmentation	All nine BGT shapes	245
Collision	All nine BGT shapes	209
Rotation	All nine BGT shapes	96
Motor Incoordination	All nine BGT shapes	66
Retrogression	All shapes excpet 4 & 6	132
Perseveration	1, 2, 3, & 5	84
Closure Difficulty	A, 4, & 7	144
Overlapping Difficulty	6 & 7	48
Angulation	2 & 3	66

that are common challenges of a free-hand sketch recognition system. In addition to shape recognition, the second step of the proposed methodology requires CNNs to identify deformation-specific similarities across different shapes, as well as, deformation-specific variances with in the same shape class samples. Since we are using pre-trained ConvNets for feature extraction in all scenarios, it is important to assess their performance in each. To achieve this purpose, we investigate a number of pre-trained CNN architectures. All the networks employed in our experiments have been pre-trained on ImageNet source data. Table 5.3 enlists the architectural details of the pre-trained ConvNet models employed in the proposed experimental protocol. The depth of the network, input image size, and layers from which the learned features have been extracted are mentioned. Brief description of each is already presented in Chapter 3.

Table 5.3: Summary of Pre-Trained CNN Architectures Employed

Model	Depth	Input Image Size	Feature Extraction Layer	Feature Dimensions
AlexNet	8	(227 x 227)	fc7	4096
VGG16	16	(224 x 224)	fc7	4096
VGG19	19	(224 x 224)	fc7	4096
SqueezeNet	18	(227 x 227)	pool10	1000
InceptionV3	22	(299 x 299)	predictions	1000
GoogLeNet	48	(224 x 224)	loss3-classifier	1000
ResNet50	50	(224 x 224)	fc1000	1000
ResNet101	101	(224 x 224)	fc1000	1000
DenseNet201	201	(224 x 224)	fc1000	1000

5.3.4 Multi-Class and Binary Classifiers

To assess the potential of CNN-based features in both scenarios i.e. shape recognition and deformation classification, we have employed a number of popular supervised machine learning classifiers. The features extracted from the pre-trained models are fed to train these classifiers independently

to observe the impact of the classifiers on the performance. It is worth mentioning that shape recognition is a multi-class (i.e. nine BGT shapes) classification problem, while deformation classification is a binary class (i.e. error or no-error) problem. Therefore, each classifier is trained accordingly. For shape recognition, four classifiers are employed. These include the Support Vector Machine (SVM) [186], Linear Discriminant Analysis (LDA) [192], Naïve Bayes (NB) [236] and Decision Trees (DT) [189]. Brief details of the classifiers and the respective hyperparameters involved in training are given below.

- Discriminant analysis is a statistical method that facilitates decision making by employing dimensionality reduction on the input data to rely on only the most discriminant values. A linear discriminant model is applied on the extracted CNN-based features. An LDA attempts to minimize the variance between the input features of a class in such a way that it maximizes the distance between the means of the distinct classes. Ranking threshold is an important hyperparameter while applying an LDA. It is the value that determines the inclusion and exclusion of an instance in the feature space. A threshold value of 0.0001 is selected.

- SVM is a non-probabilistic classifier that models a hyperplane to separate the labeled classes. A linear, one-versus-all SVM is trained for the shape recognition task. The tolerance value is set to 0.0001 and cost parameter is set to 1.

- NB is a probabilistic classification technique. We trained a multinomial NB model on the extracted CNN feature vectors. A Laplacian smoothing prior is applied to prevent the impact of zero probabilities on the decision.

- DT is a predictive model that can be used for classification. An important parameter in a classification tree is the number of splits (k) which controls the depth of the tree. The value selected for the tree splits is k=50.

For the deformation classification task, we employed an LDA classifier with similar hyperparameter values.

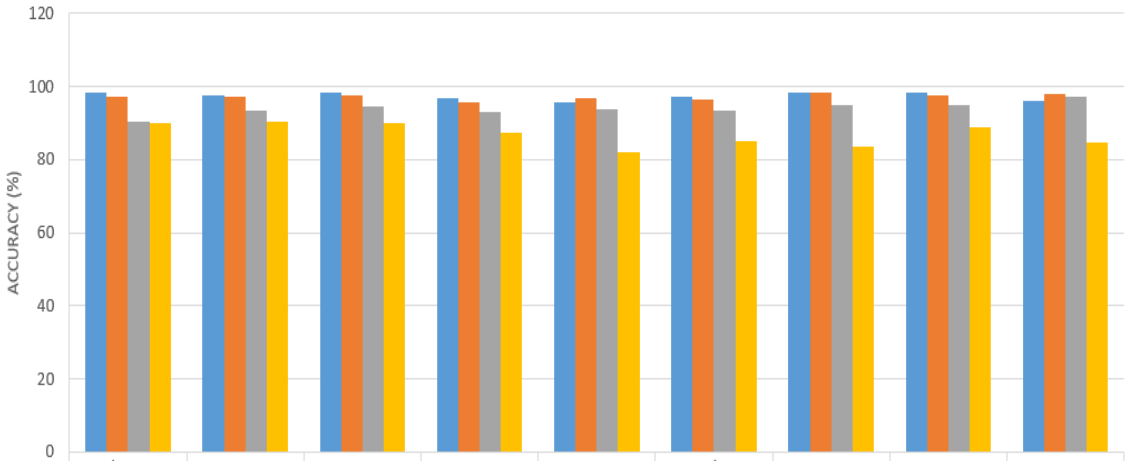
5.4 Results and Analysis

This section discusses the results of our proposed empirical analysis for the BGT shape recognition and deformation classification methodologies presented in this study.

5.4.1 Shape Recognition Results

To assess the effectiveness of pre-trained ConvNets in BGT shape recognition, we first evaluate the overall classification accuracies achieved by the combination of each CNN architecture employed with the aforementioned classifiers. As mentioned earlier, a total of 540 shape samples (with 60 samples of each shape) are divided into training and testing using 5-fold cross validation. For each fold, the overall shape accuracy is computed as $\frac{t_p+t_n}{t_p+t_n+f_p+f_n}$, where t_p, t_n, f_p, and f_n represent the

total number of True Positives, True Negatives, False Positives, and False Negatives respectively, for all nine classes. Mean classification accuracy is then computed from the accuracies achieved by each of the 5 runs of experiments. Based on the mean accuracies, Figure 5.7 shows the performance comparison for each (CNN-Classifier) combination, assessed in this study.



	AlexNet	VGG16	VGG19	SqueezeNet	InceptionV3	GoogleNet	ResNet50	ResNet101	DenseNet201
LDA	98.33	97.41	98.33	96.67	95.74	97.22	98.33	98.33	95.93
SVM	97.04	97.22	97.39	95.74	96.85	96.48	98.15	97.59	97.78
NB	90.37	93.52	94.63	93.15	93.7	93.33	95	94.82	97.04
DT	90.17	90.37	90	87.22	81.85	85.19	83.52	88.89	84.62

Figure 5.7: Overall shape classification accuracies achieved by each CNN architecture in combination with classifiers employed

It is observed that the performance of the features extracted from each pre-trained ConvNet employed is comparable with one another when fed to the same classifier. This supports our initial claim that the pre-trained CNN architectures can be successfully employed to a limited shape class dataset, even without augmentation. Furthermore, it is observed that the highest classification accuracy (i.e. 98.33%) is achieved by training an LDA with features extracted from AlexNet, VGG19, ResNet50 and ResNet101, independently. This shows that shape classification is not significantly affected by the choice of the CNN architecture employed. However, the choice of the classifier is important as both LDA and SVM outperformed NB and DT, significantly. Both LDA and SVM reported comparable accuracies across each CNN architecture, with LDA slightly outperforming in most cases.

To get a deeper insight, shape-wise classification results of our proposed technique using AlexNet-LDA combination, are reported as a confusion matrix in Table 5.4. From the confusion matrix, it is evident that almost all shape classes are successfully recognized. Few instances of mis-classification between BGT shape 1 & 2, 3 & 5, 7 & 8 and, 7 & 'A' are observed, which might have resulted due to the deformations introduced by the subjects while drawing these shapes.

Table 5.4: Confusion Matrix of Shape-Wise Classification Results Obtained By AlexNet-LDA Combination

Class	1	2	3	4	5	6	7	8	A
1	60	0	0	0	0	0	0	0	0
2	1	59	0	0	0	0	0	0	0
3	0	1	58	0	1	0	0	0	0
4	0	0	0	60	0	0	0	0	0
5	0	0	1	0	59	0	0	0	0
6	0	0	0	1	0	59	0	0	0
7	0	0	0	0	0	0	60	0	0
8	0	0	0	0	0	0	2	58	0
A	0	0	0	0	0	1	1	0	58

5.4.2 Deformation Classification Results

In this section, we discuss the deformation classification results. It is important to mention that due to the scarcity of examples of each deformation across all BGT shapes, shape-wise comparison is not feasible. Therefore, the reported results of deformation classification are error specific. Table 5.5 reports the mean classification accuracies of 5-folds for each deformation using the given CNN architectures in combination with LDA classifier. Several important observations are made while analyzing the outcomes reported in Table 5.5.

Table 5.5: Overall Deformation Classification Accuracies Achieved by each CNN Architecture in combination with LDA Classifier

	CNN Architecture								
Deformation	AlexNet	VGG 16	VGG 19	Squeeze Net	Inception V3	GoogLe Net	ResNet 50	ResNet 101	Dense Net201
Rotation	90.47%	89.52%	92.38%	91.66%	89.52%	86.66%	89.52%	**96.19%**	95.23%
Overlap	70.83%	62.5%	79.16%	58.33%	54.16%	62.5%	66.66%	**79.16%**	75.0%
Simplification	85.84%	81.13%	87.73%	84.90%	86.79%	83.01%	86.79%	**90.56%**	87.73%
Fragmentation	78.50%	84.11%	85.04%	83.17%	79.43%	73.83%	81.30%	**85.98%**	81.37%
Retrogression	94.18%	96.51%	96.51%	95.34%	94.18%	95.34%	94.18%	**97.61%**	95.34%
Perseveration	79.16%	81.25%	81.25%	79.16%	82.05%	81.25%	79.16%	**83.33%**	81.25%
Collision	84.25%	86.11%	85.18%	79.62%	87.96%	85.18%	75.92%	**90.74%**	89.81%
Closure	58.06%	67.74%	61.29%	54.83%	58.06%	61.29%	70.96%	**80.64%**	74.19%
Motor Incoordination	83.33%	80.55%	82.40%	79.62%	85.18%	84.54%	86.11%	**87.96%**	82.40%
Angulation	66.66%	70.83%	75.0%	79.16%	75.0%	62.5%	66.66%	**83.33%**	70.83%
Cohesion	61.45%	70.83%	73.95%	58.06%	67.74%	64.51%	70.96%	**80.64%**	75.0%

- Contrary to their conventional use, CNN-based features can also be employed to enhance (rather than diminish) intra-class variations and inter-class similarities.
- CNN-based features extracted from the deformation-specific augmented data are capable of representing the particular deformations sufficiently, as observed in cases like *Overlapping difficulty*, *Motor incoordination*, *Angulation* and *Rotation*. Despite having comparatively less amount of original data (as shown in Table 5.2), these deformation classes achieved

comparable results with those (e.g. *Fragmentation* and *Cohesion*) having more amount of original data.

- Unlike shape recognition, where the depth or the width of a ConvNet does not have a significant impact on the classification, in deformation classification, the choice of a suitable architecture can enhance performance.

- In general, deeper networks (i.e. ResNet101 and DenseNet201) appear to outperform wider networks (i.e. GoogLeNet & InceptionV3). In most cases, it is observed that ResNet101 outperforms other networks despite achieving variable accuracies across different deformations. DenseNet201 also demonstrates comparable results. It is an expected outcome since wider networks require more parameter tuning in each layer and thus require more training data to ensure better approximation and to avoid overfitting [237]. In comparison to wider networks, deeper networks (especially with skip-connections like ResNet), enhance approximation with considerably less number of neurons. Thus, increasing depth improves performance without increasing computational complexities. Since both the wider and the deeper networks employed in our study are pre-trained on same source data (i.e. ImageNet), the weights of ResNet101 and DenseNet201 provide a better generalization.

For a deeper insight, we also compute the 'Specificity', 'Sensitivity' and 'Precision' values in addition to accuracy. As described in the previous chapter, Sensitivity is the measure of the ability of a system to correctly classify the deformations and is calculated by the ratio $\frac{t_p}{t_p+f_n}$, while Specificity measures the ability of the system to correctly classify the non-erroneous samples and is defined as $\frac{t_n}{t_n+f_p}$. 'Precision' is the true positive relevance rate and is defined as $\frac{t_p}{t_p+f_p}$. Table 5.6 details the sensitivity, specificity and precision values of the deformation classification module using a ResNet101-LDA combination.

Table 5.6: Sensitivity, Specificity and Precision Achieved by ResNet101-LDA Combination

	Metric		
Deformation	Sensitivity	Specificity	Precision
Rotation	90.0%	96.84%	90.0%
Overlap	73.33%	88.88%	73.33%
Simplification	82.60%	92.77%	82.60%
Fragmentation	80.0%	87.35%	59.25%
Retrogression	75.0%	98.75%	75.0%
Perseveration	72.72%	86.48%	72.72%
Collision	68.18%	96.51%	83.33%
Closure	75.0%	84.21%	75.0%
Motor Incoordination	81.25%	89.13%	56.52%
Angulation	80.0%	85.71%	80.0%
Cohesion	75.0%	84.21%	75.0%

It is observed that in most cases, our proposed deformation classification scheme achieves promising results while considering the sensitivity of the system. However, in some cases, lower values of sensitivity are also obtained. For instance, in the case of *Collision*, *Perseveration* and *Overlapping difficulty*, the sensitivity of the system is below 75.0%. This can be attributed to two reasons, i.e. fewer samples and nature of the deformation. To assess the impact of amount

of sample data, we compare sensitivity values of the deformation class with the largest (i.e. *Simplification*) and the smallest (i.e. *Overlapping Difficulty*) number of samples. It is observed that *simplification* achieves a higher value of sensitivity (i.e. 82.60%) as compared to *overlapping difficulty* (i.e. 73.33%). This supports the initial hypothesis that fewer original samples contribute to the difference in performance. Nonetheless, when comparing the instances of the highest (i.e. 90.0%) and the lowest sensitivity (i.e. 68.18%) values obtained by *rotation* and *collision*, respectively, we observe that despite having fewer samples, *rotation* outperforms *collision*. This indicates that the imbalance of training samples across different deformation classes is not a conclusive reason for the varying performance. Furthermore, it also supports that the proposed augmented data is sufficiently addressing data scarcity issues as well.

The other reason for varying performance across different deformation classes can be attributed to the nature of the deformation itself. For instance, Figure 5.8-a, shows an example of a correctly drawn BGT shape 7, while Figure 5.8-b and Figure 5.8-c, demonstrate examples of the BGT shape 7 marked with 'Overlapping Difficulty' in the ground truth by trained psychologists. Visual inspection of Figure 5.8-b reveals that it is very similar to Figure 5.8-a. Consequently, our system identifies it as 'Non-Erroneous'. On the contrary, Figure 5.8-c displays a more severe deformation that is correctly identified by the system as 'Erroneous'. This shows that the ability of the system to identify deformations greatly depends on the challenging nature and severity of the deformation.

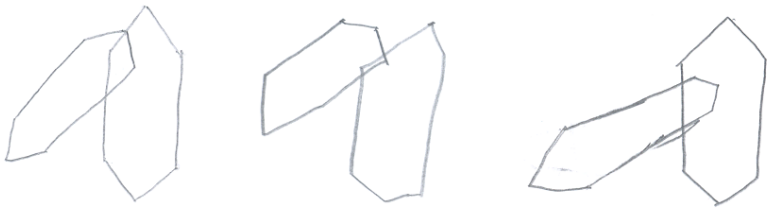


Figure 5.8: Three examples of BGT shape 7 assessed by our proposed system (a) Sample with no overlapping difficulty, correctly identified as sample with no overlapping difficulty (b) Sample with overlapping difficulty, incorrectly identified as sample with no overlapping difficulty (c) Sample with overlapping difficulty, correctly identified as sample with overlapping difficulty

As mentioned earlier, Lack's scoring standard emphasizes the importance of the occurrence of the error rather than the frequency of the error, hence if a deformation is missed in one shape template, it can be captured in another.

5.4.3 Comparative Analysis

A comparative analysis with the state-of-the-art gives a better insight into the effectiveness of any proposed method. However, despite a rich and extensive literature review, it is observed that a direct feature-wise comparison with previous studies cannot be carried out. The primary reason is that there is a paradigm shift from visual-based analysis to signal-based procedural analysis, due to the lack of better representation of the visual features and heavy reliance on extensive heuristics.

Consequently, more work has been attempted on neuropsychological tests which involve the analysis of motor impairments or delayed executive planning (e.g. Spiral tracing and Clock drawing). On the contrary, tests like BGT which provide information about the visual-perceptual orientation of a subject, received very little attention from the relevant pattern recognition community. Nonetheless, from the perspective of completeness, we attempt to compare the performance of our proposed drawing analysis methodology with some of the prominent works in this domain. A comparison summary is outlined in Table 5.7. However, it is worth mentioning that the comparison is based on addressing the challenges of the existing systems, rather than the outcome.

Table 5.7: Performance Comparison with Studies Employing Visual Analysis of Neuropsychological Drawings

Study & year	**Task**	**Samples**	**Study Objective**	**Analysis**	**Outcomes**
Smith and Hiller 1996 [43]	Necker's Cube	32/36 (Online)	To identify indicators of visual-spatial neglect	Four static geometric features extracted from drawings of patients with VSN are compared with those of healthy controls	Only two out of four features prove successful discriminators
Canham et al. 2005 [45]	ROCF	31 (Offline)	To score 18 scoring regions	Localization of scoring regions is performed using fuzzy-logic based heuristics followed by extraction and matching of static geometric features	Only 3/18 regions are localized and an average accuracy of 75% in grading of the regions is reported
Bennasar et al. 2014 [48]	CDT	648 (Offline)	Differential diagnosis of different stages of dementia	A set of geometric and spatial features are extracted and used to train cascaded classifiers	Acc. = 77.78% (for stage 3) & Acc. = 74.38% (for stage 4)
Moetesum et al. 2015 [27] (A pilot study)	BGT	18 (Offline)	To classify 11 clinical deformations for visual-perceptual dysfunctions	Shape-specific geometric features are extracted and assessed using a heuristic-based approach	Only 6/11 deformations are classified across a small subset of shapes with accuracy ranging from 63.8% to 94.2%
Harbi et al. 2017 [52]	CDT	65/100 (Online)	To score CDT drawings of healthy subjects and dementia patients	Ontology-based domain knowledge representation and fuzzy logic-based heuristic analysis of geometric features extracted from clock components is performed	99% and 95.7% classification rates are reported for identifying samples of patients and control subjects respectively.
Proposed Methodology	**BGT**	**60 (Offline)**	**To classify 11 clinical deformations for visual-perceptual dysfunctions**	**CNN based visual features are extracted from raw images and augmented samples and used to train an LDA classifier**	**All 11/11 deformations are classified across all shapes with accuracies ranging from 79.1%to 97.6%**

We first discuss the work done by Smith and Hiller in [43], that aimed to assess the quality of cubes drawn by subjects with potential VSN. Authors employed samples of children below the

age of 11 years to represent the group with under-developed cognitive skills. Assuming that the drawn cubes will be highly deformed, authors attempted to model deformations by measuring the quality of oblique angles and paralellism of the sides. Two other geometric features were also measured but failed to discriminate between the cubes of children with developed visual-spatial skills and those with under-developed ones. Although the remaining two features showed promise, however, in a later attempt [54], authors had to include several procedural features to enhance the overall performance. The study only addressed one simple shape and highlighted the insufficiency of hand-crafted visual features to assess its quality. On the contrary, we are addressing nine shapes and assessing their quality to identify several indicators of impaired brain development in children.

Another comparison can be made with [45] where authors attempted to score the ROCF test. As discussed in the literature review, ROCF is a complex drawing test with an extensive clinical scoring criteria like BGT. To score a complete ROCF response, the authors proposed a component-based assessment technique that first localized the intended scoring regions and then determined the presence or absence of a particular component by employing heuristics. Nonetheless, due to the highly unconstrained nature of the drawings, localization of only 3 out of 18 regions was performed successfully. Within the localized regions, the authors reported a mean accuracy of 75% in scoring of clinical errors. The study highlighted the challenges of localization in a component-level analysis approach due to which only partial scoring was possible. Unlike component-level analysis, we are performing the complete figure-based analysis. This avoids the challenges of localization and thus enables complete scoring of BGT.

A popularly analyzed neuropsychological drawing in the literature is the clock drawing that is assessed to determine presence (or progression) of dementia in the elderly. Similar to ROCF and BGT, the clinical scoring of CDT is also extensive. However, CDT is designed to identify impairments associated with memory and executive planning. This requires the system to assess the spatial organization of the clock components (digits and hands) within a clock circumference. In an attempt to score clock drawings of subjects suffering from different stages of dementia, authors in [48] extracted various spatial features based on the distance between the clock components. These distance-based features were then used to train a cascaded classification architecture, where the first classifier distinguished between the healthy and diseased samples, while the second and third classifiers differentiated between various stages (3 and 4) of dementia. The authors reported classification accuracies of 77.78% and 74.38%, respectively for each type of dementia. Nonetheless, extraction of distance-based features required the authors to map a layout on the digitized clock drawings. To avoid the need for an additional layout mapping, Harbi et al. in [52], proposed the use of online sample acquisition tools like a digitizer tablet while conducting tests. To prove the effectiveness of online sample acquisition for easy localization of clock components, authors acquired existing paper-based samples from the local hospital and traced all samples manually by attaching them on the screen of the digitizer tablet. This was performed due to the limitations of modifying the original test conduction protocol. On the contrary, in our proposed deformation analysis methodology, existing paper-based samples can be digitized by scanning. Furthermore, despite the ease of localization and feature extraction using online samples, CDT

scoring required an extensive ontology-based heuristic approach. Our machine learning based approach overcomes the limitations of heuristic dependencies.

Finally, we compare the current strategy with a pilot study conducted earlier in [27]. A set of hand-crafted geometric features were extracted from some of the BGT shapes to model deformations. However, it was observed that it required an exhaustive rule-based approach to estimate all possible deviations across each BGT shape. As a result, a small set of BGT drawings (18 samples) were employed to model 6 out of 11 deformations (i.e. simplification, overlap, rotation, perseveration, closure and cohesion). On the contrary, in the present study, by employing deep CNN-based features, we are able to classify all 11 deformations (simplification, fragmentation, overlap, rotation, perseveration, closure, cohesion, angulation, collision, motor incoordination and retrogression), on a relatively larger dataset (60 samples) with promising results. These findings validate the robustness and scalability of the proposed method in estimating the challenging BGT deformations.

5.5 Summary

In this case study, we applied our proposed deformation modeling strategy to identify visual-perceptual deformations produced by subjects while drawing templates of a popular neuropsychological test called BGT. Contrary to the conventional approaches that extract hand-crafted shape based features at component-level and either employ feature matching or rely on extensive heuristics, our proposed methodology models clinical manifestations using deep visual features and assesses them by training machine learning classifiers. To effectively represent a wide variety of clinical deformations without extensive heuristics, pre-trained CNN architectures are employed. Feature representation is enhanced by using deformation-specific augmentation. In the present study, we employed CNN-based features for two purposes i.e. to recognize the nine BGT shapes and to classify eleven visual-perceptual deformations. Using pre-trained ConvNets as feature extractors overcomes the issue of data scarcity which is commonly observed in health related problems like the one under consideration.

To evaluate the effectiveness of our proposed methodology, a dataset of 60 subjects was collected, the samples were digitized and individual BGT shapes were segmented before analysis to reduce possibility of errors introduced by incorrect segmentation. Different ConvNet models were employed to extract features for both shape recognition and deformation identification. Shape recognition was treated as a multi-class problem while deformation identification was considered as a binary classification problem. The experimental study validated our preliminary hypothesis that CNN-based visual features can represent domain knowledge sufficiently without an extensive rule-based approach.

Although direct comparison was not possible with any of the existing literature, nevertheless, our proposed scheme addressed several challenges outlined. The prime objective of this study was to create a benchmark for future studies in this direction. In the future extensions of this study, impact of transfer learning using CNN architectures other than those trained on ImageNet can

also be explored. Transfer learning using fine tuning has also been considered [238]. Impact of a template on capturing of deformation-specific features is another interesting direction that must be explored for better representation of deformation models. Shape-wise deformation classification can also be pursued. Observing the frequency of a particular deformation can also provide an important insight into the behaviour of the subjects with potential brain dysfunctions. This can provide a very useful exploratory direction for researchers in clinical psychology. Other BGT scoring manuals can also be modeled and evaluated as a future extension of this work.

Chapter 6

Conclusion and Future Directions

The role of handwriting and drawings in the domain of neuropsychology is well established. Similarly, handwriting analysis and sketch recognition have remained popular research areas among the computer science and pattern recognition community. Nonetheless, despite its significance limited efforts have been made to provide computerized solutions for the analysis and interpretation of neuropsychological responses of patients for the identification of graphomotor deformations. The main hindering factor in this regard has been the modeling of domain knowledge. Unlike conventional handwriting analysis and sketch recognition applications, computerized analysis of neuropsychological graphomotor impressions is highly domain specific. The sensitivity of the outcome requires an explicit understanding of the clinical procedures, while the acceptance of the assistive technology requires ease of use and limited modifications in conventional practices. This thesis addressed the problem of domain knowledge representation and proposed a method to provide an effective and robust solution while preserving the original clinical practices.

Despite the sporadic nature of the state-of-the-art in this area, the thesis categorized the proposed techniques into two broad categories based on their mode of analysis. The first category includes the techniques that analyze a graphomotor response after completion, in a manner similar to clinical practitioners and thus is termed as 'Visual Analysis based Techniques'. These techniques rely on static geometric and spatial features and require extensive heuristics to overcome insufficiency. Due to the ubiquity of modern technological devices for capturing handwriting signals, some researchers proposed the analysis of dynamic attributes involved in the process of producing a graphomotor impression. These techniques are termed as 'Procedural Analysis based Techniques'. Despite showing potential, such techniques may contradict with conventional test conduction protocols. In this thesis, we addressed both the issues of feature insufficiency and perseverance of original test conduction protocols by proposing a deep learning based solution. Convolutional neural networks are employed to extract enriched visual features from graphomotor samples provided by the domain experts. Feature enhancement is achieved by employing deformation-specific augmentation. Enhance features are then used to train supervised machine learning algorithms to estimate and classify various clinical deformations. The effectiveness of the proposed methodology is evaluated by analyzing its performance in modeling and estimation of several visual-motor and

visual-perceptual deformations that are commonly assessed by neuropsychological tests. Two case studies have been employed for this purpose.

- Identification of visual-motor deformations (tremor and micrographia) in graphomotor samples of elderly in order to detect signs of Parkinson's Disease (PD)
- Identification of visual-perceptual deformations (simplification, rotation, cohesion, perseveration, retrogression, fragmentation, motor incoordination, angulation, collision, overlapping and closure difficulty) in Bender Gestalt Test (BGT) samples of children in order to measure their visual-perceptual maturity

The analysis of the results in each scenario demonstrated the success of our proposed approach. Research contributions and key findings in each scenario are described in the subsequent sections.

6.1 Identification of Visual-Motor Deformations for Detection of Parkinson's Disease

In the first case study, we attempted to model visual-motor deformations associated with Parkinson's disease. As discussed earlier, the insufficiency of existing visual features for the characterization of PD related graphomotor deformations has resulted in a paradigm shift towards dynamic feature (kinematic, pressure, temporal, neuromotor and non-linear dynamics) analysis. This thesis proposed a method to extract rich visual features that effectively modeled PD related manifestations and provided comparable results to dynamic attributes. A popular benchmark database PaHaW [75] was employed for the evaluation of the proposed methodology. PaHaW database consists of multiple handwriting and drawing based tasks that enabled us to evaluate the impact of various templates on our proposed scheme. A method was proposed to visually assess the online signals by converting them into near-realistic offline images. The resultant images were then transformed to produce multiple representations (i.e. median residual and edge enhanced images) sensitive to fine imperfections resulting from handwriting influency and tremor (common Parkinsonian symptoms). An eight layered ConvNet architecture pre-trained on ImageNet, was employed for feature extraction from all three representations of input data. An early fusion technique was proposed to enhance features extracted from the multiple representations of data. Feature enhancement was evaluated both statistically and empirically. The enhanced features were then employed to train a binary classifier. To further enhance classification, a late fusion based approach was presented, where decisions from all task-wise classifiers were employed to predict the samples of potential PD patients. The proposed methodology achieved an overall accuracy of **83%** with a precision of **89%**. The proposed solution also achieved high sensitivity and specificity values of **84%** and **82%**, respectively. Key findings of the proposed case study are listed below:

- Extensive analysis of CNN-based visual features supports that they can effectively model visual-motor deformations present in graphomotor samples of PD patients and therefore, can be successfully employed for early detection of PD in elderly.

- The results of our evaluation showed that different tasks have a varying impact on the classification outcome. Due to this reason, the conventional approach adopted by the state-of-the art, where features from all tasks were combined and then employed to train a classifier, may result in degradation due to the negative impact of a particular task. On the contrary, our task-wise decision combining approach resulted in mitigation of this limitation.
- The proposed non-linear transformations of raw images can effectively capture fine imperfections resulting from various motor dysfunctions.
- Combining features extracted from multiple representations of the raw data significantly improved the classification results, thus indicating the success of our proposed feature enhancement technique.
- It is observed that conventional tasks like the spiral drawings, produced better results as compared to non-conventional handwriting tasks, supporting the fact that our proposed methodology works effectively on existing tasks and templates used by the clinical experts.

6.2 Identification of Visual-Perceptual Deformations for Scoring of Bender Gestalt Test

In the second case study, we attempted to model visual-perceptual deformations from drawing samples of children with learning disabilities. We evaluated the effectiveness of our proposed methodology by automating the scoring of an extensive neuropsychological test, Bender Gestalt Test (BGT) [22]. BGT scores various visual-perceptual deformations across nine different templates. Modeling human perception is a challenging task and techniques presented in the literature rely heavily on extensive heuristics to achieve this. A heuristic-based approach lacks robustness and is limited due to insufficiency to cover all possible scenarios. Consequently, we proposed an alternative AI-based solution for this purpose. Pre-trained ConvNets were used to model each deformation, while independent classifiers were employed for the classification purposes. The deformation classifiers were trained to be shape-independent and proved effective in identifying same deformations across multiple shapes (tasks). This is a noteworthy contribution of this thesis as previously CNN-based features have mostly been employed for shape recognition purposes only. Nonetheless, we not only employed CNNs to extract deformation-specific features across multiple shape classes but also used them to extract deformation-specific features from with in the same shape class. The proposed methodology was evaluated on samples provided by the domain experts. To the best of our knowledge, this was the only study that has addressed the problem of visual-perceptual deformation modeling and classification on such a scale. Previously, limited attempts with very fewer shapes and deformations were addressed. Our thesis presented identification of eleven indicators across nine different templates, (that has not been attempted previously). The proposed scheme achieved classification accuracies ranging from **79.1%** to **97.6%**. Key findings of the proposed case study are listed below:

- Although modeling of human perception is challenging, CNN-based visual features effectively modeled visual-perceptual deformations suggested by Lacks' scoring standard [31] across the nine BGT shapes.
- The proposed scheme was able to model deformation-specific intra-shape variations and inter-shape similarities effectively. This enabled the classification of same deformation across multiple shapes, as well as allowed the discrimination between deformed and non-deformed responses of the same template.
- The proposed deformation-specific augmentation proved effective in overcoming the scarcity of samples for a particular deformation, that is commonly observed in these scenarios.
- There is a need to create benchmark datasets for the evaluation of visual-perceptual deformations.

6.3 Limitations and Future Directions

This thesis presented a step towards an acceptable AI-based solution for the target domain experts. Our proposed conceptual model can provide a solid basis for an end-to-end system for the analysis of neuropsychological drawings that can be employed by clinical practitioners for standardization, validation of results and other diagnostic purposes. Since the work done in this area is yet to mature, there were several ideas that can be explored in future but have not been addressed in this thesis. Some of these are outlined below:

- This research has paved way for the integration of deep learning-based solutions in the domain of computerized analysis of neuropsychological tests, despite the scarcity of data. This is an important contribution of this thesis. Nonetheless, further exploration in this direction is required to assess the applicability of various deep learning models other than CNNs. For instance, due to the availability of online handwriting datasets, models like Recurrent Neural Networks (RNNs) [239] and their variants can be employed to characterize dynamic handwriting sequences.
- Currently, we have employed architectures pre-trained on ImageNet only. Despite achieving promising results, we can evaluate architectures that have been trained on datasets other than ImageNet, (i.e. much similar to the target data at hand).
- We have employed pre-trained ConvNets as feature extractors only. In future, the impact of fine-tuning can also be evaluated.
- The thesis exclusively evaluates the performance of static visual features. In future, the effectiveness of the proposed visual features in combination with other dynamic features can also be evaluated.

- In this thesis, we focused on tests that required subjects to copy a visual stimuli. In future, we intend to analyze projective tests like Draw-A-Person (DAP) [30] and House-Tree-Person (HTP) [152], as well.

As evident from the literature, computerized analysis of graphomotor-based neuropsychological tests is an emerging domain and requires more inter-disciplinary collaborations to provide improved solutions. The proposed study presents a direction worth exploring for various inter-disciplinary communities working to integrate technological solutions in health sector. Hence, we intend to continue our work in this direction. While the current research focused on application of intelligent techniques to identification of visual-motor and visual-perceptual disorders, the recent technological advancements can also be employed for subsequent rehabilitation of the subjects with these and related disorders. Augmented (virtual) reality systems, for instance, can be employed for learning and therapy processes for children with special needs. Such systems are known to not only enhance the motivation but also the learning experience of the subjects. In other words, AI-based systems can be employed for the complete pipeline from initial screening, diagnosis and rehabilitation, reducing the load on the domain experts.

Appendix A

Research Publications

A.1 Journal Publications

Following Impact Factor (IF) journal publications resulted from this research.

1. **M. Moetesum**, I. Siddiqi, N. Vincent, and F. Cloppet, "Assessing visual attributes of handwriting for prediction of neurological disorders—a case study on Parkinson's disease". Pattern Recognition Letters, 121:19–27, 2019. **(IF: 1.99)**

2. **M. Moetesum**, I. Siddiqi, S. Ehsan, and N. Vincent, "Deformation Modeling and Classification Using Deep Convolutional Neural Networks for Computerized Analysis of Neuropsychological Drawings". Neural Computing and Applications, 2020. **(IF: 4.66)**

A.2 Conference Publications

Following conference publications resulted from this research.

1. **M. Moetesum**, I. Siddiqi, U. Masroor, and C. Djeddi, "Automated scoring of Bender Gestalt test using image analysis techniques". In 13th International Conference on Document Analysis and Recognition (ICDAR), pages 666–670. IEEE, 2015.

2. **M. Moetesum**, I. Siddiqi, U. Masroor, N. Vincent, and F. Cloppet, "Segmentation and classification of offline hand drawn images for the bgt neuropsychological screening test". In 8th International Conference on Digital Image Processing (ICDIP), International Society for Optics and Photonics, 2016.

3. H. Nazar, **M. Moetesum**, S. Ehsan, I. Siddiqi, K. Khurshid, and K. D. M. Maier, "Classification of graphomotor impressions using convolutional neural networks: An application to automated neuro-psychological screening tests". In 14th International Conference on Document Analysis and Recognition (ICDAR), pages 432-437. IEEE, 2017.

4. **M. Moetesum**, O. Zeeshan, and I. Siddiqi, "Multi-object sketch segmentation using convolutional object detectors". In 10th International Conference on Graphics and Image Processing (ICGIP), International Society for Optics and Photonics, 2018.

5. **M. Moetesum**, I. Siddiqi, and N. Vincent, "Deformation Classification of Drawings for Assessment of Visual-Motor Perceptual Maturity". In 15th International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019.

Appendix B

Grants and Awards

B.1 Grants

Due to it's potential significance for the society in general and the research community in particular, this research has resulted in being awarded the following two research grants.

B.1.1 PAK-FRANCE PERIDOT Research Program - (2015-2017)

The project titled **Automated Detection of Neuropsychological Impairments from Image-Based Visuo-constructive Screening Tests** was awarded a research grant of **PKR 1.4 Million** by the Higher Education Commission (HEC) Pakistan and the French Ministry of Foreign Affairs, under the PERIDOT program. The project investigated the potential of hand-drawn shapes in early prediction of neuropsychological disorders. Computerized solutions were developed for screening of suspected cases using Bender Gestalt Test (BGT) as a case study. The project initiated in June 2015 in collaboration with LIPADE lab at Paris Descartes University, Paris, France. It was successfully completed in May 2017.

B.1.2 National Research Program for Universities (NRPU) - (2019-2021)

Recently, a project titled **Early Detection of Neurological Disorders through Computerized Analysis of Handwriting – An Application to Parkinson's Disease** has been awarded a grant of **PKR 2.04 Million** by the Higher Education Commission (HEC) under the National Research Program for Universities (NRPU). The project is aimed at early prediction of Parkinson's disease through computerized analysis of online handwriting. The project initiated in June 2019 and is still in progress.

B.1.3 Higher Education Commission Travel Grant - (2018)

The paper titled **Multi-Object Sketch Segmentation Using Convolutional Object Detectors** was selected for Oral Presentation at the **10th International Conference on Graphics and Image Processing (ICGIP), International Society for Optics and Photonics** in Chengdu, China in

November 2018. A travel grant of **PKR 0.24 Million** by the Higher Education Commission (HEC) Pakistan was awarded.

B.2 Awards

This research has been presented at various national and international platforms. During a presentation at the Doctoral Consortium organized during the IAPR International Conference on Document Analysis and Recognition (ICDAR), held in November 2017 in Kyoto, Japan, it was awarded the 'Best Poster Award' (certificate attached as Figure B.1). The research was declared *'The most innovative solution for addressing a major societal challenge'*.

Figure B.1: Certificate of 'Best Poster Award' at Doctoral Consortium in ICDAR 2017

Appendix C

Results of Pilot Study on BGT Scoring

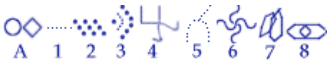
C.1 Lacks' Scoring Sheet

Standard Lacks' scoring sheet is shown in Figure C.1[1].

BENDER-GESTALT TEST Scoring Based on Lacks' Scoring System

Client: _____

Date: _____ Time to Complete: _____



Error	Description	A	1	2	3	4	5	6	7	8	Present
Rotation	Score if there is a rotation of 80° to 180° (including mirror-imaging) of the major axis of the whole figure (not a part of the figure). Do not score if *S* shifts the position of the card or paper and then draws the figure accurately.										
Overlapping Difficulty	Difficulty in reproducing the portions of the figures that should overlap. (a) Omission of the portions of the figure which overlap. (b) Simplification of figures only at the point of over lap. (c) Marked sketching or reworking only at the point of overlap. (d) Distortion of the figure at the point of overlap. (e) Figures overlap at the wrong place. (f) Failure of figures to overlap. DO NOT SCORE – parts of figures more than 1/8 in. apart, score Simplification.	■	■	■	■	■	■			■	
Simplification	Score if the figure is drawn in a simplified or easier from that is not more primitive from a maturational point of view, from the stimulus. (a) Circles for dots on figure 1. (b) Nonoverlapping parts. (c) Joining parts of figures are more than 1/8 inch apart. (d) Very simplified drawing. DO NOT SCORE – (a) figures less than 1/8 inch apart, score Close Difficulty. (b) Curves substituted for angles, not an error.										
Fragmentation	Score if the figure is broken up into parts destroying the gestalt or if the figure is incomplete (unless *S* refuses to draw the entire figure).										
Retrogression	Substitution of a more primitive gestalt form than the stimulus. (a) Loops for circles (if persistent). (b) Dashes for dots (if extreme and persistent). (c) Triangle for diamond or hexagon. (d) Square for diamond. (e) Rectangle for hexagon. DO NOT SCORE – Do not score if curves are substituted for angles or angulation of bottom of hexagon on figure 7 is omitted.					■		■			
Perseveration	There are 2 kinds of Perseveration errors. If both occur, this error is still only scored once. **TYPE A:** Inappropriate substitution of the features of a preceding stimulus, such as replacing the circles of figure 2 with the dots of figure 1 (must have made dots, not circles on figure 1); replacing the dots of figures 3 & 5 with the circles of figure 2 (must have made circles on figure 2 and dots on 1). **TYPE B:** Intradesign perseveration on continuing to draw a figure beyond the limits called for by the stimulus. For figure 1, 14 or more dots must be present, for figure 2, 13 or more columns of circles.	■				■		■	■	■	
Collision or Collision Tendency	One figure is drawn as touching or overlapping another figure (collision) or is drawn within 1/4 inch or less of another figure but does not touch (collision tendency).										
Impotence	Behavioral or verbal expressions of inability to draw a figure correctly (often accompanied by statements such as "I know this drawing is not right but I just can't make it right"). (a) Repetitious drawings or numerous erasures of figures with similar inaccuracies. (b) *S* realizes that an error has been made and tries to correct it unsuccessfully or expresses inability to correct it. DO NOT SCORE – Second attempt that corrects an error.										
Closure Difficulty	Difficulty in getting the joining parts of figures together or getting adjacent parts of a figure to touch. If figures are more than 1/8 inch apart at joining point, score Simplification.		■	■	■		■	■		■	
Motor Incoordination	Irregular (tremor-like) lines, especially with heavy pressure. Behavioral observations are important for scoring this error. Be sure *S* is drawing on a smooth surface. (a) Marked and persistent gasp, overlap, redrawing, sketching, erasures, increased pressure at points where parts of the design join one another. DO NOT SCORE – Parts of figures are more than 1/8 inch apart, score Simplification.										
Angulation	Severe difficulty in reproducing the angulation of figures. (a) Failure to reproduce angulation of a figure. (b) Angulation of the whole figure 45° to 80° rather than parts of a figure (but not greater than 80°, which would be rotation). (c) Variability of the angulation of more than half the rows of circles of figure 2. DO NOT SCORE – (a) Figure 3 should be scored leniently because its angulation is especially hard to reproduce. (b) Reversal of angulation on figure 2, score Rotation.	■	■			■	■	■	■	■	
Cohesion	Isolated decrease or increase in size of figures. Score very conservatively. This error is most frequently overscored. (a) Decrease in the size of *part* of a figure by more than 1/3 of the dimensions used in the rest of the figure. (b) Increase or decrease in the size of a figure by 1/3 of the dimensions used in the *other drawings (not* compared to the size of the stimulus cards). Exclude parts of drawing that are larger due to Perseveration.										
Time	Score if total time is greater than 15 minutes.	■	■	■	■	■	■	■	■	■	

Points:
1) Score **presence** of error, not frequency, and score conservatively. For example, even if Rotation is scored for each figure, score only 1 in the **Present** column.
2) If the subject rotates the card or paper and then draws correctly, it is correct.
3) Generally, 3 or fewer errors indicates an *absence* of visuoconstructive deficits or brain impairment; 4 errors is a *borderline* score; and 5 or 6 errors provide *some evidence* for brain impairment. The greater the number of errors, the greater the evidence for some type of brain impairment: *strong evidence* with 7 or 8 errors and *very strong evidence* with 9 to 12 errors. Five or more errors is serious, but not if the subject is lazy, impulsive, unmotivated, or uncooperative.

Total Score

Figure C.1: Lacks' scoring sheet for BGT analysis

[1] http://www.labh.it/wp_sancipriano/wp-content/uploads/2016/03/4-BenderScore.pdf

C.2 Ground Truth Labeling Tool

Scored samples acquired from the clinical practitioner were digitized by scanning. A ground truth labeling tools was developed to score the digitized samples. The user friendly interface (Figure C.2) of the tool enabled the domain expert to score the samples easily.

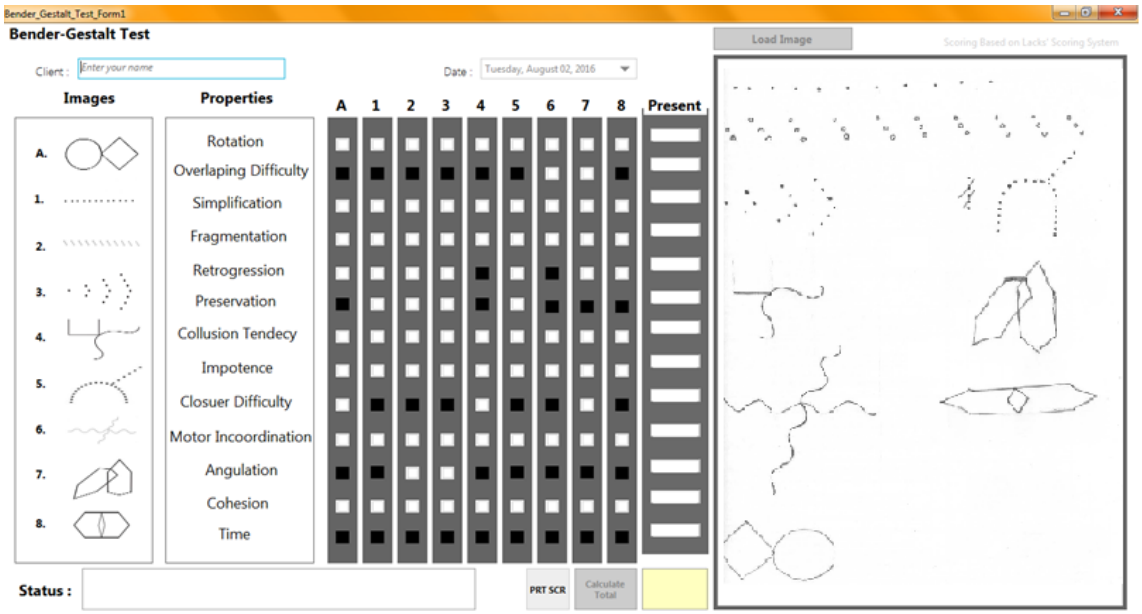


Figure C.2: Ground truth labeling tool

C.3 Results on Hand-Crafted Features and Heuristic-based Approach

Table C.1 presents the results of the experiments carried out in the pilot study ('-' refers to 'not applicable' while 'x' refers to 'not implemented'). The scores produced by the system for the BGT shapes and corresponding errors considered in the study, were compared with the scores assigned through human inspection (obtained via labeling tool). A total of 152 figures produced by 18 different subjects were used to evaluate the performance of the proposed system. Figure C.3 provides a visual illustration of the application of shape-based geometric features to score the drawings.

Table C.1: Results of Automated Scoring

Figure Number	Total Figures	Number of Correctly Scored Figures					
		Simplification	Overlap	Rotation	Presrvation	Closure	Cohesion
A	18	-	-	15	-	15	17
1	18	14	-	18	16	-	15
2	18	16	-	18	17	-	15
3	18	15	-	x	-	-	14
4	16	-	-	x	-	15	14
5	18	16	-	x	-	-	14
6	12	-	6	x	-	-	11
7	18	-	14	x	-	16	16
8	16	-	-	15	-	-	14

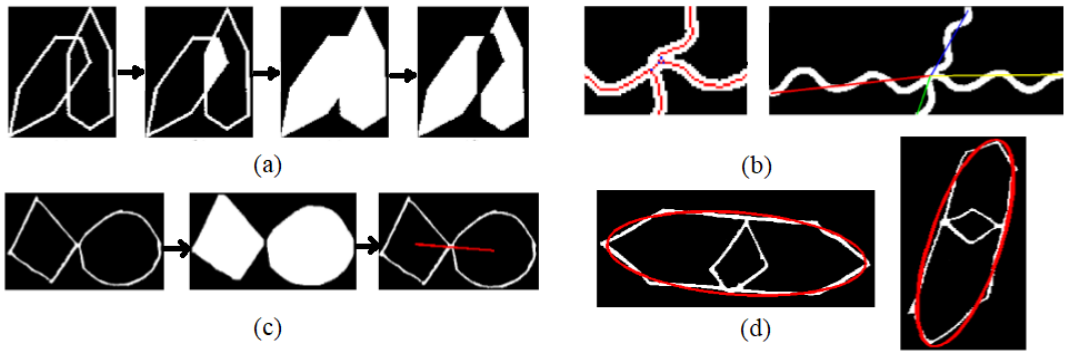


Figure C.3: Shape-based geometric features extracted from BGT (a) Shape 7 for overlapping difficulty (b) Shape 6 for overlapping difficulty (c) Shape A for rotation (d) Shape 8 for rotation [27]

Appendix D

Results of Automated Segmentation of BGT Shapes

D.1 Automated Segmentation Using Gestalt Theory based Heuristic Approach

While testing the proposed heuristic-based technique, 43 out of 72 BGT figures were correctly segmented. It was observed that the most challenging shapes for segmentation are the ones consisting of disconnected primitive components like circles/dots/dashes. These shapes had been grouped as Group-C and comprised of BGT shapes 1, 2, 3, and 5. Table D.1 gives the details of segmentation rates for individual shapes.

Table D.1: Results of Automated Segmentation Using Gestalt Theory based Heuristic Technique

Figure	Total	Extracted	Rate
A	18	18	100%
1	18	8	44%
2	18	6	33%
3	18	13	72%
4	18	15	83%
5	18	16	89%
6	18	17	94%
7	18	18	100%
8	18	18	100%

D.2 Automated Segmentation Using Convolutional Object Detectors

Automated segmentation of BGT shapes has also been performed by using state-of-the-art convolutional object detectors. Three commonly used metrics i.e. 'Precision', 'Recall' and 'F-Measure' are used to evaluate the performance of each network on a dataset of 405 shapes. Results are reported

in Table D.2. To ensure a fair comparison, various parameters (like input image resolution, ground truth encoding, matching function, training epochs, loss function and mini-batch size etc.) are kept constant across each network.

Table D.2: Results of Automated Segmentation Using Convolutional Object Detectors

Meta-Architecture	**Feature Extractor**	**Precision(%)**	**Recall(%)**	**F-Measure(%)**
SSD	InceptionV1	88.70%	37.33%	52.54%
SSD	InceptionV2	91.71%	70.13%	79.48%
Faster R-CNN	ResNet50	89.38%	95.03%	92.12%
Faster R-CNN	**ResNet101**	**92.93%**	**95.24%**	**94.07%**
R-FCN	ResNet101	89.52%	94.79%	92.08%

Appendix E

Results of Automated Shape Recognition of BGT Shapes

E.1 BGT Samples

As discussed earlier, BGT consists of nine shapes of varying degree of complexity. Figure E.1 shows two samples drawn by different subjects.

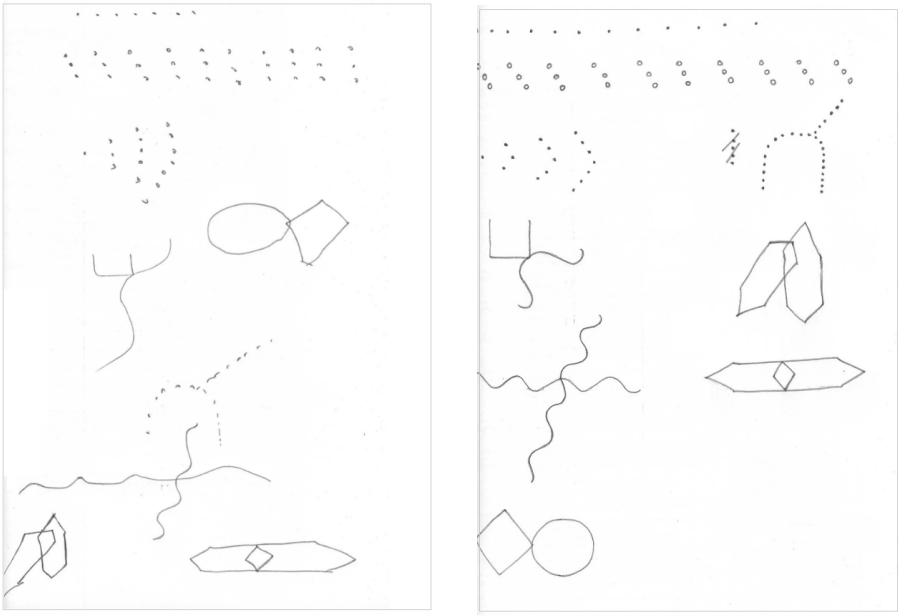


Figure E.1: Two drawn responses of BGT test

E.2 Automated Shape Recognition Using Shape Context Descriptors

18 samples of each shape class were employed to assess the effectiveness of the shape context-based recognition scheme. One image of each of the 9 classes was used as reference set and the remaining

17 images as the test set. Out of the 153 drawings presented to the system, 129 were correctly recognized achieving an overall classification rate of 84.31%. The results are as summarized in Table E.1 while the detailed confusion matrix is presented in Table E.2.

Table E.1: Classification rates on drawings from 17 subjects

Images	Recognized	Classification Rate
153	129	84.31%

Table E.2: Confusion matrix of 9 drawing classes for 17 test subjects

Class	**A**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
A	**11**	0	2	0	2	1	0	0	1
1	0	**7**	1	0	0	0	1	7	1
2	0	0	**16**	0	0	0	1	0	0
3	0	0	0	**17**	0	0	0	0	0
4	0	0	0	0	**17**	0	0	0	0
5	0	0	0	1	0	**14**	2	0	0
6	0	0	0	0	0	0	**14**	0	3
7	0	0	0	0	0	0	0	**17**	0
8	0	0	0	0	0	0	0	1	**16**

Bibliography

[1] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4):230–243, 2017.

[2] Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. Medex: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24, 2010.

[3] Daniel B Neill. New directions in artificial intelligence for public health surveillance. *IEEE Intelligent Systems*, 27(1):56–59, 2012.

[4] Guilan Kong, Dong-Ling Xu, and Jian-Bo Yang. Clinical decision support systems: a review on knowledge representation and inference under uncertainties. *International Journal of Computational Intelligence Systems*, 1(2):159–167, 2008.

[5] D Douglas Miller and Eric W Brown. Artificial intelligence in medical practice: the question to the answer? *The American journal of medicine*, 131(2):129–133, 2018.

[6] Philip Schatz and Jeffrey Browndyke. Applications of computer-based neuropsychological assessment. *The Journal of head trauma rehabilitation*, 17(5):395–410, 2002.

[7] Carolyn M Parsey and Maureen Schmitter-Edgecombe. Applications of technology in neuropsychological assessment. *The Clinical Neuropsychologist*, 27(8):1328–1361, 2013.

[8] Kenneth M Heilman and Edward Ed Valenstein. *Clinical neuropsychology*. Oxford University Press, 2003.

[9] Jeremy Hall, Ronan E O'Carroll, and Chris D Frith. 7 - neuropsychology. In Eve C. Johnstone, David Cunningham Owens, Stephen M. Lawrie, Andrew M. McIntosh, and Michael Sharpe, editors, *Companion to Psychiatric Studies (Eighth Edition)*, pages 121 – 140. Churchill Livingstone, St. Louis, eighth edition edition, 2010.

[10] Eduardo Tolosa, Gregor Wenning, and Werner Poewe. The diagnosis of parkinson's disease. *The Lancet Neurology*, 5(1):75–86, 2006.

[11] Stéphane Lehéricy, Michael A Sharman, Clarisse Longo Dos Santos, Raphaël Paquin, and Cecile Gallea. Magnetic resonance imaging of the substantia nigra in parkinson's disease. *Movement disorders*, 27(7):822–830, 2012.

[12] Muriel Deutsch Lezak, Diane B Howieson, David W Loring, Jill S Fischer, et al. *Neuropsychological assessment*. Oxford University Press, USA, 2004.

[13] Ralph M Reitan and Deborah Wolfson. Theoretical, methodological, and validational bases of the halstead-reitan neuropsychological test battery. 2004.

[14] Charles J Golden and Shawna M Freshwater. Luria-nebraska neuropsychological battery. In *Understanding psychological assessment*, pages 59–75. Springer, 2001.

[15] JM Ziviani and M Wallen. *The development of graphomotor skills*. Mosby Elsevier, 2006.

[16] Bouwien CM Smits-Engelsman and Gerard P Van Galen. Dysgraphia in children: Lasting psychomotor deficiency or transient developmental delay? *Journal of experimental child psychology*, 67(2):164–184, 1997.

[17] Naomi Weintraub and Steve Graham. The contribution of gender, orthographic, finger function, and visual-motor processes to the prediction of handwriting status. *The Occupational Therapy Journal of Research*, 20(2):121–140, 2000.

[18] Alastair D Smith. On the use of drawing tasks in neuropsychological assessment. *Neuropsychology*, 23(2):231, 2009.

[19] Laurence Likforman-Sulem, Anna Esposito, Marcos Faundez-Zanuy, Stéphan Clémençon, and Gennaro Cordasco. Emothaw: A novel database for emotional state recognition from handwriting and drawing. *IEEE Transactions on Human-Machine Systems*, 47(2):273–284, 2017.

[20] Min-Sup Shin, Sun-Young Park, Se-Ran Park, Soon-Ho Seol, and Jun Soo Kwon. Clinical and empirical applications of the rey–osterrieth complex figure test. *Nature protocols*, 1(2):892–899, 2006.

[21] Brian J Mainland and Kenneth I Shulman. Clock drawing test. In *Cognitive Screening Instruments*, pages 79–109. Springer, 2013.

[22] Lauretta Bender. A visual motor gestalt test and its clinical use. *Research Monographs, American Orthopsychiatric Association*, 1938.

[23] Annie W Hsu, Panida A Piboolnurak, Alicia G Floyd, Qiping P Yu, James E Wraith, Marc C Patterson, and Seth L Pullman. Spiral analysis in niemann-pick disease type c. *Movement Disorders*, 24(13):1984–1990, 2009.

[24] Jack A Naglieri, Timothy J McNeish, and N Achilles. Draw a person test. *Tools of the Trade: A Therapist's Guide to Art Therapy Assessments*, page 124, 2004.

[25] Daniel R Coates, Johan Wagemans, and Bilge Sayim. Diagnosing the periphery: Using the rey–osterrieth complex figure drawing test to characterize peripheral visual function. *i-Perception*, 8(3), 2017.

[26] Catherine C Price, Holly Cunningham, Nicole Coronado, Alana Freedland, Stephanie Cosentino, Dana L Penney, Alfio Penisi, Dawn Bowers, Michael S Okun, and David J Libon. Clock drawing in the montreal cognitive assessment: recommendations for dementia assessment. *Dementia and geriatric cognitive disorders*, 31(3):179–187, 2011.

[27] Momina Moetesum, Imran Siddiqi, Uzma Masroor, and Chawki Djeddi. Automated scoring of bender gestalt test using image analysis techniques. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 666–670. IEEE, 2015.

[28] Momina Moetesum, Imran Siddiqi, Nicole Vincent, and Florence Cloppet. Assessing visual attributes of handwriting for prediction of neurological disorders—a case study on parkinson's disease. *Pattern Recognition Letters*, 121:19–27, 2019.

[29] Christiane Lange-Küttner, Enno Küttner, and Marta Chromekova. Deterioration and recovery of dap iq scores in the repeated assessment of the naglieri draw-a-person (dap) test in 6-to 12-year-old children. *Psychological assessment*, 26(1):297, 2014.

[30] Florence Laura Goodenough. Measurement of intelligence by drawings. 1926.

[31] Patricia Lacks. *Bender Gestalt screening for brain dysfunction*. John Wiley & Sons Inc, 1999.

[32] Gary G Brannigan, Scott L Decker, and David H Madsen. Innovative features of the bender-gestalt ii and expanded guidelines for the use of the global scoring system. *Bender Visual-Motor Gestalt Test, Second Edition Assessment Service Bulletin*, 1, 2004.

[33] Amir M Poreh. Forum the quantified process approach: An emerging methodology to neuropsychological assessment. *The Clinical Neuropsychologist*, 14(2):212–222, 2000.

[34] Melissa Ogden-Epker and C Munro Cullum. Quantitative and qualitative interpretation of neuropsychological data in the assessment of temporal lobectomy candidates. *The Clinical Neuropsychologist*, 15(2):183–195, 2001.

[35] Pascal Derkinderen, Sophie Dupont, Jean-Sébastien Vidal, François Chedru, and Marie Vidailhet. Micrographia secondary to lenticular lesions. *Movement disorders*, 17(4):835–837, 2002.

[36] Judie Walton. Handwriting changes due to aging and parkinson's syndrome. *Forensic science international*, 88(3):197–214, 1997.

[37] Elisabetta Ambron, Luca Piretti, Alberta Lunardelli, and H Coslett. Closing-in behavior and parietal lobe deficits: three single cases exhibiting different manifestations of the same behavior. *Frontiers in psychology*, 9:1617, 2018.

[38] Massimiliano Conson, Claudia Nuzzaci, Laura Sagliano, and Luigi Trojano. Relationship between closing-in and spatial neglect: a case study. *Cognitive and Behavioral Neurology*, 29(1):44–50, 2016.

[39] Federica Molteni, Debora Traficante, Francesca Ferri, and Valeria Isella. Cognitive profile of patients with rotated drawing at copy or recall: A controlled group study. *Brain and cognition*, 85:286–290, 2014.

[40] Wayne J Camara, Julie S Nathan, and Anthony E Puente. Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice*, 31(2):141, 2000.

[41] G Groth-Marnat, F Strub, R Black, and A Luria. *Neuropsychological Assessment in Clinical Practice: A*. New York: Wiley and Sons, 2000.

[42] Gary Groth-Marnat. *Handbook of psychological assessment*. John Wiley & Sons, 2009.

[43] Stephen L Smith and Darren L Hiller. Image analysis of neuropsychological test responses. In *Medical Imaging 1996: Image Processing*, volume 2710, pages 904–915. International Society for Optics and Photonics, 1996.

[44] RO Canham, Stephen L Smith, and Andrew M Tyrrell. Automated scoring of a neuropsychological test: the rey osterrieth complex figure. In *Proceedings of the 26th Euromicro Conference. EUROMICRO 2000. Informatics: Inventing the Future*, volume 2, pages 406–413. IEEE, 2000.

[45] RO Canham, SL Smith, and AM Tyrrell. Location of structural sections from within a highly distorted complex line drawing. *IEE Proceedings-Vision, Image and Signal Processing*, 152(6):741–749, 2005.

[46] Mr C Fairhurst, T Linnell, Stephanie Glenat, RM Guest, Laurent Heutte, and Thierry Paquet. Developing a generic approach to online automated analysis of writing and drawing tests in clinical patient profiling. *Behavior Research Methods*, 40(1):290–303, 2008.

[47] Mohamed Bennasar, Rossitza Setchi, Antony Bayer, and Yulia Hicks. Feature selection based on information theory in the clock drawing test. *Procedia Computer Science*, 22:902–911, 2013.

[48] Mohamed Bennasar, Rossitza Setchi, Yulia Hicks, and Antony Bayer. Cascade classification for diagnosing dementia. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2535–2540. IEEE, 2014.

[49] Clayton R Pereira, Danillo R Pereira, Francisco A da Silva, Christian Hook, Silke AT Weber, Luis AM Pereira, and Joao P Papa. A step towards the automated diagnosis of parkinson's disease: Analyzing handwriting movements. In *2015 IEEE 28th international symposium on computer-based medical systems*, pages 171–176. IEEE, 2015.

[50] Clayton R Pereira, Danilo R Pereira, Francisco A Silva, João P Masieiro, Silke AT Weber, Christian Hook, and João P Papa. A new computer vision-based approach to aid the diagnosis of parkinson's disease. *Computer methods and programs in biomedicine*, 136:79–88, 2016.

[51] Zainab Harbi, Yulia Hicks, and Rossitza Setchi. Clock drawing test digit recognition using static and dynamic features. *Procedia Computer Science*, 96:1221–1230, 2016.

[52] Zainab Harbi, Yulia Hicks, and Rossitza Setchi. Clock drawing test interpretation system. *Procedia computer science*, 112:1641–1650, 2017.

[53] MC Fairhurst and Stephen L Smith. Application of image analysis to neurological screening through figure-copying tasks. *International journal of bio-medical computing*, 28(4):269–287, 1991.

[54] Stephen L Smith and Basilio R Cervantes. Dynamic feature analysis of vector-based images for neuropsychological testing. In *Medical Imaging 1998: Physiology and Function from Multidimensional Images*, volume 3337, pages 304–313. International Society for Optics and Photonics, 1998.

[55] A Garbi, SL Smith, D Heseltine, and P Thomson. Automated and enhanced assessment of unilateral visual neglect. 1999.

[56] Céline Rémi, Carl Frélicot, and Pierre Courtellemont. Automatic analysis of the structuring of children's drawings and writing. *Pattern Recognition*, 35(5):1059–1069, 2002.

[57] Samuel Chindaro, Richard Guest, Michael Fairhurst, and Jonathan Potter. Assessing visuo-spatial neglect through feature selection from shape drawing performance and sequence analysis. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(07):1253–1266, 2004.

[58] Perla Werner, Sara Rosenblum, Gady Bar-On, Jeremia Heinik, and Amos Korczyn. Handwriting process variables discriminating mild alzheimer's disease and mild cognitive impairment. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 61(4):P228–P236, 2006.

[59] Alex Hazell and Stephen L Smith. Towards an objective assessment of alzheimer's disease: The application of a novel evolutionary algorithm in the analysis of figure copying tasks. In *Proceedings of the 10th annual conference companion on Genetic and evolutionary computation*, pages 2073–2080. ACM, 2008.

[60] Stephen L Smith and Michael A Lones. Implicit context representation cartesian genetic programming for the assessment of visuo-spatial ability. In *2009 IEEE Congress on Evolutionary Computation*, pages 1072–1078. IEEE, 2009.

[61] Jeremia Heinik, Perla Werner, Tal Dekel, Ilya Gurevitz, and Sara Rosenblum. Computerized kinematic analysis of the clock drawing task in elderly people with mild major depressive disorder: an exploratory study. *International psychogeriatrics*, 22(3):479–488, 2010.

[62] Puspa Inayat Khalid, Jasmy Yunus, Robiah Adnan, Mokhtar Harun, Rubita Sudirman, and Nasrul Humaimi Mahmood. The use of graphic rules in grade one to help identify children at risk of handwriting difficulties. *Research in developmental disabilities*, 31(6):1685–1693, 2010.

[63] Maria Francesca De Pandis, Manuela Galli, Sara Vimercati, Veronica Cimolin, Maria Vittoria De Angelis, and Giorgio Albertini. A new approach for the quantitative evaluation of the clock drawing test: preliminary results on subjects with parkinson's disease. *Neurology research international*, 2010, 2010.

[64] Ney Renau-Ferrer and Céline Rémi. A generic approach for recognition and structural modelling of drawers' sketching gestures. In *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, page 1. Citeseer, 2011.

[65] Manuela Galli, Sara Laura Vimercati, Giacomo Stella, Giorgia Caiazzo, Federica Norveti, Francesca Onnis, Chiara Rigoldi, and Giorgio Albertini. A new approach for the quantitative evaluation of drawings in children with learning disabilities. *Research in developmental disabilities*, 32(3):1004–1010, 2011.

[66] François Beuvens and Jean Vanderdonckt. Usigesture: An environment for integrating pen-based interaction in user interface development. In *2012 Sixth International Conference on Research Challenges in Information Science (RCIS)*, pages 1–12. IEEE, 2012.

[67] SL Vimercati, M Galli, MF De Pandis, A Ancillao, G Stella, and G Albertini. Quantitative evaluation of graphic gesture in subjects with parkinson's disease and in children with learning disabilities. *Gait & Posture*, (35):S23–S24, 2012.

[68] Narges Tabatabaey-Mashadi, Rubita Sudirman, and Puspa Inayat Khalid. An evaluation of children's structural drawing strategies. *Jurnal Teknologi*, 61(2), 2012.

[69] Peter Drotár, Jiří Mekyska, Zdeněk Smékal, Irena Rektorová, Lucia Masarová, and Marcos Faundez-Zanuy. Prediction potential of different handwriting tasks for diagnosis of parkinson's. In *2013 E-Health and Bioengineering Conference (EHB)*, pages 1–4. IEEE, 2013.

[70] Peter Drotár, Jiří Mekyska, Irena Rektorová, Lucia Masarová, Zdeněk Smékal, and Marcos Faundez-Zanuy. A new modality for quantitative evaluation of parkinson's disease: In-air movement. In *13th IEEE International Conference on BioInformatics and BioEngineering*, pages 1–4. IEEE, 2013.

[71] Peter Drotár, Jiří Mekyska, Irena Rektorová, Lucia Masarová, Zdeněk Smékal, and Marcos Faundez-Zanuy. Decision support framework for parkinson's disease based on novel handwriting markers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(3):508–516, 2014.

[72] Réjean Plamondon, Christian O'Reilly, and Claudéric Ouellet-Plamondon. Strokes against stroke—strokes for strides. *Pattern Recognition*, 47(3):929–944, 2014.

[73] Christian O'Reilly, Réjean Plamondon, and Louise-Hélène Lebrun. Linking brain stroke risk factors to human movement features for the development of preventive tools. *Frontiers in aging neuroscience*, 6:150, 2014.

[74] Narges Tabatabaey-Mashadi, Rubita Sudirman, Richard M Guest, and Puspa Inayat Khalid. Analyses of pupils' polygonal shape drawing strategy with respect to handwriting performance. *Pattern Analysis and Applications*, 18(3):571–586, 2015.

[75] Peter Drotár, Jiří Mekyska, Irena Rektorová, Lucia Masarová, Zdeněk Smékal, and Marcos Faundez-Zanuy. Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease. *Artificial intelligence in Medicine*, 67:39–46, 2016.

[76] Catherine Taleb, Maha Khachab, Chafic Mokbel, and Laurence Likforman-Sulem. Feature selection for an improved parkinson's disease identification based on handwriting. In *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, pages 52–56. IEEE, 2017.

[77] Josep Garre-Olmo, Marcos Faúndez-Zanuy, Karmele López-de Ipiña, Laia Calvó-Perxas, and Oriol Turró-Garriga. Kinematic and pressure features of handwriting and drawing: preliminary results between patients with mild cognitive impairment, alzheimer disease and healthy controls. *Current Alzheimer research*, 14(9):960–968, 2017.

[78] Jan Mucha, Vojtech Zvoncak, Zoltan Galaz, Marcos Faundez-Zanuy, Jiri Mekyska, Tomas Kiska, Zdenek Smekal, Lubos Brabenec, Irena Rektorova, and Karmele Lopez-de Ipina. Fractional derivatives of online handwriting: A new approach of parkinsonic dysgraphia analysis. In *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, pages 1–4. IEEE, 2018.

[79] Gennaro Vessio. Dynamic handwriting analysis for neurodegenerative disease assessment: A literary review. *Applied Sciences*, 9(21):4666, 2019.

[80] Mounim A El-Yacoubi, Sonia Garcia-Salicetti, Christian Kahindo, Anne-Sophie Rigaud, and Victoria Cristancho-Lacroix. From aging to early-stage alzheimer's: Uncovering handwriting multimodal behaviors by semi-supervised learning and sequential representation learning. *Pattern Recognition*, 86:112–133, 2019.

[81] Rosa Senatore, Antonio Della Cioppa, and Angelo Marcelli. Automatic diagnosis of neurodegenerative diseases: An evolutionary approach for facing the interpretability problem. *Information*, 10(1):30, 2019.

[82] Peter Drotár, Jiří Mekyska, Irena Rektorová, Lucia Masarová, Zdenek Smékal, and Marcos Faundez-Zanuy. Analysis of in-air movement in handwriting: A novel marker for parkinson's disease. *Computer methods and programs in biomedicine*, 117(3):405–411, 2014.

[83] Donato Impedovo, Giuseppe Pirlo, and Gennaro Vessio. Dynamic handwriting analysis for supporting earlier parkinson's disease diagnosis. *Information*, 9(10):247, 2018.

[84] Peter Drotár, Jiří Mekyska, Irena Rektorová, Lucia Masarová, Zdeněk Smékal, and Marcos Faundez-Zanuy. Decision support framework for parkinson's disease based on novel handwriting markers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(3):508–516, 2015.

[85] Max Wertheimer. Gestalt theory. 1938.

[86] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[87] Sonja von Campenhausen, Bernhard Bornschein, Regina Wick, Kai Bötzel, Cristina Sampaio, Werner Poewe, Wolfgang Oertel, Uwe Siebert, Karin Berger, and Richard Dodel. Prevalence and incidence of parkinson's disease in europe. *European Neuropsychopharmacology*, 15(4):473–490, 2005.

[88] Lonneke ML De Lau and Monique MB Breteler. Epidemiology of parkinson's disease. *The Lancet Neurology*, 5(6):525–535, 2006.

[89] Yue Zhou, Mary E Jenkins, Michael D Naish, and Ana Luisa Trejos. The measurement and analysis of parkinsonian hand tremor. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 414–417. IEEE, 2016.

[90] José L Contreras-Vidal, Patricia Poluha, Hans-Leo Teulings, and George E Stelmach. Neural dynamics of short and medium-term motor control effects of levodopa therapy in parkinson's disease. *Artificial Intelligence in Medicine*, 13(1-2):57–79, 1998.

[91] Arend WA van Gemmert, Hans-Leo Teulings, and George E Stelmach. The influence of mental and motor load on handwriting movements in parkinsonian patients. *Acta psychologica*, 100(1-2):161–175, 1998.

[92] Sara Rosenblum, Margalit Samuel, Sharon Zlotnik, Ilana Erikh, and Ilana Schlesinger. Handwriting as an objective tool for parkinson's disease diagnosis. *Journal of neurology*, 260(9):2357–2361, 2013.

[93] Eun-Joo Kim, Byung Hwa Lee, Key Chung Park, Won Yong Lee, and Duk L Na. Micrographia on free writing versus copying tasks in idiopathic parkinson's disease. *Parkinsonism & related disorders*, 11(1):57–63, 2005.

[94] Hui-Ing Ma, Wen-Juh Hwang, Shao-Hsia Chang, and Tsui-Ying Wang. Progressive micrographia shown in horizontal, but not vertical, writing in parkinson's disease. *Behavioural neurology*, 27(2):169–174, 2013.

[95] Seth L Pullman. Spiral analysis: a new technique for measuring tremor with a digitizing tablet. *Movement Disorders*, 13(S3):85–89, 1998.

[96] Christel Bidet-Ildei, Pierre Pollak, Sonia Kandel, Valérie Fraix, and Jean-Pierre Orliaguet. Handwriting in patients with parkinson disease: Effect of l-dopa and stimulation of the sub-thalamic nucleus on motor anticipation. *Human movement science*, 30(4):783–791, 2011.

[97] Alfredo Berardelli, JC Rothwell, PD Thompson, and M Hallett. Pathophysiology of bradykinesia in parkinson's disease. *Brain*, 124(11):2131–2146, 2001.

[98] Michael P Broderick, Arend WA Van Gemmert, Holly A Shill, and George E Stelmach. Hypometria and bradykinesia during drawing movements in individuals with parkinson's disease. *Experimental brain research*, 197(3):223–233, 2009.

[99] Sonia Kandel, Jean-Pierre Orliaguet, and Louis-Jean Boe. Detecting anticipatory events in handwriting movements. *Perception*, 29(8):953–964, 2000.

[100] Qi He, Kuiyu Chang, and Ee-Peng Lim. Anticipatory event detection via sentence classification. In *2006 IEEE International Conference on Systems, Man and Cybernetics*, volume 2, pages 1143–1148. IEEE, 2006.

[101] AL Smiley-Oyen, KA Lowry, and JP Kerr. Planning and control of sequential rapid aiming in adults with parkinson's disease. *Journal of Motor Behavior*, 39(2):103–114, 2007.

[102] Khalid Ali and Huw R Morris. Parkinson's disease: chameleons and mimics. *Practical neurology*, 15(1):14–25, 2015.

[103] Alban Letanneux, Jeremy Danna, Jean-Luc Velay, François Viallet, and Serge Pinto. From micrographia to parkinson's disease dysgraphia. *Movement Disorders*, 29(12):1467–1475, 2014.

[104] Peter Drotár, Jiří Mekyska, Zdeněk Smékal, Irena Rektorová, Lucia Masarová, and Marcos Faundez-Zanuy. Contribution of different handwriting modalities to differential diagnosis of parkinson's disease. In *2015 IEEE International Symposium on Medical Measurements and Applications (MeMeA) Proceedings*, pages 344–348. IEEE, 2015.

[105] Clayton R Pereira, Silke AT Weber, Christian Hook, Gustavo H Rosa, and João P Papa. Deep learning-aided parkinson's disease diagnosis from handwritten dynamics. In *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 340–346. Ieee, 2016.

[106] S Sheikhi. Clinical use of bender-gestalt test in brain lesions diagnosis and its comparison with magnetic resonance imaging (mri). *J Urmia Nurs Midwifery Fac*, 5(1):15–21, 2007.

[107] Acácia Aparecida Angeli dos Santos and Lília Maíse de Jorge. Bender test with dyslexics: comparison of two systems of punctuation. *Psico-USF*, 12(1):13–21, 2007.

[108] Ryana A Allen and Scott L Decker. Utility of the bender visual-motor gestalt test—second edition in the assessment of attention-deficit/hyperactivity disorder'. *Perceptual and motor skills*, 107(3):663–677, 2008.

[109] Roselaine Berenice Ferreira, Cristiane Friedrich Feil, and Maria Lucia Tiellet Nunes. Bender visual-motor gestalt test in the children's clinical assessment. *Psico-USF*, 14(2):185–192, 2009.

[110] Livia de Freitas Keppeke, Isa de Pádua Cintra, and Teresa Helena Schoen. Bender visual-motor gestalt test in adolescents: Relationship between visual-motor development and the tanner stages. *Perceptual and motor skills*, 117(1):257–275, 2013.

[111] Elizabeth M Koppitz. The bender gestalt test for young children. 1964.

[112] James JA Cavanaugh. Preventing reading failure: Prediction, diagnosis, intervention, by jeannette jansky and katrina dehirsch., 1973.

[113] Max L Hutt and Sylvia Monheit. Effectiveness of the hutt adaptation of the bender-gestalt test configuration scale in differentiating emotionally disturbed adolescents. *Psychological reports*, 56(2):439–443, 1985.

[114] Réjean Plamondon and Sargur N Srihari. Online and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1):63–84, 2000.

[115] Alex Graves and Jürgen Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in neural information processing systems*, pages 545–552, 2009.

[116] David Doermann, Karl Tombre, et al. *Handbook of document image processing and recognition*. Springer, 2014.

[117] Daniel Keysers, Thomas Deselaers, Henry A Rowley, Li-Lun Wang, and Victor Carbune. Multi-language online handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1180–1194, 2016.

[118] Youssouf Chherawala, Partha Pratim Roy, and Mohamed Cheriet. Combination of context-dependent bidirectional long short-term memory classifiers for robust offline handwriting recognition. *Pattern Recognition Letters*, 90:58–64, 2017.

[119] Konstantinos Ntirogiannis, Basilis Gatos, and Ioannis Pratikakis. A combined approach for the binarization of handwritten document images. *Pattern recognition letters*, 35:3–15, 2014.

[120] Kai Chen, Hao Wei, Jean Hennebert, Rolf Ingold, and Marcus Liwicki. Page segmentation for historical handwritten document images using color and texture features. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 488–493. IEEE, 2014.

[121] Konstantinos Zagoris, Ioannis Pratikakis, and Basilis Gatos. Unsupervised word spotting in historical handwritten document images using document-oriented local features. *IEEE Transactions on Image Processing*, 26(8):4032–4041, 2017.

[122] Sheng He and Lambert Schomaker. Beyond ocr: Multi-faceted understanding of handwritten document characteristics. *Pattern Recognition*, 63:321–333, 2017.

[123] Napa Sae-Bae and Nasir Memon. Online signature verification on mobile devices. *IEEE Transactions on Information Forensics and Security*, 9(6):933–947, 2014.

[124] Moises Diaz, Andreas Fischer, Miguel A Ferrer, and Réjean Plamondon. Dynamic signature verification system based on one real signature. *IEEE transactions on cybernetics*, 48(1):228–239, 2016.

[125] Chawki Djeddi, Imran Siddiqi, Labiba Souici-Meslati, and Abdellatif Ennaji. Text-independent writer recognition using multi-script handwritten texts. *Pattern Recognition Letters*, 34(10):1196–1202, 2013.

[126] Imran Siddiqi, Chawki Djeddi, Ahsen Raza, and Labiba Souici-Meslati. Automatic analysis of handwriting for gender classification. *Pattern Analysis and Applications*, 18(4):887–899, 2015.

[127] Ali Mirza, Momina Moetesum, Imran Siddiqi, and Chawki Djeddi. Gender classification from offline handwriting images using textural features. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 395–398. IEEE, 2016.

[128] Momina Moetesum, Imran Siddiqi, Chawki Djeddi, Yaacoub Hannad, and Somaya Al-Maadeed. Data driven feature extraction for gender classification using multi-script handwritten texts. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 564–569. IEEE, 2018.

[129] Sargur Srihari Graham Leedham. A survey of computer methods in forensic handwritten document examination. In *Proceeding the Eleventh International Graphonomics Society Conference, Sccottsdale Arazona*, pages 278–281, 2003.

[130] Arpita Chakraborty and Michael Blumenstein. Preserving text content from historical handwritten documents. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 329–334. IEEE, 2016.

[131] Yi Li and Wenzhao Li. A survey of sketch-based image retrieval. *Machine Vision and Applications*, 29(7):1083–1100, 2018.

[132] Sargur N Srihari and Sangjik Lee. Automatic handwriting recognition and writer matching on anthrax-related handwritten mail. In *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*, pages 280–284. IEEE, 2002.

[133] Joan Puigcerver, Alejandro H Toselli, and Enrique Vidal. Icdar2015 competition on keyword spotting for handwritten documents. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1176–1180. IEEE, 2015.

[134] Abdelâali Hassaïne, Somaya Al Maadeed, Jihad Aljaam, and Ali Jaoua. Icdar 2013 competition on gender prediction from handwriting. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1417–1421. IEEE, 2013.

[135] Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. Icdar 2013 document image binarization contest (dibco 2013). In *2013 12th International Conference on Document Analysis and Recognition*, pages 1471–1476. IEEE, 2013.

[136] Florence Cloppet, Veronique Eglin, Marlene Helias-Baron, Cuong Kieu, Nicole Vincent, and Dominique Stutzmann. Icdar2017 competition on the classification of medieval handwritings in latin script. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1371–1376. IEEE, 2017.

[137] Luiz G. Hafemann, Robert Sabourin, and Luiz S. Oliveira. Learning features for offline handwritten signature verification using deep convolutional neural networks. *Pattern Recognition*, 70:163 – 176, 2017.

[138] Jenny Ziviani. *Chapter 11. The Development of Graphomotor Skills*, pages 217–236. 12 2006.

[139] Virgil Mathiowetz. Assessing abilities and capacities: Motor behavior. In *Occupational therapy for physical dysfunction*, pages 186–211. Lippincott Williams & Wilkins, 2008.

[140] JHP Van der Meulen, JJ Denier van der Gon, CCAM Gielen, RHJM Gooskens, and J Willemse. Visuomotor performance of normal and clumsy children, fast goal-directed arm-movements with and without visual feedback. *Developmental Medicine & Child Neurology*, 33(1):40–54, 1991.

[141] Stephen Grossberg and Rainer W Paine. A neural model of cortico-cerebellar interactions during attentive imitation and predictive learning of sequential handwriting movements. *Neural Networks*, 13(8-9):999–1046, 2000.

[142] BCM Smits-Engelsman, PH Wilson, Y Westenberg, and Jaak Duysens. Fine motor deficiencies in children with developmental coordination disorder and learning disabilities: An underlying open-loop control deficit. *Human movement science*, 22(4-5):495–513, 2003.

[143] John P Wann. Trends in the refinement and optimization of fine-motor trajectories: Observations from an analysis of the handwriting of primary school children. *Journal of motor Behavior*, 19(1):13–37, 1987.

[144] Niamh Tunney, Leslie F Taylor, Mandy Gaddy, Amie Rosenfeld, Neal Pearce, Jeff Tamanini, and Alison Treby. Aging and motor learning of a functional motor task. *Physical & Occupational Therapy in Geriatrics*, 21(3):1–16, 2004.

[145] Sarah J Harris and David J Livesey. Improving handwriting through kinaesthetic sensitivity practice. *Australian Occupational Therapy Journal*, 39(1):23–27, 1992.

[146] Pimjai Sudsawad, Catherine A Trombly, Ann Henderson, and Linda Tickle-Degnen. Testing the effect of kinesthetic training on handwriting performance in first-grade students. *American Journal of Occupational Therapy*, 56(1):26–33, 2002.

[147] Mathew Thomas, Abhishek Lenka, and Pramod Kumar Pal. Handwriting analysis in parkinson's disease: Current status and future directions. *Movement disorders clinical practice*, 4(6):806–818, 2017.

[148] Janet Summers. Joint laxity in the index finger and thumb and its relationship to pencil grasps used by children. *Australian Occupational Therapy Journal*, 48(3):132–141, 2001.

[149] Janet E Yakimishyn and Joyce Magill-Evans. Comparisons among tools, surface orientation, and pencil grasp for children 23 months of age. *American Journal of Occupational Therapy*, 56(5):564–572, 2002.

[150] Julie L Dennis and Yvonne Swinth. Pencil grasp and children's handwriting legibility during different-length writing tasks. *American Journal of Occupational Therapy*, 55(2):175–183, 2001.

[151] José Eduardo Martinelli, Juliana Francisca Cecato, Marcos Oliveira Martinelli, Brian Alvarez Ribeiro de Melo, and Ivan Aprahamian. Performance of the pentagon drawing test for the screening of older adults with alzheimer's dementia. *Dementia & neuropsychologia*, 12(1):54–60, 2018.

[152] John N Buck. The h-t-p technique. a qualitative and quantitative scoring manual. *Journal of clinical psychology*, 4(4):317–396, 1948.

[153] Karen Laurie Roston, Jim Hinojosa, and Howard Kaplan. Using the minnesota handwriting assessment and handwriting checklist in screening first and second graders' handwriting legibility. *Journal of Occupational Therapy, Schools, & Early Intervention*, 1(2):100–115, 2008.

[154] O Tucha, L Mecklinger, J Thome, A Reiter, GL Alders, H Sartor, M Naumann, and KW Lange. Kinematic analysis of dopaminergic effects on skilled handwriting movements in parkinson's disease. *Journal of neural transmission*, 113(5):609–623, 2006.

[155] Maria Teresa Angelillo, Donato Impedovo, Giuseppe Pirlo, Lucia Sarcinella, and Gennaro Vessio. Handwriting dynamics as an indicator of cognitive reserve: An exploratory study. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 835–840. IEEE, 2019.

[156] Atilla Ünlü, Rüdiger Brause, and Karsten Krakow. Handwriting analysis for diagnosis and prognosis of parkinson's disease. In *International Symposium on Biological and Medical Data Analysis*, pages 441–450. Springer, 2006.

[157] Sanne Broeder, Evelien Nackaerts, Alice Nieuwboer, Bouwien CM Smits-Engelsman, Stephan P Swinnen, and Elke Heremans. The effects of dual tasking on handwriting in patients with parkinson's disease. *Neuroscience*, 263:193–202, 2014.

[158] W Ding, LJ Ding, FF Li, Y Han, and L Mu. Neurodegeneration and cognition in parkinson's disease: a review. *Eur Rev Med Pharmacol Sci*, 19(12):2275–81, 2015.

[159] Claudio De Stefano, Francesco Fontanella, Donato Impedovo, Giuseppe Pirlo, and Alessandra Scotto di Freca. A brief overview on handwriting analysis for neurodegenerative disease diagnosis. In *WAIAH@ AI* IA*, pages 9–16, 2017.

[160] C Kotsavasiloglou, N Kostikis, Dimitrios Hristu-Varsakelis, and M Arnaoutoglou. Machine learning-based classification of simple drawing movements in parkinson's disease. *Biomedical Signal Processing and Control*, 31:174–180, 2017.

[161] Stephan Müller, Oliver Preische, Petra Heymann, Ulrich Elbing, and Christoph Laske. Diagnostic value of a tablet-based drawing task for discrimination of patients in the early course of alzheimer's disease from healthy individuals. *Journal of Alzheimer's Disease*, 55(4):1463–1469, 2017.

[162] Stephan Müller, Oliver Preische, Petra Heymann, Ulrich Elbing, and Christoph Laske. Increased diagnostic accuracy of digital vs. conventional clock drawing test for discrimination of patients in the early course of alzheimer's disease from cognitively healthy individuals. *Frontiers in aging neuroscience*, 9:101, 2017.

[163] Joseph R Cockrell and Marshal F Folstein. Mini-mental state examination. *Principles and practice of geriatric psychiatry*, pages 140–141, 2002.

[164] Ingrid Arevalo-Rodriguez, Nadja Smailagic, Marta Roqué i Figuls, Agustin Ciapponi, Erick Sanchez-Perez, Antri Giannakou, Olga L Pedraza, Xavier Bonfill Cosp, and Sarah Cullum. Mini-mental state examination (mmse) for the detection of alzheimer's disease and other dementias in people with mild cognitive impairment (mci). *Cochrane Database of Systematic Reviews*, (3), 2015.

[165] Masoumeh Farokhi and Masoud Hashemi. The analysis of children's drawings: social, emotional, physical, and psychological aspects. *Procedia-Social and Behavioral Sciences*, 30:2219–2224, 2011.

[166] Helen D Pratt and Donald E Greydanus. Intellectual disability (mental retardation) in children and adolescents. *Primary Care: Clinics in Office Practice*, 34(2):375–386, 2007.

[167] Julian F Miller. Cartesian genetic programming. In *Cartesian Genetic Programming*, pages 17–34. Springer, 2011.

[168] Serge Nicolas, Bernard Andrieu, Jean-Claude Croizet, Rasyid B Sanitioso, and Jeremy Trevelyan Burman. Sick? or slow? on the origins of intelligence as a psychological object. *Intelligence*, 41(5):699–711, 2013.

[169] Jürgen Kornmeier and Michael Bach. The necker cube—an ambiguous figure disambiguated in early visual processing. *Vision research*, 45(8):955–960, 2005.

[170] Vera Miler Jerkovic, Vladimir Kojic, Natasa Dragasevic Miskovic, Tijana Djukic, Vladimir S Kostic, and Mirjana B Popovic. Analysis of on-surface and in-air movement in handwriting of subjects with parkinson's disease and atypical parkinsonism. *Biomedical Engineering/Biomedizinische Technik*, 64(2):187–194, 2019.

[171] Sara Rosenblum, Batya Engel-Yeger, and Yael Fogel. Age-related changes in executive control and their relationships with activity performance in handwriting. *Human movement science*, 32(2):363–376, 2013.

[172] Moussa Djioua and Réjean Plamondon. Studying the variability of handwriting patterns using the kinematic theory. *Human movement science*, 28(5):588–601, 2009.

[173] Christian O'Reilly and Réjean Plamondon. Development of a sigma–lognormal representation for on-line signatures. *Pattern Recognition*, 42(12):3324–3337, 2009.

[174] Giuseppe Pirlo, Moises Diaz, Miguel Angel Ferrer, Donato Impedovo, Fabrizio Occhionero, and Urbano Zurlo. Early diagnosis of neurodegenerative diseases by handwritten signature analysis. In *International Conference on Image Analysis and Processing*, pages 290–297. Springer, 2015.

[175] Heloise Hse and A Richard Newton. Sketched symbol recognition using zernike moments. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 1, pages 367–370. IEEE, 2004.

[176] Réjean Plamondon, Christian O'Reilly, Céline Rémi, and Thérésa Duval. The lognormal handwriter: learning, performing, and declining. *Frontiers in psychology*, 4:945, 2013.

[177] Thérésa Duval, Céline Rémi, Réjean Plamondon, Jean Vaillant, and Christian O'Reilly. Combining sigma-lognormal modeling and classical features for analyzing graphomotor performances in kindergarten children. *Human movement science*, 43:183–200, 2015.

[178] PA Bromiley, NA Thacker, and E Bouhova-Thacker. Shannon entropy, renyi entropy, and information. *Statistics and Inf. Series (2004-004)*, 2004.

[179] James F Kaiser. On a simple algorithm to calculate the'energy'of a signal. In *International conference on acoustics, speech, and signal processing*, pages 381–384. IEEE, 1990.

[180] Ujjwal Bhattacharya, Réjean Plamondon, Souvik Dutta Chowdhury, Pankaj Goyal, and Swapan K Parui. A sigma-lognormal model-based approach to generating large synthetic online handwriting sample databases. *International Journal on Document Analysis and Recognition (IJDAR)*, 20(3):155–171, 2017.

[181] Réjean Plamondon, Xiaolin Li, and Moussa Djioua. Extraction of delta-lognormal parameters from handwriting strokes. *Frontiers of Computer Science in China*, 1(1):106–113, 2007.

[182] Philip Sedgwick. Pearson's correlation coefficient. *Bmj*, 345:e4483, 2012.

[183] Leann Myers and Maria J Sirois. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences*, 12, 2004.

[184] Harold Hotelling et al. A generalized t test and measure of multivariate dispersion. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*. The Regents of the University of California, 1951.

[185] Patrick E McKnight and Julius Najab. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1, 2010.

[186] Lipo Wang. *Support vector machines: theory and applications*, volume 177. Springer Science & Business Media, 2005.

[187] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[188] Geoffrey M Jacquez. A k nearest neighbour test for space–time interaction. *Statistics in medicine*, 15(18):1935–1949, 1996.

[189] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[190] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[191] Joao P Papa, Alexandre X Falcao, and Celso TN Suzuki. Supervised pattern classification based on optimum-path forest. *International Journal of Imaging Systems and Technology*, 19(2):120–131, 2009.

[192] Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18:1–8, 1998.

[193] Ruihu Wang. Adaboost for feature selection, classification and its relation with svm, a review. *Physics Procedia*, 25:800–807, 2012.

[194] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[195] Clifton D Sutton. Classification and regression trees, bagging, and boosting. *Handbook of statistics*, 24:303–329, 2005.

[196] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[197] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[198] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.

[199] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[200] Sheng He and Lambert Schomaker. Deep adaptive learning for writer identification based on single handwritten word images. *Pattern Recognition*, 88:64–74, 2019.

[201] Sebastiano Battiato, Giovanni Maria Farinella, Oliver Giudice, and Giovanni Puglisi. Aligning shapes for symbol classification and retrieval. *Multimedia Tools and Applications*, 75(10):5513–5531, 2016.

[202] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1105–1113, 2016.

[203] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[204] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

[205] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[206] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[207] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[208] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[209] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[210] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[211] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[212] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[213] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[214] Esther J Smits, Antti J Tolonen, Luc Cluitmans, Mark van Gils, Bernard A Conway, Rutger C Zietsma, Klaus L Leenders, and Natasha M Maurits. Standardized handwriting to assess bradykinesia, micrographia and tremor in parkinson's disease. *PloS one*, 9(5):e97614, 2014.

[215] Sigurlaug Sveinbjornsdottir. The clinical symptoms of parkinson's disease. *Journal of neurochemistry*, 139:318–324, 2016.

[216] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[217] Christopher G Goetz, Werner Poewe, Olivier Rascol, Cristina Sampaio, Glenn T Stebbins, Carl Counsell, Nir Giladi, Robert G Holloway, Charity G Moore, Gregor K Wenning, et al. Movement disorder society task force report on the hoehn and yahr staging scale: status and recommendations the movement disorder society task force on rating scales for parkinson's disease. *Movement disorders*, 19(9):1020–1028, 2004.

[218] Javier Galbally, Réjean Plamondon, Julian Fierrez, and Javier Ortega-Garcia. Synthetic on-line signature generation. part i: Methodology and algorithms. *Pattern Recognition*, 45(7):2610–2621, 2012.

[219] Javier Galbally, Julian Fierrez, Javier Ortega-Garcia, and Réjean Plamondon. Synthetic on-line signature generation. part ii: Experimental validation. *Pattern Recognition*, 45(7):2622–2632, 2012.

[220] Miguel A Ferrer, Moises Diaz-Cabrera, Aythami Morales, Javier Galbally, and Marta Gomez-Barrero. Realistic synthetic off-line signature generation based on synthetic on-line data. In *2013 47th International Carnahan Conference on Security Technology (ICCST)*, pages 1–6. IEEE, 2013.

[221] Miguel A Ferrer, Moises Diaz-Cabrera, and Aythami Morales. Synthetic off-line signature image generation. In *2013 international conference on biometrics (ICB)*, pages 1–7. IEEE, 2013.

[222] Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

[223] Ronald L Iman and James M Davenport. Approximations of the critical region of the fbietkan statistic. *Communications in Statistics-Theory and Methods*, 9(6):571–595, 1980.

[224] P. Nemenyi. *Distribution-free Multiple Comparisons*. 1963.

[225] Luigi Trojano and Guido Gainotti. Drawing disorders in alzheimer's disease and other forms of dementia. *Journal of Alzheimer's Disease*, 53(1):31–52, 2016.

[226] Guido Gainotti and Luigi Trojano. Constructional apraxia. In *Handbook of clinical neurology*, volume 151, pages 331–348. Elsevier, 2018.

[227] Thomas Richard Miles. The bangor dyslexia test. 1997.

[228] Gary S Wilkinson and Gary J Robertson. Wide range achievement test (wrat4). *Lutz, FL: Psychological Assessment Resources*, 2006.

[229] Momina Moetesum, Imran Siddiqi, Uzma Masroor, Nicole Vincent, and Florence Cloppet. Segmentation and classification of offline hand drawn images for the bgt neuropsychological screening test. In *Eighth International Conference on Digital Image Processing (ICDIP 2016)*, volume 10033, page 100334N. International Society for Optics and Photonics, 2016.

[230] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[231] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[232] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[233] Momina Moetesum, Osama Zeeshan, and Imran Siddiqi. Multi-object sketch segmentation using convolutional object detectors. In *Tenth International Conference on Graphics and Image Processing (ICGIP 2018)*, volume 11069, page 1106929. International Society for Optics and Photonics, 2019.

[234] Serge Belongie, Greg Mori, and Jitendra Malik. Matching with shape contexts. In *Statistics and Analysis of Shapes*, pages 81–105. Springer, 2006.

[235] Haris Bin Nazar, Momina Moetesum, Shoaib Ehsan, Imran Siddiqi, Khurram Khurshid, Nicole Vincent, and Klaus D McDonald-Maier. Classification of graphomotor impressions using convolutional neural networks: An application to automated neuro-psychological screening tests. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 432–437. IEEE, 2017.

[236] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[237] Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940, 2016.

[238] Momina Moetesum, Imran Siddiqi, and Nicole Vincent. Deformation classification of drawings for assessment of visual-motor perceptual maturity. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 941–946. IEEE, 2019.

[239] Jerome T Connor, R Douglas Martin, and Les E Atlas. Recurrent neural networks and robust time series prediction. *IEEE transactions on neural networks*, 5(2):240–254, 1994.