Video-based Neurological Deficit Analysis

A

Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Yan Zhuang

May 2022

APPROVAL SHEET

This

Dissertation

is submitted in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Author: Yan Zhuang

This Dissertation has been read and approved by the examing committee:

Advisor: Gustavo Rohde

Advisor:

Committee Member: Aidong Zhang

Committee Member: Andrew Southerland

Committee Member: Nikolaos Sidiropoulos

Committee Member: Miaomiao Zhang

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

May 2022

Abstract

The configuration and movement of the face can indicate the presence or absence of various neurological diseases. A central brain lesion such as a stroke will cause pathological, asymmetric weakness on the lower facial muscles on the contralateral side. However, recognizing facial weakness in existing pre-hospital settings still remains a challenge, largely due to variability in training and experience of non-neurologist providers. The proposed research develops an automated, accurate, and quantitative video-based digital screening tool for facial weakness analysis that can enable fast patient triage and augment standard pre-hospital stroke care. The proposed approach not only achieves equivalent performance to paramedics, but also provides visualizable and interpretable results. In addition, to increase the model's robustness to illumination changes, we leverage patch-wise (local) image gradient distributions and transport-based metric for illumination-invariant face analysis. The experiment results demonstrate that the proposed method outperforms other alternatives in several face analysis tasks with challenging illumination conditions.

Acknowledgements

First and foremost, I am deeply grateful to my esteemed advisor Prof. Gustavo K. Rohde for his invaluable supervision, precise guidance, and constant support during my PhD study. Prof. Rohde sets a good example of an outstanding researcher and provides insightful advice and feedback that shape my thinking during the training. I am also thankful for Prof. Andrew Southerland for the great collaboration opportunity on the BANDIT project as well as his crucial support. I would like to express my gratitude to my dissertation committee members: Dr. Aidong Zhang, Dr. Nikolaos Sidiropoulos, and Dr. Miaomiao Zhang for their valuable advise and encouragement.

I own my thanks to several talented post-doctoral researchers: Dr. Shiying Li and Dr. Liam Cattell for their insightful and constructive input and discussion, and my lab mates: Mohammad Shifat-E-Rabbi, Abu Hasnat Mohammad Rubaiyat, and Xuwang Yin for their precious help. I would like to thank various colleagues with whom I had the opportunity to work: Dr. Chad Aldridge, Dr. Mohamed Hassan, Dr. Mark McDonald, Dr. Omar Uribe, and Dr. Natasha Ironside.

I want to thank the present and former director of the computer engineering program: Prof. Barry Johnson and Prof. Joanne Dugan for providing generous support and accepting me to the program. I also appreciate the continuous help from our program administrative assistant Ms. Natalie Edwards.

Finally, I would like to thank my family and friends. First and foremost, I would like to thank my wife Hui Li for her unconditional love and unyielding support. I want to thank my parents for their unlimited support and having faith in me. Without them, I cannot achieve the goal. I would also like to thank Prof. Wenyao Xu and Prof. Feng Lin for their guidance, encouragement, and help.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Medical imaging & computer vision

Machine learning-based technologies including medical imaging have transformed the healthcare sector. Many medical imaging products obtained administration approvals from the US FDA and European CE clearance, and become available in the market [1]. Regions such as India and Thailand start to experimentally deploy AI-guided diagnosis systems for diabetic retinopathy [2]. Other medical imaging analysis techniques focusing on echocardiogram, CT, X-ray, and MRI scans also led to new methods for diagnosing atrial fibrillation, liver dysfunction, and diabetic retinopathy as well skin, breast, and colon cancer [3, 4, 5, 6, 7, 8, 9, 10, 11].

On the other hand, owing to the fact that camera-based devices become more accessible than ever in the form of laptops, smartphones, and tablets as well as the recent radical development of deep learning techniques for computer vision [12], as an emerging research direction, medical computer vision [13] constitutes a critical ingredient for healthcare domain [14], which unlocks a large amount of opportunities for the rise of virtual and contactless care, especially during the COVID-19 pandemic [15]. To be precise, this arising research field, spanning visual detection, tracking, recognition, and analysis of interested clinical events, is able to provide computer-assisted understanding of clinical-relevant diagnostic information regarding complex human behaviors in real-world settings. Successful applications are ranging from medical scene perception in hospitals to physical/physiological activities monitoring [16, 17, 18, 19, 20], for examples, ICU patient mobility monitoring, hand hygiene protocol compliance monitoring, daily living activities classification in elderly living spaces, vital signs measurement, sleep monitoring, fall detection, and gait

estimation. It has the potential to transform the healthcare landscape rapidly, enabling pervasive, non-invasive, low-cost, and long-term monitoring and understanding of the target subject's health conditions [14]. However, clinical deployment of machine vision-based applications still encounter several prominent challenges such as environmental variations, patient-to-patient pathological manifestation and appearance variations, lack of model generalizability and transparency, and big data v.s. rare event [14, 21, 22, 13]. If these problems could be solved successfully, one would create a more accessible and affordable virtual healthcare system that can overcome the physical and geographical barrier and begin to make a border social impact. A more detailed survey regarding medical computer vision and medical imaging analysis are available in [13, 22, 23, 24, 25].

Specifically, in the realm of neurology, machine learning algorithms have been developed to screen for autism, identify language deficits in patients with dementia, detect the electrographic onset of seizures, and monitor speech and motor manifestations of Parkinsons disease [26, 27, 28, 29, 30]. When it comes to neurological function, machine learning has been used to assess facial motion, direction of gaze, and arm strength [31, 32] individually. The Brain Attack and Deficit Identification Tool (BANDIT) project aims to develop an "all-in-one" tool that can detect multiple common neurological deficits such as facial weakness, limb drift, and abnormal eye movement at the same time, by modeling configuration information of the face, eyes, limbs with the event of interest (e.g., stroke) using medical computer vision and medical imaging. Significance of the novel approach is to provide clinically relevant information for cases when expert neurologists are not immediately available for patient triage. This proof-of-concept study provides an illuminating example, showing that automated video-based neurological deficits analysis is able to deliver standard neurology care at a distance via video visits, which not only overcomes physical barriers to provide patients to access convenient medical care [33], but also addresses shortage of neurologists in remote and rural areas [34]. In addition, it could be naturally integrated to existing tele-neurology networks such as NIH StrokeNet [35].

1.2 Stroke

Stroke remains the second leading cause of death worldwide, accounting for 6.6 million deaths in 2019 and 11 percent of all deaths in 2020 [36], with more than 795,000 people suffering a stroke every year [37]. However, mortality alone does not fully capture the societal, financial, and individual burden of stroke given that stroke is also one of the leading causes of long term disability worldwide, as measured by disability-adjusted life years [38]. In addition to the immediate mortality and

morbidity of stroke, the impact of stroke is also heavily intertwined with the global burden of dementia, which shares many of the same chronic vascular risk factors and is a research priority for the National Institutes of Health and World Health Organization, among others [39]. To prevent severe consequences of stroke, rule of thumb for stroke treatment is "time is the essence". Hence, early recognition of stroke improves patient outcomes. Numerous studies have shown that reduction in time to treatment by as little as fifteen minutes significantly decreases mortality and disability in stroke patients [40, 41, 42].

Therefore, Emergency Medical Services (EMS) in the prehospital setting plays a significant role in earlier recognition of stroke and more frequent treatment with acute stroke therapy [43, 44, 45, 46]. However, the identification of stroke by non-neurologists can be challenging, as illustrated by the finding that paramedics fail to detect stroke as high as 56% of patients when not using a diagnostic aid [47]. Pre-hospital screening tools including the Cincinnati Prehospital Stroke Scale (CPSS) and NIH Stroke Scale (NIHSS) administered by emergency medical providers facilitate early detection and triage of acute stroke, but their accuracy and pervasiveness are variable. As a result, the patient or bystander failed to recognize the stroke symptom and only approximately 50% of stroke patients request EMS service for urgent help [48]. Even with the help from EMS, recognition of stroke remains a challenge for EMS personnel [49, 50], who is without comprehensive neurological training. All contribute to the high miss rate for EMS stroke alerts in the field [51, 52, 53, 47]. Furthermore, shortage of neurological expertise and the designated stroke care team, especially in rural and underserved areas, leads to treatment delay as well [54]. Furthermore, the accuracy of currently available stroke diagnostic aids are even more limited in rural and low access areas. One recent study [55] shows the disparities between rural and urban regions in terms of stroke treatment. The Joint Commission on Accreditation of Health Care Organizations certified primary stroke centers (PSCs) account for 2.4% of rural hospitals, compared to 18.7% of their urban counterparts. As a result, many patients either fail to receive timely treatment or are ineligible for acute stroke therapy as time ran out. Additionally, the COVID-19 pandemic further strains prehospital stroke care by increasing demands on emergency providers, and possibly an increased risk of severe strokes in patients with SARS-CoV-2 infection [56, 57].

To address the aforementioned issues, recent efforts introduce the neurological expertise through ambulance-based telemedicine [58, 59]. One study previously demonstrated that mobile video teleneurologic assessment in the field is feasible and correlative to an in-person examination by a vascular neurologist [60]. However, while mobile teleneurologic assessment is a promising addition to the current paradigm, it is not an easily generalizabile solution. The number of patients who

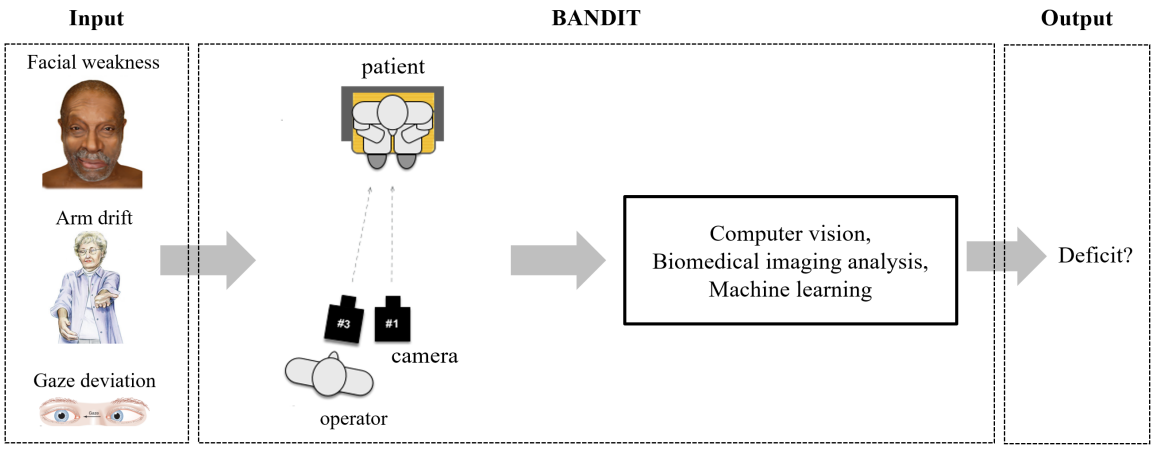

Figure 1.1: The BANDIT project.

need to be screened far exceeds the number of available neurologists and stroke specialists to help with screening, and with an ever increasing shortage of neurologists, both in the US and worldwide, this disparity will likely worsen [61, 62, 63].

In order to realize the full potential of prehospital stroke screening, a great need is desired to find a solution to the problem of neurological deficit detection beyond remote evaluation by a neurologist. Innovation that improves detection of acute stroke in the field will result in earlier and more frequent treatment with acute stroke therapies, better patient mortality and disability outcomes, and decreased social burden of stroke.

1.3 The BANDIT Project

The most commonly used stroke identification tools in the prehospital setting are CPSS and NIHSS, which include assessment of facial weakness, limb weakness or drift, and dysarthria (i.e. impaired speech) [64, 65]. Detection of these deficits renders the screen positive, prompting notification and triage of suspected stroke to a nearby hospital [65]. Unfortunately, signs of stroke can be difficult to recognize without comprehensive neurologic training [49, 50]. Their efficacy is similarly limited by inaccurate and inconsistent stroke detection by non-neurologists [66].

Inspired by the existing CPSS and NIHSS examination protocols performed by the neurologists, the Brain Attack and Deficit Identification Tool (BANDIT) project targets to automate these clinical examination protocols using medical computer vision and imaging analysis. The screening tool is automated, accurate, quantitative, easily-deployable, and low-cost, as shown in Fig. 1.1. The proposed system would allow the non-neurologist users to identify neurological deficits quickly in the

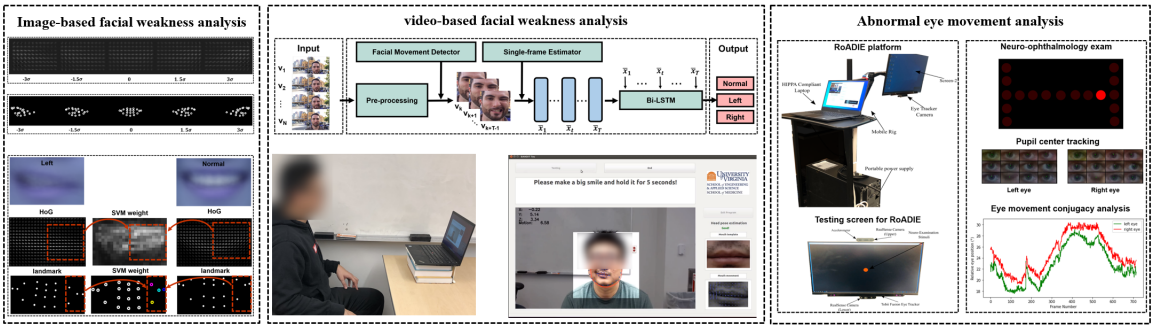

Figure 1.2: Facial weakness detection and eye deficit detection for the BANDIT project.

field and minimize inter-operator variability and operator error of standard neurological examination. The development process is performed within a supervised learning framework. Namely, given an input examination video, the framework tracks the motion of faces, eyes, and limbs, assesses their configurations on-the-fly, and provides the quantitative and interpretable evaluation results. The expected outcomes of BANDIT is to augment standard neurology care in rural and underserved areas, expedite the diagnostics and treatment process, and provide more accurate triage of neurological patients.

The ongoing BANDIT study focuses on the detection and quantification of facial weakness and abnormal eye movement analysis. First, BANDIT systematically investigates several options for facial weakness detection on static images [67, 68] and devises an novel automated facial weakness video identification framework [69, 70]. The validation experiments of the current prototype [70] against human raters show the proposed algorithm's accuracy rate is equivalent to paramedics and close to trained residents. BANDIT also investigates various methods for quantifying abnormal eye movement, by presenting a video-based eye movement assessment workflow for abnormal eye movement detection. The proposed workflow utilizes a ResNet-based pupil detector to detect the center of the pupil for both eyes in a given video and analyzes the conjugacy of eye movement [71]. Experimental evaluations demonstrate that the proposed system is able to achieve equal performance compared to the COTS eye tracker such Tobii eye tracker [72]. In addition, it has been integrated into the Rolling Apparatus to Detect Impairment of the Eyes (RoADIE) platform, which is equipped with various imaging and signal modalities to acquire patient video data at the emergency room [73].

1.4 Dissertation overview

This thesis focuses on studying the topics of vision-based facial weakness detection and quantification via medical computer vision and computational imaging. To detect and quantify this neurological sign, it first systematically investigates different options for facial weakness analysis, including the shape-based features, appearance-based features, combined, or CNN feature maps on static images [67, 68] or video data [69, 70]. It leverages various methods to improve models transparency. Specifically, the pilot study presents an automatic pathological facial weakness detection tool based on a single RGB image [67]. The proposed system is able to extract the facial landmarks and classify facial weakness using a supervised learning method. The learning method projects the shape-based features onto a much lower subspace, where the low-dimensional representations not only greatly reduce the feature dimension, but also produce the visualizable and interpretable results to understand facial weakness as shown in Fig. 1.2 top left panel. In a subsequent study [68], we experimentally evaluate the performance of multiple state-of-the-art landmark feature extraction methods for measuring facial weakness, showing that landmark-based methods can suffer from inaccuracies in face landmarks localization. Moreover, we then compare different facial weakness classification schemes using various handcrafted-based features and deep learning approaches. Evaluation results demonstrate that a combination of shape and appearance-based features produce the best results. By visualizing the maximum weight of the SVM classifier for the HoG features and landmarks features, it shows that the clinically meaningful information can be detected by the proposed model, as shown in Fig. 1.2 bottom left panel. Utilizing the knowledge learned from two previous studies, we propose to assess facial weakness from a simple video examination procedure. To achieve this, we build an automated framework for facial weakness detection using videos from a regular RGB camera [69, 70]. Fig. 1.2 (top middle panel) depicts such a framework. For a given video, it first extracts the most discriminant shape and appearance information regarding facial weakness, by learning a subspace model that transforms the high dimensional shape and appearance-based features into the low dimensional representations. Then a recurrent neural network models the temporal dynamics of both low-dimensional shape and appearance-based features of each frame through a Bi-LSTM network. The system is evaluated on a "in-the-wild" video facial weakness dataset that is verified by three board-certified neurologists. Experimental evaluation shows that it is able to outperform other state-of-the-art alternatives, achieve the equal performance to paramedics, and provide visualizable and interpretable results that increases model transparency. In addition, a live and real-time prototype with interactive GUI is implemented on a regular laptop as a proof-of-concept [74]. The important

implication of this study is that the proposed method opens a new opportunity of providing clinical assistance to non-neurologists (e.g, paramedics) to increase the coverage of standard neurological prehospistal care.

When experimenting various facial weakness detection and quantification algorithms on the patient dataset acquired in real-world clinical settings (e.g., outpatient clinical room and inpatient ward), illumination variability is an important factor that affects the algorithm's performance. Thus a new low-level optimal transport-based illumination-invariant image descriptor is proposed [75], which beyond the scope of facial weakness analysis. The method is based on mathematical modeling of local gradient distributions using the Radon Cumulative Distribution Transform (R-CDT). The proposed work demonstrates that lighting variations cause certain types of deformations of local image gradient distributions which, when expressed in R-CDT domain, can be modeled as a subspace. Face recognition is then performed using a nearest subspace in R-CDT domain of local gradient distributions, as illustrated in Fig. 1.3. Experiment results demonstrate the proposed method outperforms other alternatives in several face recognition tasks with challenging illumination conditions.

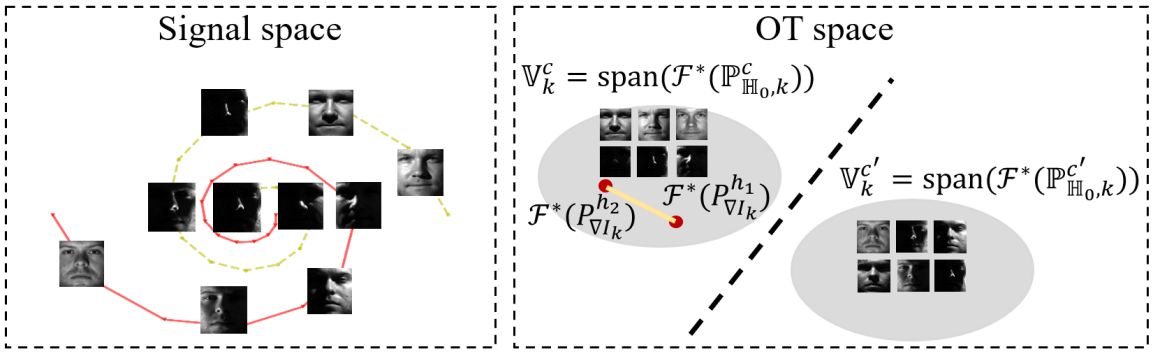

Figure 1.3: Local-Wasserstein feature sets for illumination-invariant face recognition.

1.5 Dissertation structure

Section 1 first provides an overview of the dissertation. The subsequent **Section 2** and **Section 3** detail methods regarding image-based and video-based facial weakness analysis as well as the experimental evaluation results. In **Section 4**, to address varying illumination issues encountered in the real-patient dataset, the thesis devises a novel optimal transport-based representation for illumination-invariant face analysis task. The following **Section 5** provides the discussion regrading

the knowledge learned from this study and future work. The final chapter **Section 6** concludes the dissertation.

Chapter 2

Image-based facial weakness detection and quantification

2.1 Overview

Although the in-person examination or video visit by a neurologist is still the gold standard, systems for identifying and quantifying facial weakness from static frames have emerged over the past decade [76, 77, 78, 79, 67, 80, 81, 82, 83, 84, 85]. In this study, goal of image-based analysis is not to clinically assess facial weakness symptom, but rather to study the best representation of facial weakness. By investigating the optimal feature representation of facial weakness on static images, the knowledge learned is a necessary and crucial step in the process of arriving at a full-fledged video-based technique to clinically diagnosis facial weakness.

While a large amount of techniques have been developed for facial weakness assessment, broadly speaking, these approaches can be categorized into landmark-based features [76, 77, 78, 79, 67, 80, 81], which encode information generally associated with position of key facial features such as shape, and texture-based features [82, 83, 84, 85] that focus on intensity information such as skin color, creases, texture, etc. Top and bottom right panel of Fig. 2.1 shows the shape-based features and texture-based features. The most commonly used methods for facial weakness detection heavily rely on extracting facial landmarks and calculating handcrafted geometric features, such as computing distance and angle features between facial landmarks as well as asymmetry ratios [76, 77, 78, 79, 67, 80, 81].

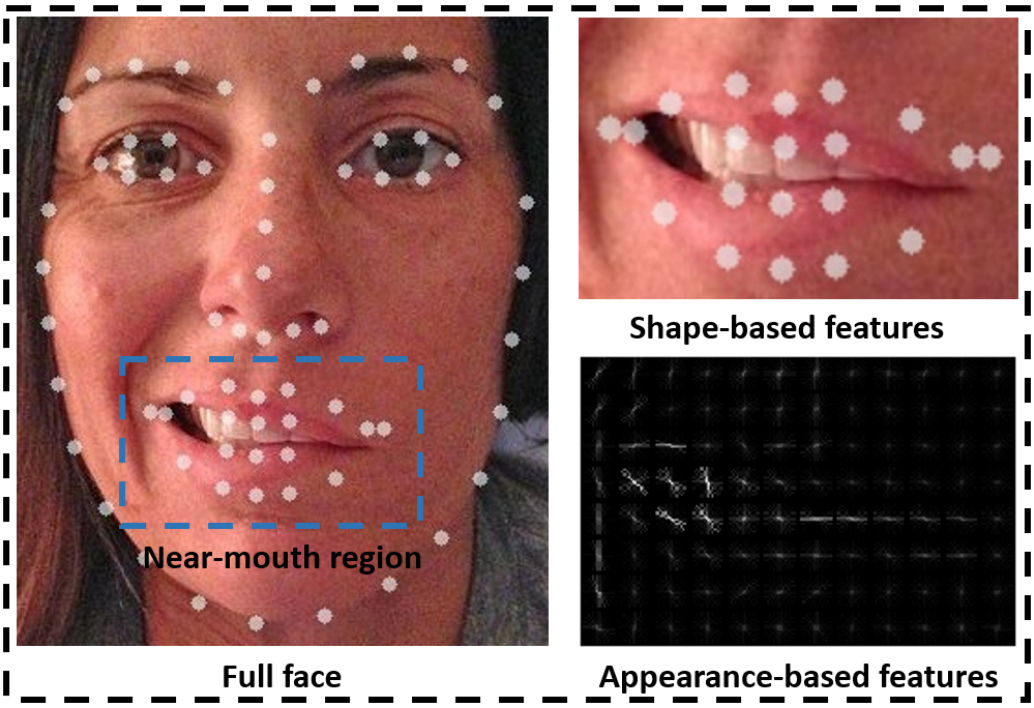

Figure 2.1: Examples of left facial weakness, shape-based, and texture-based features.

The pilot study, rather than calculating handcrafted features directly, presents a supervised learning system that is able to learn the discriminant information regrading facial weakness using on a statistical shape analysis method [67]. Specifically, after the facial landmarks extraction and alignment, the proposed learning approach projects the shape-based features onto a much lower subspace. The compact and low-dimensional representation not only greatly reduces the feature dimension, but also produces the visualizable and interpretable results. However, in this study we also notice that the accuracy of facial landmark extraction approaches is insufficient in some cases, because facial landmarks fail to delineate the shape of the mouth accurately, as demonstrated in the left panel of Fig. 2.1. Thus, a subsequent study [68] first experimentally evaluates the performance of multiple state-of-the-art landmark feature extraction methods in terms of facial weakness classification accuracy, showing that inaccurate face landmarks localization issues can lead to classification accuracy drop. Then, we use a straightforward linear SVM classification framework to empirically demonstrate that a combination of landmarks and texture-based features produces the best results when compared to use either landmarks or textures features separately. At end of the study, to improve the interperablity of the proposed model, we visualize the maximum weight of the linear SVM classifier for the HoG features and landmarks features, in order to show that the clinically meaningful pathological information can be detected by the proposed model.

2.2 Related work

Telestroke A recent strategy has been the introduction of neurological expertise in the prehospital setting through telemedicine [58, 59], in order to enhance emergency response. The iTREAT study previously demonstrated that such mobile video tele-neurological assessment in the field is feasible [58, 59]. Other research has shown that remote evaluation of a patient in the ambulance via telemedicine highly correlates with on-board examination by a vascular neurologist [60]. This setup was shown to be cost-effective in starting the evaluation of stroke prior to arrival in the hospital, saving time and potentially improving patient outcomes, when compared with the traditional stroke evaluation process. However, while mobile tele-neurological assessment is a promising improvement of the current paradigm it is not a comprehensive solution. The number of patients who need to be screened far exceeds the number of available vascular neurologists. With an ever increasing shortage of neurologists, both in the US and worldwide, this disparity will likely worsen [61, 62, 63, 35]. This setup was complex and cost of maintenance is high. Several studies showed that the operation cost of telestroke system in the stroke center is more than $10k USD [86, 87, 88].

Facial weakness detection Facial analysis systems have a long history in the medical applications. For instances, Almutiry et. al. [89] developed a system to recognize Parkinson's disease by examining facial emotional expression. Kruszka et. al. [90] also devised a system that was able to leverage facial configuration landmarks-based geometric features to discovery the genetic disease such as the 22q11.2 deletion syndrome. Readers can find a more comprehensive literature review in [20]. In the realm of neurology, several works have been developed to facial weakness analysis [76, 77, 78, 79, 67, 81]. Depending on the features used in their studies, they can generally be divided into two categories: (1) landmark-based approaches; (2) intensity-based approaches. The most popular method is to use the facial landmarks to measure facial configuration features, such as distance and angle between facial landmarks, and use these configuration features to perform classification. For instance, Gaber *et al.* [77] implemented a Kinect based system to quantify facial paralysis by calculating a symmetry index for the eyebrows, eyes, and mouth. Similarly, Park *et al.* [78] proposed to calculate an asymmetry index by measuring the displacement difference between the right and left side of the forehead and mouth area. In addition to landmark-based features, several work have been proposed to detect facial weakness based intensity-based features. To be specific, Guo *et al.* [85] proposed a Convolution Neural Network (CNN) method to perform facial weakness severity classification using a pre-trained InceptionV1 model. He *et al.* [82, 83] employed

the optical flow and local binary pattern on three orthogonal planes (LBP-TOP) to perform the facial weakness classification for video data.

2.3 Predictive models

The first section illustrates preprocessing steps include face detection, landmark extraction, and face normalization that removes the rigid transformations such as translation, rotation, and scaling effects. The following two sections present two aforementioned studies: [67] and [68] accordingly. The work [67] presents a statistical shape analysis method for facial weakness analysis, while the work [68] demonstrates the inaccurate landmarks localization issues, and provides a solution accordingly, by using a combination of texture-based and shape-based features.

2.3.1 Prepossessing

The preprocessing step's goal is to detect the existence of a face, extract the corresponding facial landmarks, and align facial landmarks as well as the pixel intensities to remove translation, scaling, and rotation effects in the images. To begin, a human face can be represented by 68 pairs of points, in which each pair of points defines the location of a specific facial landmark, as illustrated in the left panel of Fig. 2.1. We employs a robust facial landmark tracking system [91], including a face detector that can detect faces and a facial landmark detector that can locate facial landmarks. Then, the images need to be aligned, due to the fact that the image data was collected from public repositories such as Google Image and YouTube, and the facial images contained in them are subject to random location, orientation, and size variations. In order to remove these variations, begin by averaging the location of each set of landmarks to produce an average template, the preprocessing step applies a rigid body estimation method to estimate the rotation, scaling, and translation parameters that align each image in the data set to the average template. Extracted landmarks, aligned landmarks, raw intensity information, and aligned intensity values are shown in Fig. 2.2 (a), Fig. 2.2 (b), Fig. 2.2 (c), and Fig. 2.2 (d), respectively.

2.3.2 A shape-based method for facial weakness analysis

Approach one utilizes the penalized Linear Discriminant Analysis (pLDA) on the aligned landmarks to model the pathological deformations regarding facial weakness by maximizing the separation for three different classes: normal, left facial weakness, or right facial weakness [67]. Compared with the

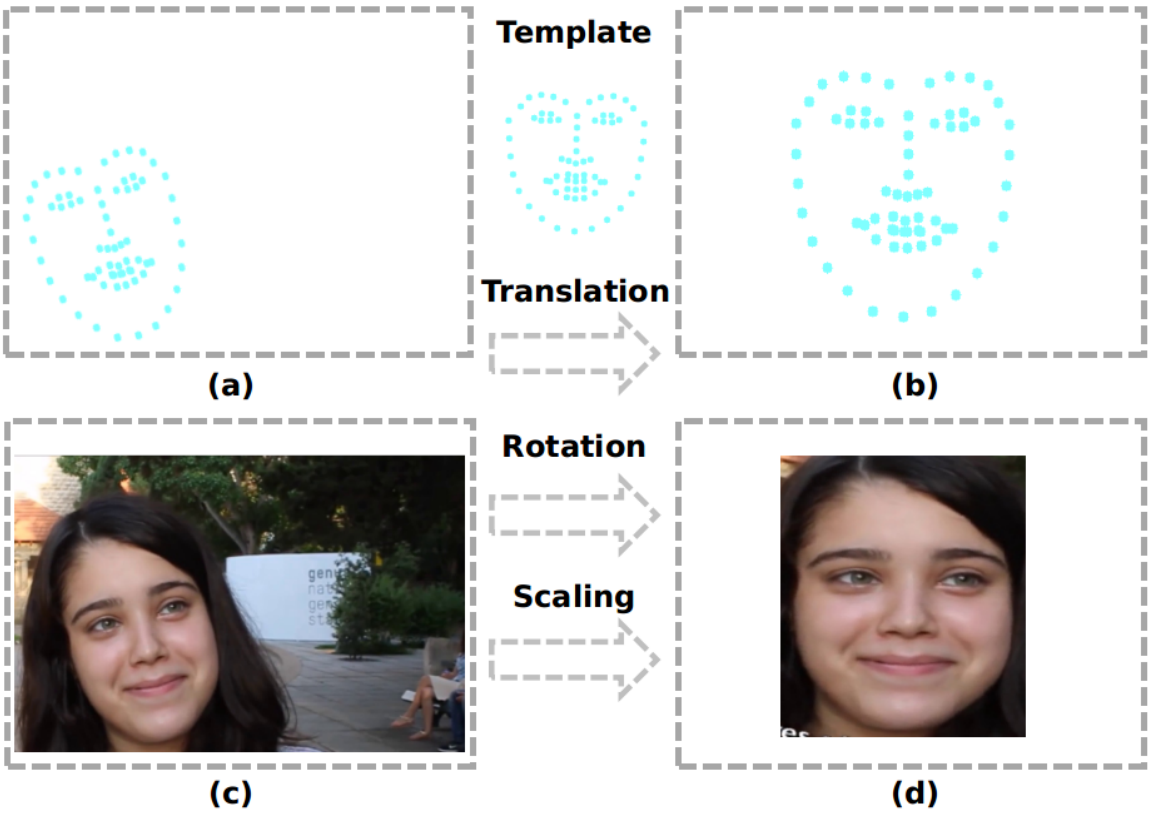

Figure 2.2: The facial landmarks and intensity normalization.

LDA method, which is typically burdened by ill condition problems when applied to high dimensional spaces [92], the pLDA approach is able to address the potential issues of singularity, using an extra penalized term. In addition, the pLDA has the advantage of exacting meaningful discriminating features with respect to class-related information, thus increasing model transparency.

Specifically, given a set of subject landmarks denoted as $\mathbf{l}_n$, $n = 1, \cdots, N$, with N being the number of subjects, the standard LDA method is formulated as:

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \left(\frac{\boldsymbol{w}^T S_B \boldsymbol{w}}{\boldsymbol{w}^T S_W \boldsymbol{w}} \right) \tag{2.1}$$

where $S_B = \sum_c N_c(\mu_c - \tilde{l})(\mu_c - \tilde{l})^T$ and $S_w = \sum_c \sum_{i \epsilon c}(\mathbf{l}_i - \mu_c)(\mathbf{l}_i - \mu_c)^T$ define the 'between class scatter matrix' and 'within classes scatter matrix'. μ_c is the center of class c, $\tilde{l}$ is the center of all dataset, and N_c is number of instance in class c. The goal of LDA is to maximize the objective function above and the solution of this optimization problem is identical to the solution of the eigenvalue problem:

$$S_B \boldsymbol{w} = \lambda S_W \boldsymbol{w} \tag{2.2}$$

Because of the potential issues of singularity or near singularity in S_W, it is customary to replace it by $S_W + \alpha I$, with I the identify matrix, Wang *et. al.* [93] noticed that such modification amounts to solving a "penalized" version of the problem. That is, the equation $S_T \boldsymbol{w} = \lambda (S_W + \alpha \boldsymbol{I})$ is the solution of the following optimization problem:

$$\max_{\boldsymbol{w}} \left(\frac{\boldsymbol{w}^T S_T \boldsymbol{w}}{\boldsymbol{w}^T (S_W + \alpha \boldsymbol{I}) \boldsymbol{w}} \right) \tag{2.3}$$

which equals to adding a penalty term defined as $\alpha - \left(\frac{\boldsymbol{w}^T \boldsymbol{w}}{\boldsymbol{w}^T S_T \boldsymbol{w}} \right)$ to the original problem. Once a set of directions $\boldsymbol{w}_i$, $i = 1, \cdots, M$ has been obtained from the pLDA procedure, we assemble a projection matrix W where each column is composed of an eigenvector from the problem above $\boldsymbol{w}_i$. The first two eigenvectors are kept for classification. We configure the pLDA penalty weigh term α as 0.1 and only keep first two pLDA directions, because first two eigenvectors contain the most meaningful discriminating information, which captures the changes of mouth shape and the changes of mouth sizes, respectively.

Therefore, a low dimensional discriminant representation for a given subject is then obtained as $\hat{l}_n = W^T \mathbf{l}_n$. After pLDA analysis, we utilize a KNN classifier for the classification [94] on the low dimensional vectors $\hat{l}_n$. The configuration setup for the KNN classifier are 5 neighbors with uniform weight. The classification result has three classes: normal, left facial weakness, and right facial weakness.

Evaluation results The proposed method is evaluated on a "in-the-wild" facial weakness dataset. Specifically, to build such an image dataset, we gather images of healthy controls as well as facial weakness patients from publicly available online repositories such as Google Image. Specifically, images of people with normal smiles and with unilateral facial weakness are collected and organized in the following groups: normal, left facial weakness, and right facial weakness. The images are then reviewed independently by two senior resident neurologists. Each image is given a numerical score ranging from "1" to "5", which "1" denotes the likelihood that pathology is absent and "5" denotes high likelihood that pathology is present. Additionally, images for which pathology (e.g., facial weakness in our case) is suspected are further classified as left or right denoting the side of the droop. Only images with the same score (e.g., 1 for normal, 5-left for left facial weakness, 5-right for right facial weakness) are used for the study. The total number of faces showing normal smile and facial weakness is 333. Only images with consistent rating by both raters as likely normal or likely abnormal are included for analysis. Of the 199 images analyzed, 18 images are excluded due

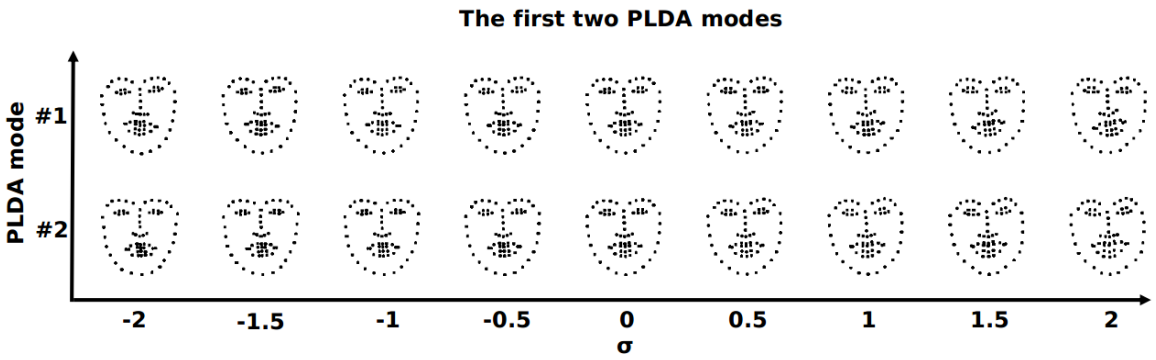

Figure 2.3: Modes of variations along the first two pLDA directions.

to limitations in facial landmark extraction. Of the remaining 181 images, 87 are rated as likely normal and 94 are rated as likely having pathological facial weakness (44 are left facial weakness and 50 are right facial weakness). The stratified 5-fold cross-validation is used to evaluate the predictive model. Specifically, the datasets is randomly divided into 5 groups with balanced samples. Four groups of samples are used for the training process, and the left a single group serves as the testing data. This process repeats for 5 times.

The proposed method performs with the overall accuracy of 94.5%, precision of 94.8%, sensitivity of 94.6%, and specificity of 96.8%. This algorithm achieves high specificity for all three classes, indicating fewer occurrence of false negative cases. The high precision rate indicates that most of the true positive cases can be identified correctly. Fig. 2.3 illustrates the first two pLDA discriminant components for the test dataset. The first row is the first pLDA discriminant component, while the second row is second pLDA discriminant component. From Fig. 2.3, it is clear that the first discriminant component is able to detect the mouth shape differences among three classes, since the control subjects have the symmetric mouth shape while the patient mouth shapes are more likely to be asymmetric. The second pLDA component is about the mouth size differences. The clinical interpretation of second pLDA component is that stroke patient loses the control of facial muscle on the affected side. Therefore, the patient's mouth cannot be fully open when they are smiling, thus the mouth size of patient is relatively smaller than that of health control.

Fig. 2.4 and 2.5 are the histograms of testing data projected onto first 2 pLDA directions, showing that the first pLDA discriminant component is able to classify the left facial weakness, normal, and right facial weakness, while the second pLDA discriminant component is able to classify the normal and abnormal cases (left deficit and right deficit).

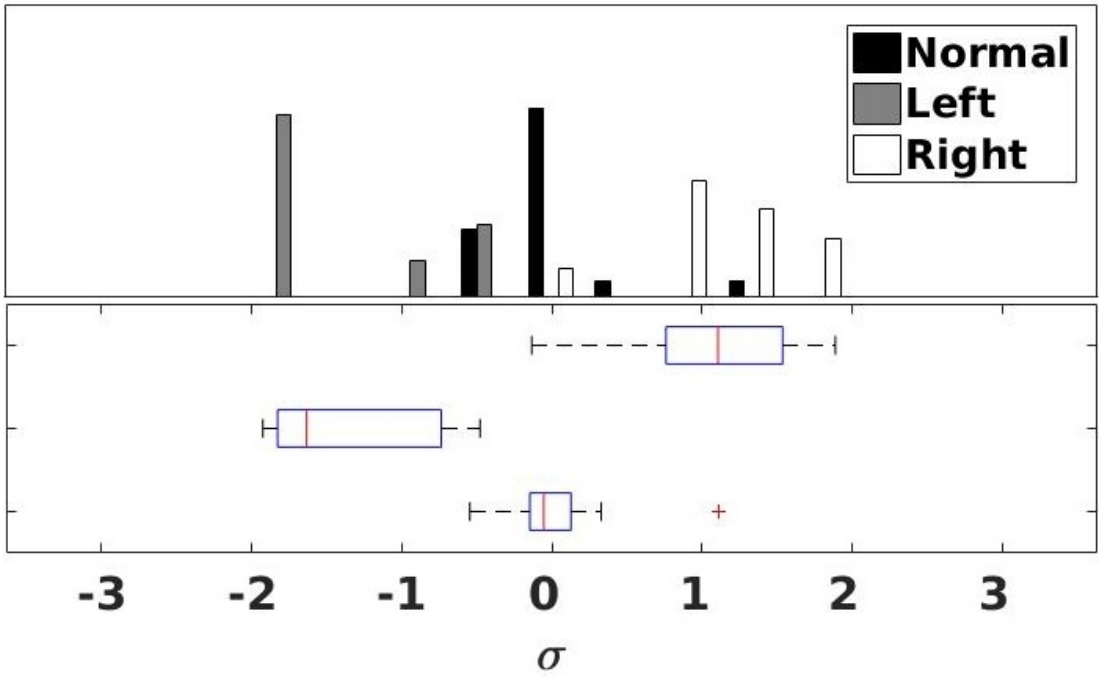

Figure 2.4: Histograms of data projected onto the first principle pLDA direction.

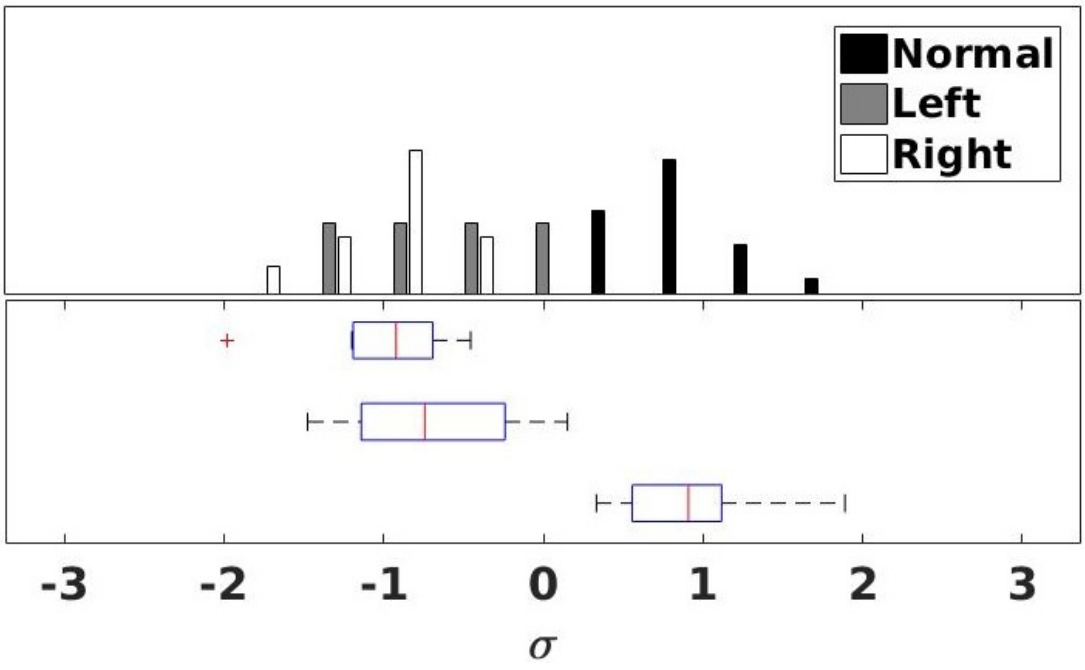

Figure 2.5: Histograms of data projected onto the second principle pLDA direction.

Figure 2.6: Inaccurate landmark localization issues.

2.3.3 A texture-based approach for facial weakness analysis

Previous study [67] shows that landmark extraction algorithms, however, are at times ineffective, thus landmark-based methods can suffer from inaccurate face landmarks localization issues as shown in Fig. 2.6. To address this issue, we empirically demonstrate that approaches that incorporate the texture-based features, such as Histogram of Oriented Gradients (HoG) features [95], tend to be more accurate [68]. Therefore, the second study examines the performance of several state-of-the-art facial landmark extractors for facial weakness detection, then presents the idea that incorporating the texture-based information consistently improves the performance for all different landmark schemes on a larger neurologist-verified facial weakness image benchmark dataset [68].

First, we evaluate the performance of the state-of-the-art landmark extractors in terms of facial weakness classification accuracy. Three face detection and facial landmark extraction schemes are assessed in our study: (1) Deformable Part Model (DPM)[96] and Coarse-to-Fine Shape Searching (CFSS) algorithm [97]; (2) Deformable Part Model (DPM) and Ensemble of Regression Trees (ERT) algorithm [91]; (3) Single Stage Headless (SSH) [98] algorithm and Hourglass Network (HN) [99]. The DPM+CFSS and the DPM+ERT are two learning-based schemes that placed first and second in a recent facial tracking comparison paper [100]. The SSH+HN is a deep learning based scheme that has achieved the best performance in another comparison paper [101]. Evaluation results of different schemes utilizing these landmarks extractors on the facial weakness detection task are summarized in Table 2.1. The main take-way message is that using state-of-the-art landmark schemes, even from

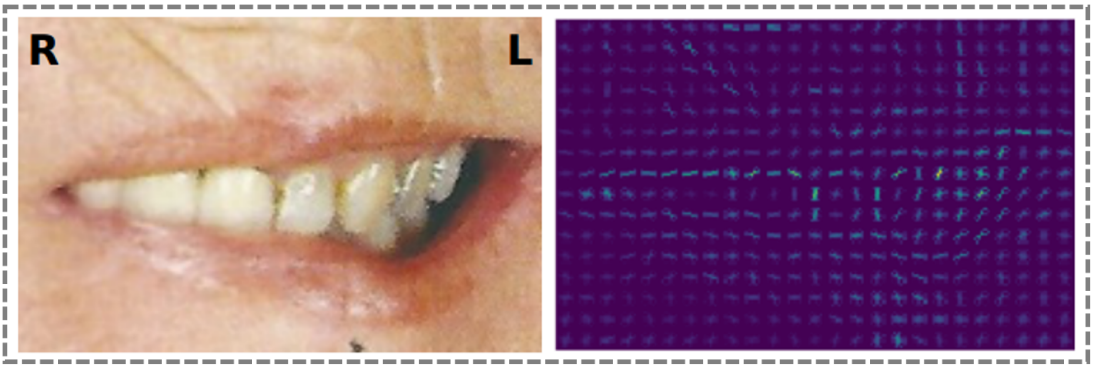

Figure 2.7: A HoG features example for the near-mouth region.

the state-of-the-art facial landmark extractors, are not able to achieve the best performance in terms of classification accuracy. A detailed discussion is available in later section.

We explore the idea that combines landmarks-based features and texture-based features to improve facial weakness classification accuracy. In other words, our contribution [68] illustrates that incorporating texture-based features consistently improves the facial weakness classification when compared to the approaches that solely use landmark-based features or texture-based features. To demonstrate that, we first extract landmarks feature and organize it as a standard vector $\mathbf{l}$. The HoG features, which refer to the statistical distribution of the gradient orientation as show in Fig. 2.7, are extracted. More specifically, an image is divided into several small subareas (called cells). For each cell, the gradient information including magnitude $\nabla\hat{I} = [\nabla_x I, \nabla_y I]$ and phase $\theta = \tan^{-1}\frac{\nabla_y I}{\nabla_x I}$ are calculated, then the magnitude of $[\nabla_x I,\ \nabla_y I]$ is rearranged into separated orientation bins. Several cells are grouped into a block, where L_2 norm based contrast normalization is performed on the histograms from this block to remove local lighting difference. Finally, the HoG features are formulated by concatenating these histograms from each block into a larger vector containing all histograms estimated, denoted as $\mathbf{g}$. The HoG features are extracted using the following parameters configuration: the number of orientation bins in each cell is 9 and a cell consists of 8 by 8 pixels. In addition, since both the landmark features and the HoG features are high-dimensional, we compute the PCA coefficients from the training data set for the facial landmark features and the HoG features, respectively. We reduce the dimension of the features to the number of components that can cover 90% of the variance. After PCA transformation, we concatenate the two vectors into a single vector $\mathbf{x} = [\hat{\mathbf{l}}, \hat{\mathbf{g}}]$ that will serve as input to the predictive model.

In short, given a set of training examples denoted as $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n, \cdots, \mathbf{x}_N]$, $n = 1, \cdots, N$, with $\mathbf{x}_n = [\hat{\mathbf{l}}, \hat{\mathbf{g}}]$ being one particular feature vector (each image corresponds to one feature vector)

and N being the total number of images. $y_t \in \{1, \cdots, c, \cdots, C\}$ is the ground truth label for t_{th} subject, and C is the total number of classes. In our case C equals to 3. We formulate the linear SVM classifier using "one-versus-one" setting, meaning that this method constructs $K = C(C-1)/2$ binary classifier. Each of these binary classifiers is trained using the data from two classes. Therefore, suppose that the training data from the i_{th} and j_{th} classes, this binary classification problem can be formulated to the following optimization problem [102]:

$$\min_{w^{ij}, b^{ij}, \xi^{ij}} \frac{1}{2}(w^{ij})^T w^{ij} + C \sum_t \xi_t^{ij} \tag{2.4}$$

$$\text{subject to} (w^{ij})^T x_t + b^{ij} \geq 1 - \xi_t^{ij} \text{ if } y_t = i, \tag{2.5}$$

$$(w^{ij})^T x_t + b^{ij} \leq -1 + \xi_t^{ij} \text{ if } y_t = j, \tag{2.6}$$

$$\xi_t^{ij} \geq 0. \tag{2.7}$$

where $C \sum_t \xi_t^{ij}$ is the penalty parameter and b^{ij} is the bias term. The training process aims to obtain the optimal w^{ij} and b^{ij} [102]. In the testing phase, each of the K binary classifiers outputs a classification result, the final decision function uses a majority voting strategy to obtain the largest vote as the prediction.

Deep learning comparison method To compare with Guo *et al.* [85], we implement a transfer learning based facial weakness classification framework. Rather than directly using a pretrained ImageNet model, we choose to train a Convolutional Neural Network (CNN) model on a human emotion recognition task [103], since it is a highly related classification task, which consists of 28709 training and 3589 testing 48x48 pixel grayscale images of aligned faces with one of seven labels: angry, disgust, fear, happy, sad, surprise, and neutral. The base model, consisting of four convolutional layers and two dense layers, is able to achieve 65% accuracy for the emotion classification task. Then we fine-tune the model using the facial weakness dataset that will be discussed below. Specifically, we modify the original last dense layer to output three classes instead of seven. Thereafter, we fine-tune the entire network using our data set. An earlier stop strategy is used for the learning process. The optimization goal is to maximize the categorical accuracy using an Adam optimizer with the leaning rate of 0.001.

Evaluation results To comprehensively evaluate the proposed system, we created a facial weakness dataset benchmark that consists of 437 images of people with facial weakness and health con-

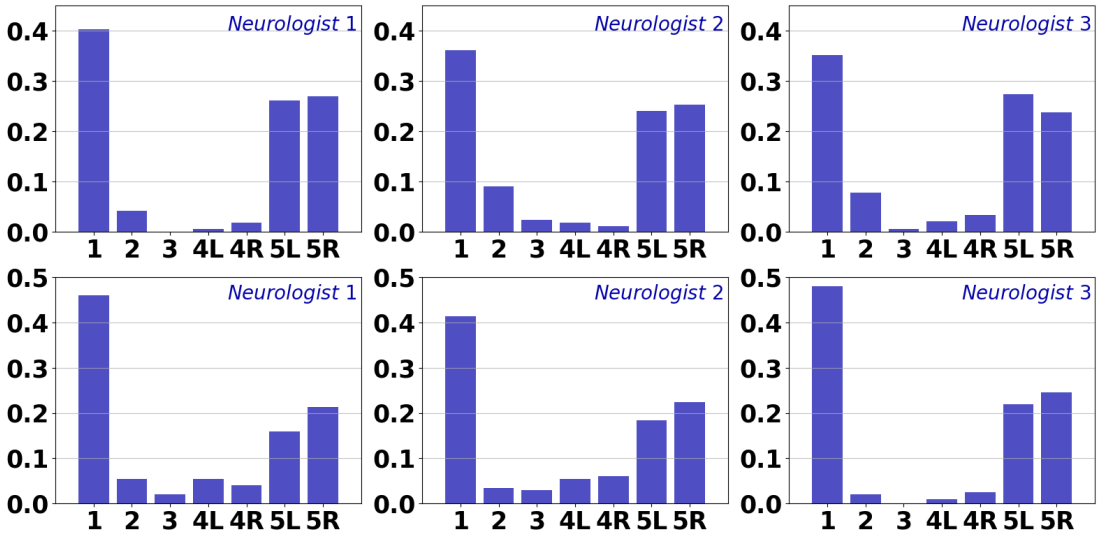

Figure 2.8: Rating distribution for images sub-dataset and video sub-datset.

trols. Firstly, three senior medical students collect 333 images (image sub-dataset) and 203 videos (video sub-dataset) of facial weakness and normal controls from public available repositories such as Google Images and YouTube. Specifically, images and videos of people with a smile are collected to guarantee muscle activation for assessment of pathological asymmetry. These data are then reviewed by six neurology experts for verification. First, three experienced neurology trainees rate the image sub-dataset. Each image is given a numerical score ranging from '1' to '5', which '1' denotes facial weakness absence, '2' denotes high likelihood that facial weakness is absent, '3' denotes undetermined cases, '4' denotes high likelihood that facial weakness is present, '5' denotes that facial weakness is present. Additionally, images with facial weakness are further classified as left (L) or right (R) denoting the laterality (side of the weakness). Only images with the same score (e.g., '1' for normal, '5L' for left facial weakness, '5R' for right facial weakness) by all three raters are used in this study for analysis. 95 images are excluded due to disagreement. Among these 238 images, 88 images are rated as normal (facial weakness absence), 77 images are rated as left facial weakness, and 73 images are rated as right facial weakness. Fig. 2.8 shows the rating distributions. The first row in Fig. 2.8 shows the rating distribution of the three experienced neurology trainees for all 333 images. The corresponding Fleiss Kappa value is 72.0% [CI 68.3%-75.6%]. In addition, in order to augment the above-mentioned image sub-dataset, we sample one frame from each video in the video sub-dataset and add them to the image sub-dataset. The videos are reviewed independently by three board-certified neurologists. The video rating follows the same rating protocol, given a numerical score ranging from '1' to '5', which '1' denotes facial weakness absence, '5R' denotes right

	Proposed method									CNN	
	Near eye			Near mouth			Full face			Near mouth	Full face
	Landmarks	HoG	Landmarks + HoG	Landmarks	HoG	Landmarks + HoG	Landmarks	HoG	Landmarks + HoG	Image space	Image space
DPM+CFSS (N=435)	43.2 ± 3.4%	59.3 ± 2.3%	61.4 ± 3.5%	85.5 ± 3.2%	87.6 ± 2.9%	90.1 ± 3.7%	68.3 ± 5.5%	80.7 ± 2.0%	81.1 ± 4.2%	74.4 ± 3.0%	71.4 ± 5.4%
DPM+ERT (N=435)	45.5 ± 3.2%	66.4 ± 2.1%	66.9 ± 2.9%	87.1 ± 2.8%	90.3 ± 1.4%	91.3 ± 0.9%	82.3 ± 3.5%	84.1 ± 4.2%	86.7 ± 3.7%	82.0 ± 2.9%	76.5 ± 3.4%
SSH+HN (N=413)	53.0 ± 4.3%	69.2 ± 2.6%	72.4 ± 2.5%	93.2 ± 2.6%	92.0 ± 1.6%	94.9 ± 3.5%	87.2 ± 1.7%	86.2 ± 2.5%	90.3 ± 4.0%	82.2 ± 4.9%	80.1 ± 2.0%
HA (N=437)	51.7 ± 4.5%	66.9 ± 5.2%	67.6 ± 4.6%	90.1 ± 1.9%	92.9 ± 1.5%	94.5 ± 2.1%	90.3 ± 2.7%	86.0 ± 1.6%	92.2 ± 1.1%	77.2 ± 1.6%	76.2 ± 7.7%

Table 2.1: Performance evaluation

facial weakness, and '5L' denotes left facial weakness. We choose the videos with a median score that larger than '4' as the left/right facial weakness and the videos less than a median score of '2' as the normal, resulting in 48 videos with left facial weakness, 53 videos with right facial weakness, and 98 videos with normal. Finally the frame with maximum smile activation in each video is manually selected by a graduate student and further verified by an experienced neurology trainee. This resulted in 199 additional images. The second row in Fig. 2.8 shows the rating distribution for three board-certified neurologists for the 203 videos. Note that the x-axis is the score rated by neurologists and the y-axis is the corresponding data percentage. The corresponding Fleiss Kappa value is 73.3% [CI 68.8%-77.9%]. Finally, total together the image dataset ends up with 437 images including 186 normal images, 125 left facial weakness images, and 126 right facial weakness images. Each image only contains one person. Moreover a graduate manually annotates the facial landmarks using a semi-automatic image annotation tool [104]. This tool first locates facial landmarks on the image, then the graduate student manually inspects and corrects the facial landmarks, which are denotes as Hand-annotated facial landmarks (HA, N = 437, Normal: 186, Left: 125, and Right: 126). Three ROIs: near-eye region, near-mouth region, and the full face, are evaluated in the experiment. After the normalization, near-eye region image is resized as 64 by 224 pixels, near-mouth region is resized as 128 by 200 pixels, and the full face is resized as 256 by 256 pixels. The stratified 5-fold cross-validation is used in the experiments shown below. The accuracy averaged over these 5 times is reported. The experiments results are summarized in Table 2.1. We present the evaluation results of the proposed model using concatenated landmark features and HoG features across different ROIs (e.g, near-eye region v.s. near-mouth region v.s. full face), as well as using either landmark features or HoG features separately.

First, we report the performance of the face detection and facial landmark extraction algorithms that used in this study. The DPM fails to detect faces in 2 images, while SSH fails to detect faces in 24 images. We find that the SSH experiences difficulty detecting face in the images where the person is closed to the camera and the forehead is missing from the image. In total, the DPM+CFSS and the DPM+ERT successfully extract 435 landmarks from the image dataset (N = 435, Normal: 186, Left: 124, and Right: 125) while the SSH+HN successfully extracts 413 facial landmarks from the image

dataset (N = 413, Normal: 176, Left: 118, and Right: 119). Table 2.1 demonstrates the performance of various classification schemes using different landmarks extraction algorithms (e.g, DPM+CFSS v.s. DPM+ERT v.s. SSH+HN v.s. HA) and ROIs (near-eye v.s. near-mouth v.s. full face) in terms of accuracy. It is worth noting that HA achieves a better performance than the DPM+CFSS and the DPM+ERT method (N=435). The SSH+HN method also has a good performance except that 24 images are excluded for analysis due to face detection failures (N=413). Another observation from Table 2.1 is that the HoG features have higher classification accuracy than landmark features for the most cases. This is because the HoG features are able to tolerate some local lighting and translation variations. In some cases where the facial landmarks can only delineate the coarse shape of the mouth, some important information with respect to facial weakness such as the mouth corner and lip edge shape is missing, causing incorrect prediction if the geometric features calculated from the mouth landmarks are used. When using the combination of the landmark features and HoG features, the classification accuracy can be improved. This indicates that combining the landmark features and the HoG features is able to provide extra useful information for classification, thus may mitigate some error caused by inaccurate landmarks extraction algorithms. To be specific, only using the landmark features discard some intensity information such gradient and edge that are significant clues for facial weakness classification. For instance, the contour of the mouth can be represented using edges and gradients. The mouth of health control subject has smooth contour while the facial weakness subject has irregular and sharp mouth edges. Combining the HoG features, however, allows the classifier to make use of these gradient and edge information for classification. Moreover, the HoG features can be treated as a nonlinear function that maps the edge orientation of the original image into a specified orientation arrangement. This histogramming process as well as block normalization enables the HoG features can handle some local lighting and translation variations [105]. Finally experiment evaluations show that near-mouth region has a higher pathological manifestation of facial weakness than near-eye region and full face in this dataset. Table 2.1 also provides the evaluation results for deep learning method. This follows the same trend that near-mouth region is more informative than full face. The deep learning based method has a lower performance in our dataset, one possible reason is that the data set used in our case is relatively small and the deep learning method typically requires a substantially larger amount of training data to be effective.

Table 2.2 provides the confusion matrix of our method and the deep learning based method for the near-mouth region. In additional to have a higher performance, the proposed approach makes fewer laterality errors such as misclassifying the left facial weakness as right facial weakness or vice versa. This is clinically meaningful because correctly determining the affected (weakness) side is able

Landmarks + HoG	Predict Normal	Predict Left	Predict Right	CNN	Predict Normal	Predict Left	Predict Right
Actual Normal	175	5	6	Actual Normal	180	2	4
Actual Left	9	114	1	Actual Left	24	93	7
Actual Right	11	2	112	Actual Right	26	7	92

Table 2.2: Confusion matrix

to help clinical professionals to locate which part of the brain may be damaged. Another observation is that deep learning based method is more likely to predict the result as the normal, one reason is that the dataset have more normal samples than left and right weakness samples.

Model transparency In addition, to improve the model transparency and interpretability, we visualize the weights of the linear SVM classifier for HoG features and landmarks features, because several studies suggested that a larger weight in the linear SVM classifier indicates more significant role in the decision function [106][107]. The visualization allows us to find the most significant texture and shape information regarding facial weakness classification.

We only focus on the near-mouth region in this section. Fig. 2.9 shows the average images, average HoG features, and average mouth landmarks over the training examples for left and right facial weakness and those without facial weakness. An observation is that these images can be treated as the 'template' for each class to show the texture (e.g., HoG features) and shape (e.g, landmark features) information variations across three different classes. As discussed above, the linear SVM classifier used in this study employs an one-to-one classification setting, meaning that $K = C(C-1)/2 = 3$ binary classifiers are constructed. We denotes w^{ij} as the weight of the binary classifier for class i against class j, where $i \in \{\text{Normal}, \text{Left}, \text{Right}\}$, $j \in \{\text{Normal}, \text{Left}, \text{Right}\}$ and $i \neq j$. Below we detail the procedures to visualize the maximum weight and its corresponding orientation in each cell for these three binary classifiers, which will be able to help us to understand what is learned for the training dataset. When extracting the HoG features, the near-mouth region is divided into 16 by 25 cells. A histogram with 9 bins is constructed to represent each cell and each bin is associated with a linear SVM weight. Then, in each cell we only show the maximum SVM weight, resulting in a 16 by 25 'pixels' image. In Fig.2.10, the first column shows the maximum SVM weight of three binary classifier ('Normal v.s Left', 'Normal v.s. Right', and 'Left v.s. Right' from

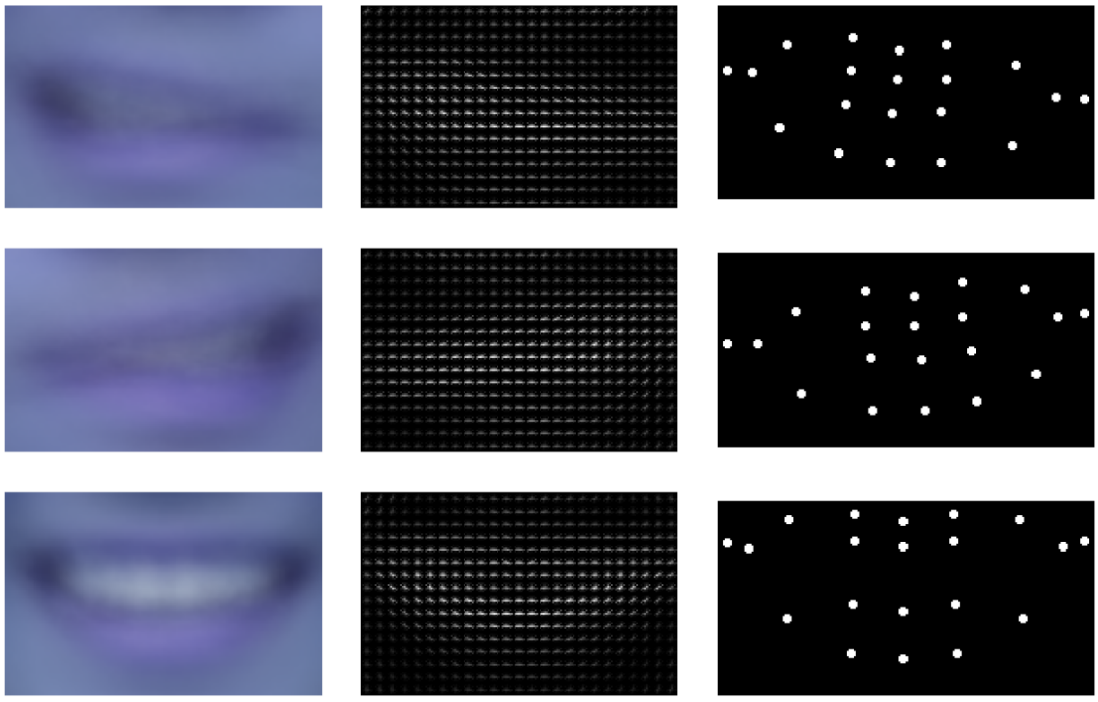

Figure 2.9: The average mouth image, HoG features, and landmarks for left, right, and normal cases.

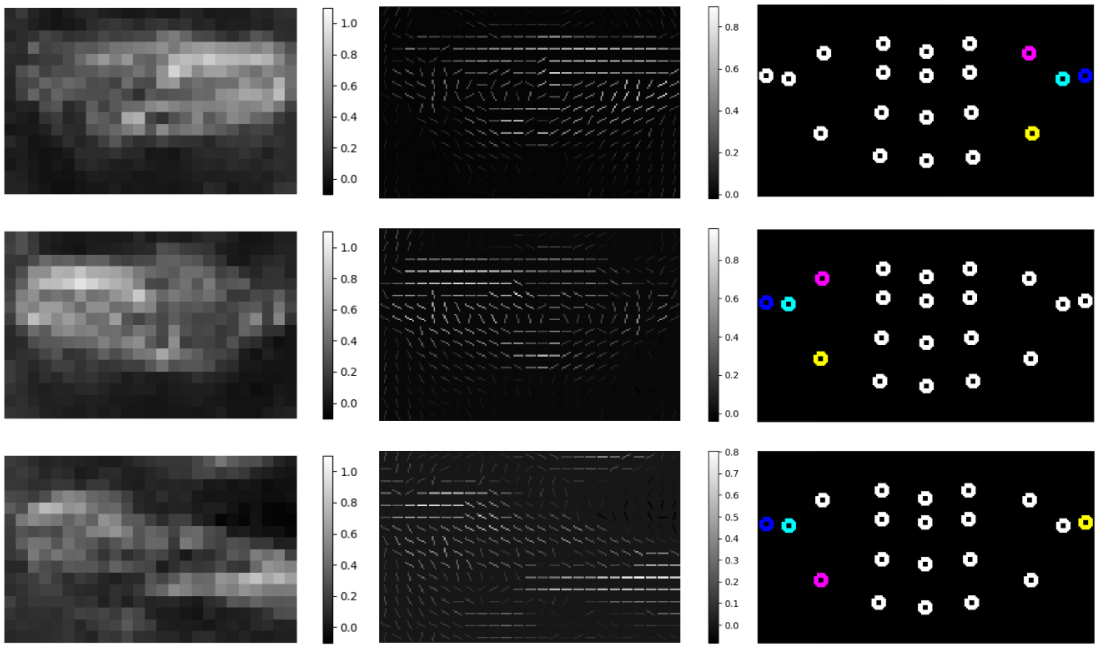

Figure 2.10: The significant cells for HoG features and the significant landmarks.

top to bottom) for the HoG magnitude. Similarly the corresponding orientation is shown in the second column in Fig.2.10. This visualization process enables us to understand what information is crucial for facial weakness classification. For instance, when classifying 'Normal v.s. Left', the cells on the left side of the mouth have larger weight. This makes sense as the patient with left facial weakness is unable to show the smile on the left side while a normal control can. Therefore, gradient information (magnitude and orientation of the edge) on the left side of the normal is more informative to classify the left facial weakness. Likewise in the third column of Fig. 2.10 we rank the importance of mouth landmarks using different colors (dark blue > cyan > magenta > yellow). For example, when classifying the 'Normal v.s. Left', the landmark on left mouth corner (in blue color as shown in the upper-left corner of Fig. 6) is more important than others. This is because the position of this landmark is in a lower position for the subject with left facial weakness compared with the healthy controls due to the fact that the subject loses the control of facial muscles on the left side and cannot raise the left mouth corner accordingly. While the intact side (right side) of the facial has some deformation, but it may not be as useful as the deformed landmarks of the symptomatic side (left side). The figure in upper-left corner of the Fig. 6 shows that the linear SVM classifier is able to learn this most discrimination information on the affected side from our dataset. Thus, during the testing phase of the classification procedure, the linear SVM classifier will assign a larger weight on this landmark to classify 'Normal v.s. Left '. Likewise, we identify three other important landmarks and rank them based on their weights, as shown in upper-left corner of Fig. 2.10, all of them are located on the affected side.

2.4 Summary

Through the two studies discussed above, we demonstrate that landmarks-based approach is an effective method for facial weakness analysis [67]. However, inaccurate landmark localization issues cause decreased classification accuracy. To address these issues, the second study [108] shows that incorporating texture-based information together with landmarks-based features leads to improved performance in terms of facial weakness classification accuracy. The knowledge learned from this chapter lays a solid foundation, in order to build a full-fledged video-based solution for facial weakness detection and quantification.

Chapter 3

Video-based facial weakness detection and quantification

3.1 Overview

The previous section outlines several methods for facial weakness detection on static images [67, 68], including identifying the optimal features regarding facial weakness analysis. However, assessing facial weakness solely from a single static image is somewhat limited, because of the availability of the static image containing pathological information. While a doctor or trained technician, could potentially acquire such an image, it would be preferable if a simple video examination procedure could be used instead. In addition, the video examination contains more information (many frames) it could be potentially more robust in identifying the signs of deficit. Fig. 3.1 demonstrates a facial weakness video example. At the beginning when the subject is in neutral expression, clearly it is difficulty to assess the presence of facial weakness. However, when the subject is asked to show a smile, the facial weakness symptom becomes more distinguishable.

Figure 3.1: A video examination example facial weakness.

Figure 3.2: Illustration of a typical user scenario.

Therefore, using the knowledge learned from Chapter 2 as building blocks, we investigate approaches for video-based facial weakness classification, including two major studies: majority voting-based approach [69] and RNN-based approach [70]. The majority voting method performs the classification for each single frame and aggregates the frame-wise classification results using a majority voting strategy as the overall classification result for the input video. However, this method has several drawbacks, such as misusing the frames in which subjects display no symptoms of facial weakness or failing to take the temporal information into account. To address these issues, we build a novel automated and quantitative facial weakness screening framework, which not only automatically identifies the frames with maximum muscle activation but also models the temporal dynamics of both shape and appearance-based features of each target frame through a Bi-directional Long Short-term Memory network (Bi-LSTM). To evaluate the proposed systems, we assemble, curate, and verify an "in-the-wild" video dataset, which is verified by three board-certified neurologists and rated by multiple paramedics and neurology trainees. Evaluation results show that the proposed framework achieves equivalent performance to paramedics and provides visualizable and interpretable results.

Significance of the proposed work is that it can be beneficial to assist the paramedics to identify the facial weakness in the field or, more importantly, whenever expertise in neurology is not available

either for emergency stroke patient triage or chronic disease management, leading to increasing coverage and earlier treatment. One usage example is shown in Fig. 3.2. Furthermore, this study, as a proof-of-concept, shows that such technology could be a potential solution to the lack of neurologists nationally and globally, serving as an essential blueprint for future innovations in the field.

3.2 Related work

Several works have been proposed to automate face weakness assessment in video data. The literature leads to three main approaches depending on the type of features used: (1) shape-based approaches; (2) appearance-based approaches; (3) depth-based approaches.

Analyzing the anomalies in the facial geometric cues have been most popular approach to perform facial weakness detection. Researchers often detected facial landmarks and directly measured the facial geometric features, e.g, the distance and angles between landmarks, which were used to perform the classification on a single RGB image or a sequence of images [78, 79, 77, 81, 67]. For example, Gaber *et al.* [77] implemented a Kinect-based system to quantify facial paralysis by calculating a symmetry index for the eyebrows, eyes, and mouth. Guo *et al.* [81] computed the location and displacement of the landmarks to formulate the shape-based features for classifying the facial weakness. However, the drawback of shape-based methods is that the current facial landmark extraction algorithms are typically trained and calibrated using normal facial configuration and may suffer from poor accuracy for patients with facial weakness.

The face also exhibits specific texture and appearance information. Together with the emergence of deep learning methods, researchers have investigated the appearance-based and texture-based features as an alternative. In [82, 83], the author employed the optical flow and local binary pattern on three orthogonal planes (LBP-TOP) to perform the facial weakness classification for videos. Guo *et al.* [85] devised a convolutional neural network (CNN) method to perform facial weakness severity classification on static image. Li *et al.* [84] extracted the intensity values and LBP features from various parts of the face as input features to a two-stage support vector machine (SVM) classifier to assess facial paralysis. Zhuang *et al.* [69] developed a facial weakness classification system using the histogram of oriented gradients features. Other works also investigated the facial weakness classification using a combination of shape and appearance-based features. Haase *et al.* [109] and Modersohn *et al.* [110] located the local patches on the face and used a combination of the shape and appearance-based features from the local patches to assess the asymmetry for static images. More recently, Xu *et al.* proposed a Dual-path LSTM network to evaluate the facial weakness [111].

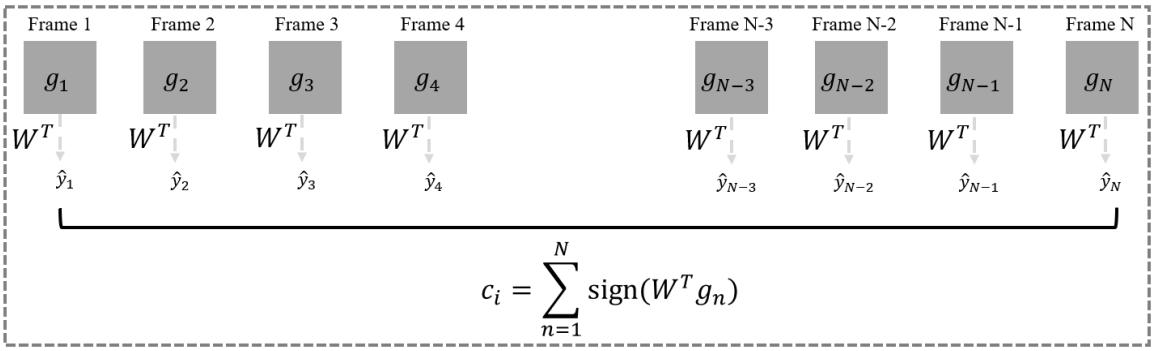

Figure 3.3: A majority voting based approach for facial weakness video classification.

The local and global features for each frame were extracted by two autoencoder networks, then the LSTM modeled the temporal relationship and outputed the classification result. Storey *et al.* [112] proposed a 3D-ResNet network to classify facial weakness using the near-mouth region.

Several studies leveraged depth information to assess facial asymmetry. Bandini *et al.* [113] utilized a RGB-D camera to conduct the facial movement analysis for patients with amyotrophic lateral sclerosis. Alagha *et al.* [114] analyzed the dynamics of facial expressions in unilateral facial palsy using a sequences of 3D images. Desrosiers *et al.* [115] proposed to extract Dense Scalar Field features from a sets of 3D face images to evaluate the facial weakness treatment. The downsides of depth-based techniques can be hindered by the need of dedicated hardware and the fact that 3D reconstruction from depth information is often a delicate procedure that can have accuracy issues.

3.3 Video-based facial weakness detection using majority voting

Overview The previous chapter demonstrates the effectiveness of gradient information used for facial weakness analysis of static images. To here, goal of the first study [116] is to classify an examination video, or a sequence of individual frames, using the most straightforward approach, which is to perform frame-by-frame classification and to combine a sequence of classification results into one single result. Fig. 3.3 illustrates the proposed idea. Using this idea, we develop a simple but effective video classification method [116], named "F-DIT-V", for facial weakness detection. In the following sections, we will provide the details regarding F-DIV-V.

Preprocessing and feature extraction Given an input video, F-DIT-V first decomposes it into a total of N individual frames. For each frame, F-DIT-V extracts the corresponding facial

landmarks and aligns each frame to remove undesired variations in terms of the size, displacement, and orientation. After the alignment, the frames are cropped around the subject's mouth, since facial weakness is primarily manifested in the near-mouth region. The HoG features $\mathbf{g}$ are extracted and will be used for frame-wise classification.

Predictive model The predictive model consists of a pLAD classifier and a voting classifier. The pLDA algorithm outputs the classification result for each frame, and the voting classifier aggregates image-level predictions into a single classification result: normal, left facial weakness, or right facial weakness. The goal of the standard LDA method is to maximize the Fisher criterion, while the pLDA algorithm [93] used in this study aims at maximizing the Fisher criterion and minimizing a penalty term at the same time, which is formulated as:

$$\max_{\boldsymbol{w}} \left(\frac{\boldsymbol{w}^T S_T \boldsymbol{w}}{\boldsymbol{w}^T \left(S_W + \alpha \boldsymbol{I}\right) \boldsymbol{w}} \right) \tag{3.1}$$

where $S_B = \sum_c N_c(\mu_c - \bar{\mathbf{g}})(\mu_c - \bar{\mathbf{g}})^T$ represents the "between scatter matrix", $S_W = \sum_c \sum_{i \epsilon c}(\hat{\mathbf{g}}_i - \mu_c)(\hat{\mathbf{g}}_i - \mu_c)^T$ defines "within classes scatter matrix", μ_c is the center of class c, $\bar{\mathbf{g}}$ is the center of all dataset, N_c is number of instance in class c, $S_T = S_B + S_W$, α is the penalty term, and I is the identify matrix. Once a set of directions $\boldsymbol{w}_k$, $k = 1, \cdots, K$ has been obtained from the pLDA procedure, we assemble a projection matrix W where each column is composed of an eigenvector from the problem above $\boldsymbol{w}_i$. The classification result for each individual frame is predicted as $y_n = W^T \mathbf{g}_n$, $n = 1, \cdots, N$.

After obtaining the classification result for each frame, a voting classifier combines the image-wise results into a single prediction result. Specifically, the voting classifier first counts how many frames belongs to each class in the given video with N frames:

$$C_i = \sum_{n=1}^{N} \text{sign}(W^T \mathbf{g}_n) \tag{3.2}$$

where i represents facial weakness class (e.g. normal, left facial weakness, right facial weakness) and C_i is the total number of frames for i_{th} class. Then, the voting classifier determines the presentence of facial weakness. If the facial weakness present, it compares the duration of facial weakness (left or right) with a predefined threshold τ and predicts the classification result.

Experiment evaluation Given the absence of a publicly available annotated stroke deficit video dataset, we gathered videos of healthy controls and patients with facial weakness from publicly

available online repositories such as YouTube. In these videos, there exists a large amount of variations of pose, lighting, distance, appearance, and environment as commonly observed "in-the-wild". The videos were reviewed independently by two senior resident neurologists. Each video was given a numerical score ranging from "1" to "5", which "1" denotes high likelihood that pathology is absent and "5" denotes high likelihood that pathology is present. Only videos rated concordantly were used for the study, resulting in 37 videos with left facial weakness, 38 videos with right facial weakness, and 60 normal videos. The total number of frames for each video ranges from 34 to 200. Video pre-processing was comprised of face landmark extraction, alignment, and cropping, resulting in a set of 64 by 128 pixel near-mouth region images for each video. Each individual frame in the video was labeled and verified by one senior neurology resident and one M.D. student to train pLDA classifier. The stratified five-fold cross-validation was used in the experiment. In this work, the number of orientation bins in a cell is 9 and a cell consists of 8 by 8 pixels. The number of cells in each block is 4. Since the extracted HoG features are high-dimensional, we compute the principal component coefficients from training dataset to reduce feature dimension and avoid over-fitting. In our case, we reduce the dimensions of the features to the components that can cover 95% of the variance. We compared the proposed algorithm with state-of-art alternatives, including the LBP-TOP method and a RNN-based method. Specifically, the LBP-TOP method was configured using following the setup: the parameters of the radii along horizontal axis (X), vertical axis (Y), and time axis (T) were 1, 1, and 2 respectively, and the number of neighbor points in the XY plane, XT plane, and YT plane were 8. For each video, the LBP-TOP model outputs a 59×3 vector which specified LBP features along the XY, XT, and YT planes. Finally, this vector was fed into a linear SVM classifier for classification. The RNN-based method exploits and models the temporal relationship between each individual frame and the sequence label. In our study, the RNN model was implemented using a long-short term memory (LSTM) network with 100 hidden units. The last hidden output was used as input to a linear layer to perform classification. The network was optimized by minimizing the cross entropy loss using an Adam optimizer.

The accuracy , precision , recall, and specificity of the proposed F-DIT-V method are 92.9%($\pm$3.2%), 93.6%($\pm$2.9%), 92.8%($\pm$3.4%), and 94.2%($\pm$2.6%), respectively, which achieves the best performance. Only very few videos are misclassified, because the pLDA classifier is unable to predict enough correct number of frames for a given video, leading the voting classifier to perform the incorrect voting and predict a incorrect result. The LBP-TOP method achieves mediocre performance with accuracy of 80.7%($\pm$6.2%), precision of 83.8%($\pm$5.8%), recall of 80.8%($\pm$6.1%), and specificity of 83.0%($\pm$5.4%). The LBP-TOP method has a lower performance evaluation because it

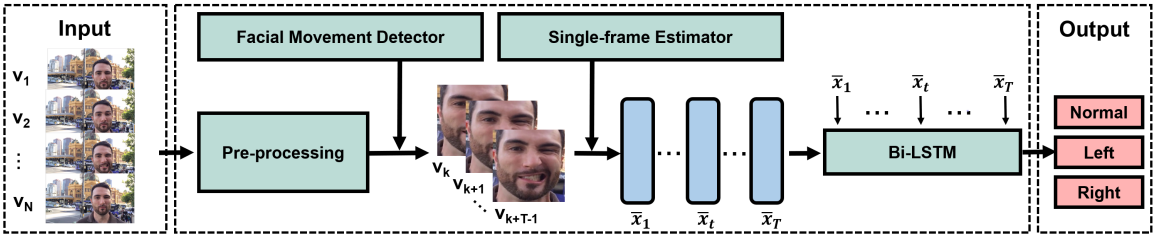

Figure 3.4: Architecture of the RNN-based approach.

only extracts textual information from three orthogonal planes, meaning that some information is lost which may result in a decrease in performance. The LSTM-based method has the lowest performance with the highest deviation. The accuracy, precision, recall, and specificity are 77.9%(±7.8%), 78.0%(±7.7%), 77.8%(±7.7%), and 81.9%(±8.8%) respectively. The relative poor performance of LSTM network on our dataset is caused by the fact that the number of videos in our dataset is relative small while the LSTM-based method normally requires a substantially large amount of training dataset to be effective. However, it is worth noting that the temporal information used in LSTM-based method is not incorporated in our framework. Our next research will include leveraging the temporal information to enhance the performance of our proposed method in term of accuracy and robustness.

3.4 Video-based facial weakness detection and quantification using RNN

Overview Although the F-DIT-V approach [116] is an effective solution for video-based facial weakness classification, it is worth further exploring a new method that is able to: (1) automatically identify a video segment of interested, where the subject's muscle activation is maximized, (2) and exploit the spatial temporal relationship among video frames. Therefore, we propose a new framework [70] that is capable of not only addressing the aforementioned issues, but also leveraging all useful information and knowledge that are learned from previous studies, as shown in Fig. 3.4. To be specific, for a given input video, the proposed framework extracts facial landmarks and performs landmarks and intensity normalization that removes translation, rotation, and scaling variations. Next a facial movement detector employs the optical flow approach to measure the face movement intensity and locates an video segment where a smile configuration is clearly evident, because our intent is not to use all the frames but only use the frames that have the maximum muscle activation

Table 3.1: Notation

	Notation
T	Number of target frames
x	Input feature (either shape-based or appearance-based feature)
W_{plda}	Projection matrix for pLDA analysis
W_{pca}	Projection matrix for PCA analysis
$\hat{x}$	Output of either shape-based or appearance-based features from the single frame estimator
$\bar{x}$	Concatenated shape-based and appearance-based features
t	Frame/time index
h_t^l	Hidden state of Bi-LSTM in layer l at frame/time t
$f(\cdot,\cdot)$	LSTM hidden state update function
W_y	Projection matrix of affine layer
$\hat{y}$	Classification prediction

for assessing facial weakness. Once the desired video segment is obtained, the shape and texture-based features are extracted from the target frames inside the desired video segment. Then the single-frame estimator projects the extracted high-dimensional shape and texture-based features onto a low-dimensional subspace. Finally the framework classifies the input video using the low-dimensional representation of shape and texture-based features via a Bi-LSTM network. The final output of the framework is left facial weakness (left), right facial weakness (right), or normal control (normal). The following paragraphs present details regarding the RNN-based framework for video-based facial weakness classification.

preprocessing The preprocessing step decomposes the video into a sequence of individual frames and detects faces within each video sequence correspondingly. Then a standard face alignment algorithm is employed to align the facial landmarks and pixel intensities as discussed in our previous studies [68] to eliminate the translation, scaling, and rotation effects, due to the fact that the facial images contained in videos are subject to random location, orientation, and size variations. After the alignment, only the region of interest (ROI) is kept. It is worth noting that we evaluate the facial weakness classification scheme on two ROIs: the near-mouth region and full face in the experiment section. Finally, we make use of the aligned 68 anatomically significant facial landmarks as shape-based features and HoG features as the appearance-based features, as shown in Fig. 3.5, respectively.

Facial movement detector After the face normalization, rather than utilizing every frame in the whole video, we are particular interested in a video segment, in which the subject's facial expression is maximized. Hence, the goal of the facial movement detector is to locate the desired video segment that has the maximum muscle activation, as shown in Fig. 3.6. Because the facial movement can

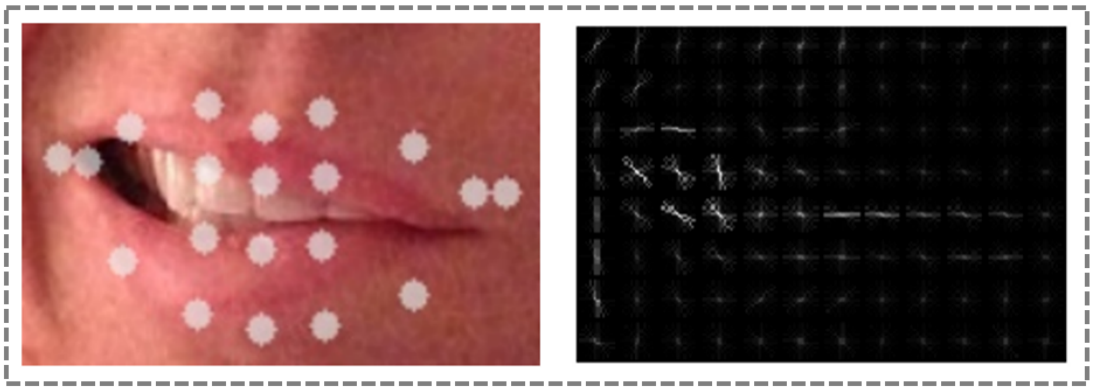

Figure 3.5: The shape-based (landmarks) features and texture-based (HoG) features.

be characterized by the displacement of pixels, the optical flow method is employed to identify the full smile activation and locate the desired video segment. The first frame serves as the reference frame I_0, The optical flow estimation of the subsequent frames are computed with respected to the reference frame I_0. For example, consider a pair of reference frame I_0 and target frame I_n (n_{th} frame), and a point $u = [u_x, u_y]^T$ on I_0 at (x, y), the goal of optical flow estimation is to find a "similar" point's location of $v = u + d = [u_x + d_x, u_y + d_y]^T$ on I_1 [117]. The optical flow vector d, which describes the movement of point u in I_1 with respect to I_0, consists of two components: d_x and d_y. The corresponding magnitude is computed as $|d|_{xy} = \sqrt{d_x^2 + d_y^2}$. The total pixel movement magnitude for frame n, therefore, can be computed as $D_n = \sum_x \sum_y |d|_{xy}$. Next, we use a sliding window to calculate the total magnitude for the T frames inside the window. Then we select the window that has the largest magnitude as the target window. All T target frames inside the detected window are used for further analysis. The index of the detected window can be obtained by $\text{argmax}_k \frac{1}{T} \sum_{n=k}^{k+T} D_n$.

Single-frame estimator The single frame estimator aims at extracting the most discriminant shape and appearance information related to facial weakness from a single frame. In order to construct such an estimator, we model the pathological meaningful shape and appearance variation on an neurologist-verified image dataset that is independent of the video dataset in a supervised-learning fashion. We utilize a composition of the principle component analysis (PCA) [118] and penalized linear discriminant analysis (pLDA) [93] method to perform the statistical shape and texture analysis, which is able to learn the discriminating pattern to separate between multiple classes (normal v.s. left v.s. right). This is can be represented mathematically as $\hat{x} = W_{\text{plda}}^T W_{\text{pca}}^T x$, where x denotes either the shape feature vector or appearance feature vector. These projection

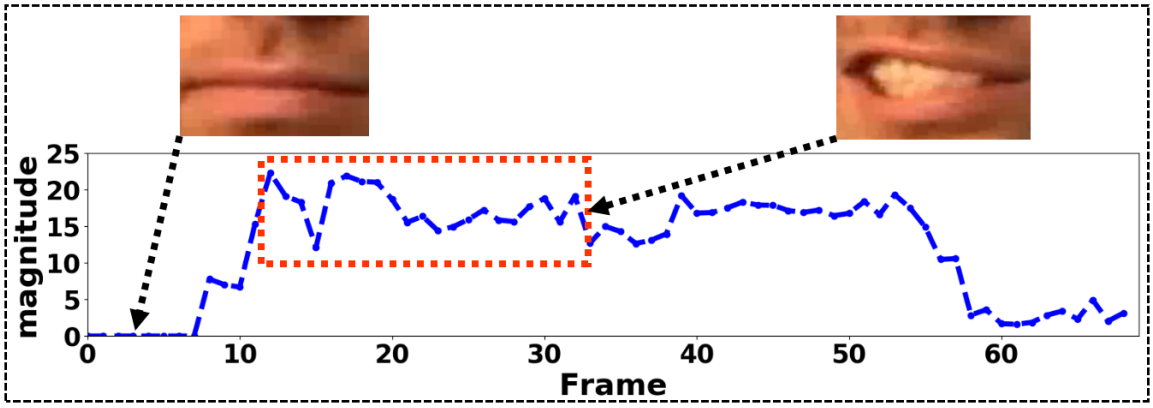

Figure 3.6: The facial movement detector locates the video segment with the maximum muscle activation.

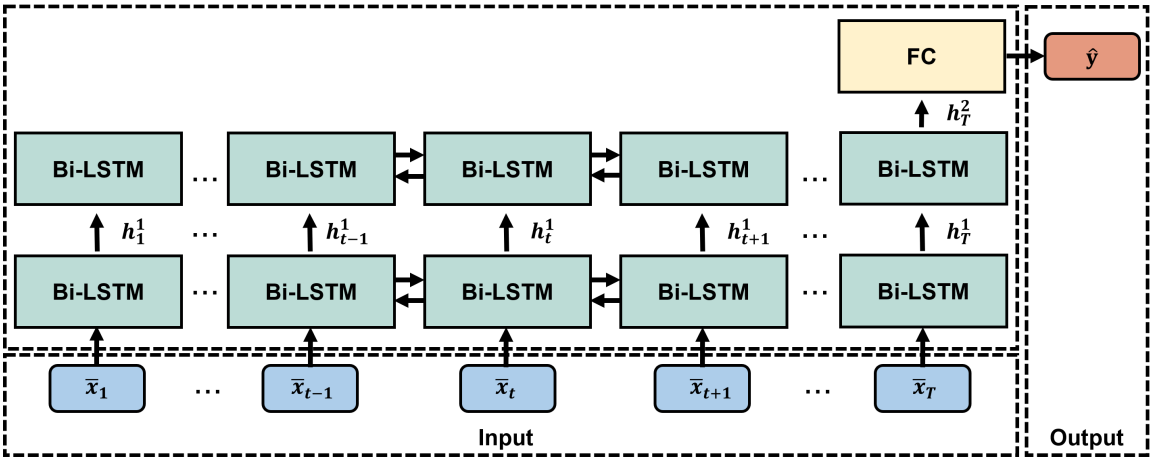

Figure 3.7: The network structure of the Bi-LSTM used in this study.

matrices, W_{pca} and W_{plda}, can be estimated using the standard PCA and pLDA techniques [118, 93]. It is worth noting that each column in the projection matrices W_{plda} and W_{pca} is composed of an eigenvector and defines a set of directions in the low-dimensional PCA and pLDA subspace. Using this single frame estimator not only allows us to identify the most discriminant pattern of facial weakness and provides the visualizable and interpreable result, but also facilitates the fast computation by representing high-dimensional data in a compact form. To this end, the single frame estimator generates a feature sequence by directly concatenating the estimated shape and appearance-based features for each frame in the detected target video segment $\{v_t\}_{t=k}^{k+T-1}$ as $\{\bar{x}_t\}_{t=1}^{T}$, where k is the index of the first target frame.

Temporal modeling using RNN The temporal modeling algorithm in our study seeks to predict the label $\hat{y}$ for the feature sequence $\{\bar{x}_t\}_{t=1}^{T}$ via a recurrent neural network (RNN) based approach. The RNN network is able to learn the temporal relationship between the image sequences and maps the learned temporal information to a sequence label [119]. We use a two-layer Bi-LSTM network to implement the RNN as shown in Fig. 3.7. To be precise, given the input feature sequence$\{\bar{x}_t\}_{t=1}^{T}$, the first Bi-LSTM layer computes the hidden state $h_t^1 = [\overrightarrow{h}_t^1, \overleftarrow{h}_t^1]$ by concatenating the forward hidden state $\overrightarrow{h}_t^1$ and backward hidden state $\overleftarrow{h}_t^1$ at time t, while the $\overrightarrow{h}_t^1$ and $\overleftarrow{h}_t^1$ can be calculated as:

$$\begin{aligned} \overrightarrow{h}_t^1 &= f(\bar{x}_t, \overrightarrow{h}_{t-1}^1) \\ \overleftarrow{h}_t^1 &= f(\bar{x}_t, \overleftarrow{h}_{t+1}^1) \end{aligned} \tag{3.3}$$

where $f(\cdot, \cdot)$ refers to the standard LSTM update equation in a LSTM cell [119], $\bar{x}_t$ is the input at time t, h_{t-1}^1 and h_{t+1}^1 are hidden state at $t-1$ and $t+1$, and the superscript l of h_t^l is the l_{th} layer of LSTM network. Likewise, using equation (3.3), the $h_t^2 = [\overrightarrow{h}_t^2, \overleftarrow{h}_t^2]$ can be computed at time t. Finally, a dense layer outputs the classification result $\hat{y}$ for the given input sequence as:

$$\hat{y} = W_y \cdot h_T^2 + b_y \tag{3.4}$$

where W_y and b_y are the weight matrix and bias item of the fully-connected layer.

Deep learning-based comparables To evaluate the performance of the proposed algorithm, we compared it with the LBPTOP based approach [82], 3DResNet (or 3DPalsyNet) [112], Dual-path LSTM [111], 2DCNN+RNN [120], and Two-stream LSTM [121]. To be specific, we use the following parameters to configure the LBPTOP approach: the radius along X axis, Y axis, and T axis are 1, 1, and 2 respectively, and the number of neighbor points are 8 for the XY plane, XT plane, and YT plane. The 3DResNet and Dual-path LSTM are implemented as discussed in [112, 111]. The shallow three-dimensional convolutional neural networks (3DCNN) serves as the performance baseline for deep learning based method, which consists of two convolutional layers. The number of filters and kernel size of first convolutional layer are 32 and 5x5x5, while the second layer are 48 and 3x3x3 respectively. A batch normalization (BN) layer is used after each convolutional layer. Then a ReLu layer, a dropout layer with drop rate of 0.2, and a maxpooling layer with kernel size of 2 generate the feature maps. Finally, a composition of two affine layers makes the classification. Another popular deep learning based video classification alternative makes use of a combination of CNN and RNN.

To be specific, the CNN network learns the optimal representation of each individual frame while the RNN network models the changes of these spatial features along the temporal axis. We also implement a CNN+RNN (or CNN+LSTM) network based on [120]. The CNN has 4 convolutional layers with kernel numbers of 64, 128, 256, and 512, whose kernel size are 5x5, 3x3, 3x3, and 3x3 respectively. Then each convolutional layer is followed by a BN layer and a ReLu layer. Then a dropout layer with dropout ratio of 0.5 is applied. A global maxpooling layer with kernel size of 2 reduces the dimension of feature map from the CNN. A composition of two fully connected layer with 1024 nodes generates the feature map for each individual frame. Then the RNN network is implemented as a two-layer LSTM, each layer has 1024 hidden units. A fully connected layer with 512 nodes is applied on the final output of the LSTM network to produce the classification result. More recently, multiple studies indicate that incorporating the motion information, e.g, the optical flow [121, 122, 123], increases the video classification performance. We adopt this two-stream LSTM architecture as discussed in [121] for our study. To be precise, the two-stream LSTM utilizes one spatial CNN network to extract the spatial information from each frame and another motion CNN network to extract the motion information from the stacked optical flow, respectively. Then a LSTM network concatenates the outputs from these two CNN networks and learns the long-term dependency. In our implementation of the two-stream LSTM, its spatial CNN network shares the similar configuration with CNN+RNN method as discussed above. The motion CNN network has three convolutional layers. For each convolutional layer, a BN layer and a ReLu layer are applied thereafter. The kernel sizes for each convolutional layer are 5x5, 3x3, and 3x3. The number of convolutional kernel is 64, 128, and 256. After the ReLu layer, a maxpooling layer with kernel size of 2 is used. Then an affine layer with 1024 nodes produces the feature map. Finally, the feature maps from motion CNN network and spatial CNN network are concatenated and input to a two-layer LSTM with 1024 hidden units, which generates the classification prediction result.

Experiments overview We first describe the acquisition of two independent facial weakness datasets and the corresponding verification process by three board-certified neurologists followed by the experiment setup description. The subsequent section provides experiment results including the statistical shape and appearance analysis of the single frame estimator and performance comparisons with other approaches. Furthermore, the comparison between the proposed method and human raters is provided. At the end of this section, a prototype of our proposed approach as a proof-of-concept is present.

Figure 3.8: The facial weakness image dataset.

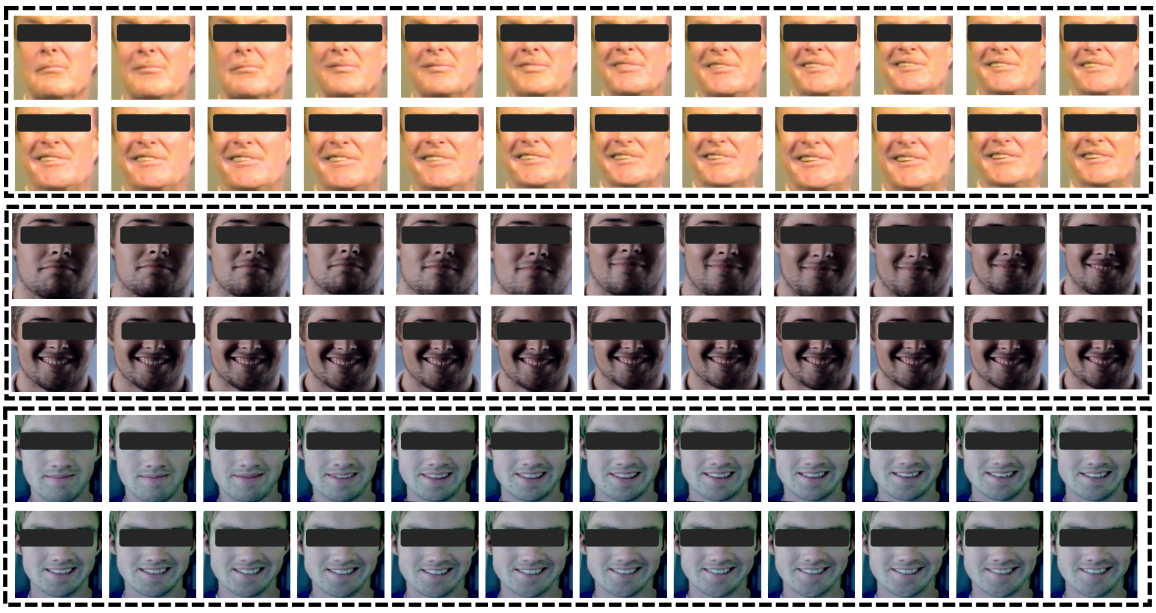

Figure 3.9: The facial weakness video dataset.

Data collection and validation Due to the fact that no facial weakness image and video dataset is publicly available, we assembled two independent datasets including an image dataset and a video dataset from publicly available repositories such as Google Images and Youtube. The image dataset was used to build the single-frame estimator, while the video dataset was used to evaluated the proposed video classification algorithm. It is worth noting that since neurologist diagnosis is still the gold standard for confirming the presence or absence of facial weakness, the image dataset and video dataset were verified by three board-certified neurologists using a modified NIH stroke score (NIHSS) [124]. To minimize the possibility that three clinical raters made the same mistake, the median score of their ratings serves as the ground truth. The image dataset acquisition and verification is described in [68]. In total, the image dataset consists of 236 images, including 88 normal control images, 76 left facial weakness images, and 72 right facial weakness images, as shown in Fig. 3.8. Then it is augmented by flipping the images horizontally. The video dataset was collected by three senior medical students. The data acquisition protocol included the videos containing exactly one subject who faces the camera directly with full face in view and has a neutral expression at the beginning and shows a prominent smile thereafter to ensure the presence of maximum muscle activation. Three board-certified neurologists independently reviewed the videos. To classify presence or absence of face deficits, the 5-point scale NIHSS scores were converted into a modified 3-point scale: 1 denotes pathology absent, 2 denotes pathology indeterminate, and 3 denotes pathology present. Then we computed a median score from the three neurologists to serve as the ground truth. The videos with a median score of one or three were selected in our study, resulting in 43 left facial weakness videos, 50 right facial weakness videos, and 96 normal videos (72 men v.s. 117 women and 155 light-skinned v.s. 34 dark-skinned in terms of demographics information), as shown in Fig. 3.9.

Experiments setup This section specifies the experimental setup. We set the window length T, which equates to the length of the subject having maximum muscle activation, as 0.83 seconds (20 frames in our case). The ensemble of regression trees (ERT) algorithm [91] was chosen to perform facial landmark extraction owing to its high performance [100]. Two ROIs were analyzed in our study:near-mouth region and the full face. After the normalization, near-mouth region was resized as 128 by 200 pixels and the full face was resized as 256 by 256 pixels. The parameter configuration for the HoG features was: the number of orientation bins in each cell was 12 and a cell consists of 16 by 16 pixels. In terms of parameter setup for the PCA method, the components that can cover 98% of the variance of HoG features were kept and the components that can cover 96% of the

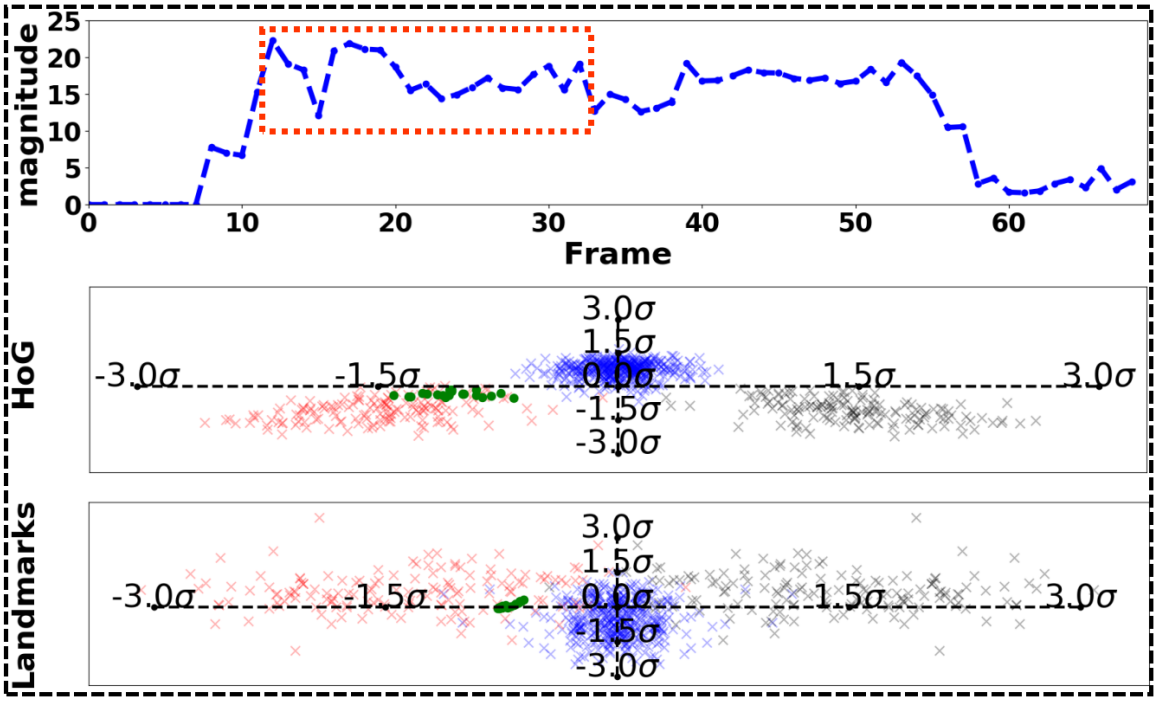

Figure 3.10: Left video classification example.

variance of landmark features were kept. For the pLDA method, the α for HoG features was set to 10 and for landmark features was set to 0.1 according to [93]. The Bi-LSTM network consisted of two Bi-LSTM layers, each Bi-LSTM layer had 64 hidden units. We utilized the Adam optimizer to optimize the network. The learning rate for Adam optimizer was set to 0.001. The β_1 and β_2 were set to 0.9 and 0.999, respectively. We report the accuracy averaged over these 5 times using stratified 5-fold cross-validation scheme.

Video classification results We demonstrates three examples of video classification, namely left facial weakness v.s. normal v.s. right facial weakness in details, as shown in Fig. 3.10, Fig. 3.11, and Fig. 3.12, which illustrates three concrete examples about how the proposed approach works. First, the framework measures muscle activation and detects a relevant video segment as highlighted in the red window. Then, the single-frame estimator projects the shape and appearance-based features of each target frame inside this video segment onto the two-dimensional pLDA subspace as shown in the middle and bottom row. Each green dot represents a target frame inside the video segment. This shows that the learned optimal representation of facial weakness from the image dataset by the single-frame estimator is not overfitting and is effective for the video dataset.

Then we evaluate the performance of the proposed system in terms of accuracy. Table 3.2 presents the evaluation results of the proposed method for two different ROIs (near-mouth v.s. full face).

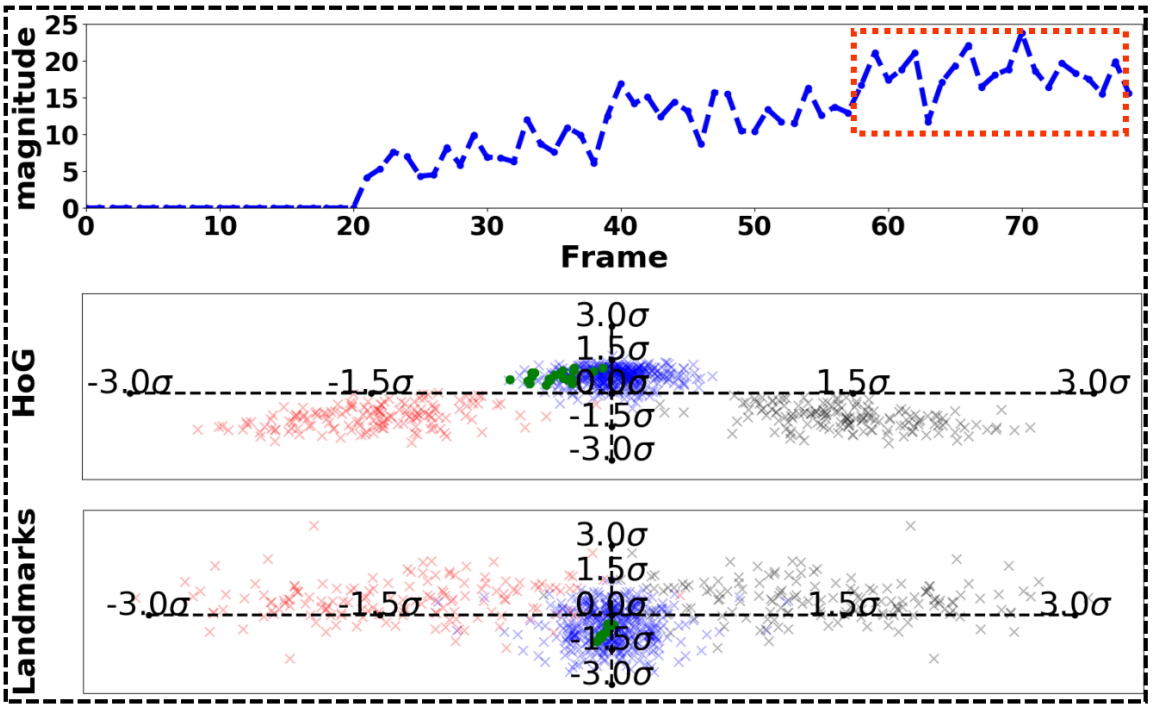

Figure 3.11: Normal video classification example.

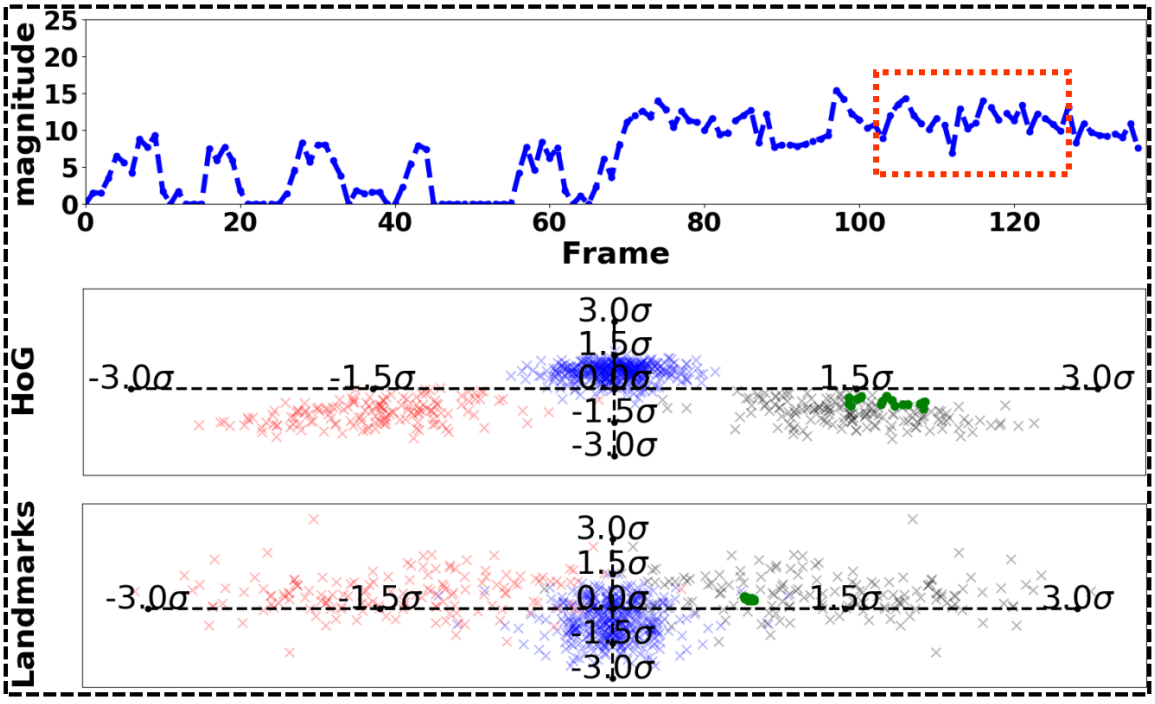

Figure 3.12: Right video classification example.

Table 3.2: Performance of the proposed method

	Acc.	Sens.	Spec.
Face	88.3±1.8%	82.5±2.7%	91.2±1.3%
Near-mouth	94.3±2.1%	91.4±3.2%	95.7±1.6%

The proposed method achieves the accuracy of 88.3%, sensitivity of 82.5%, and specificity of 91.2% for the full face region, and accuracy of 94.3%, sensitivity of 91.4%, and specificity of 95.7% for near-mouth region. However, we also note that there are several misclassified cases due to the lighting and appearance variations, which degrade the image quality and cause inaccurate ROI segmentation. The proposed method shows that the near-mouth region contains more discriminant information to identify facial weakness than that of full face, which is also verified by other studies [81, 82].

In addition, Table 3.3 presents the comparison results with other existing facial weakness detection alternatives. First, we note that the proposed algorithm outperforms other methods in terms of accuracy, sensitivity and specificity. To quantitatively assess the performance difference between the proposed algorithm and other comparison methods, a statistical test (Cochran's Q test, which is a generalized method of McNemar's test used for evaluating multiple classifiers [125, 126]) is conducted to show that the performance difference of our method with other methods shown in Table 3.3 is statistically significant (p-value <0.001). Secondly, the performance varies significantly among the other deep learning based methods. The 3DResNet achieves the best performance. When comparing with the shallow 3DCNN, a deep learning baseline method for comparison, the higher performance of 3DResNet demonstrates that both of the depth of the network and the architecture of the network affect classification accuracy. In terms of RNN-based methods' performance, the results demonstrate that LSTM network architecture is effective to learn the temporal discriminate information to classify facial weakness. The Dual-path LSTM and two-stream LSTM obtain a better performance compared with the CNN+LSTM approach, this is because the Dual-path LSTM takes local information from the image patches into account and the two-stream LSTM incorporates motion information from the optical flow features. Both operation increase classification performance [123].

Comparison with clinical raters In order to compare the performance of our algorithm with human raters, three EMS paramedics and three upper level residents rated our video dataset using the same protocol described above. The relative experience of the three paramedics included an Emergency Medical Technicians (EMT) with seven years of experience total and five years of experience as an advanced life support provider, a nationally registered paramedic with over 10 years

Table 3.3: Comparison results

	Acc.	Sens.	Spec.
LBP-TOP [82]	84.9±7.8%	77.3±11.8%	88.7±5.9%
Shallow 3DCNN	69.3±1.9%	54.0±2.9%	76.9±1.4%
Dual-path LSTM [111]	77.1±5.8%	65.6±8.7%	82.8±4.3%
3DResNet [112]	82.0±4.8%	73.0±7.2%	86.5±3.6%
CNN+LSTM [120]	71.4±5.1%	57.1±7.7%	78.58±3.8%
Two-stream LSTM [121]	80.5±9.9%	70.7±14.9%	85.3±7.48%
Bi-LSTM (ours)	94.3±2.1%	91.4±3.2%	95.7±1.6%

Table 3.4: Performance comparison with human raters

	Bi-LSTM	Paramedics	Residents
Accuracy	94.3% [90.2%-98.4%]	92.6% [90.1%-94.7%]	97.9% [96.4%-99.1%]
Sensitivity	91.4% [85.1%-97.7%]	87.8% [83.9%-91.7%]	96.4% [93.9%-98.5%]
Specificity	95.7% [92.6%-98.8%]	99.3% [98.2%-100.0%]	99.7% [98.9%-100.0%]

of experience, and an entry level EMT with one year of experience. The three neurology residents have three years of highly-focused and systematic neurology training. Table 3.4 shows the accuracy, sensitivity, and specificity with 95% confidence interval for the proposed framework, paramedics, and senior neurology trainees respectively. The results demonstrate that the senior neurology trainees achieve the highest performance for all three evaluation metrics. Our algorithm performance is reaching the level of residents and equivalent to the paramedics. The paramedics have the higher specificity as compared to the specificity of the proposed framework, while the proposed algorithm has better sensitivity. We also perform statistical tests for the comparison between the proposed algorithm and human raters using the McNemar's test [125, 126]. The statistical tests show that there is no significant difference between the performance of our algorithm and the paramedics' performance (p-value = 0.091) while there is a significant difference between our algorithm and the residents (p-value <0.001). Furthermore, in order to quantify the disagreement among the human rates, we compute the Fleiss Kappa scores (a measure of agreement among individuals within a group) for the paramedics and resident. The paramedics achieved a Fleiss Kappa statistic of 0.806 (95% CI [0.724, 0.888], p-value <0.001), while the residents had greater agreement at 0.921 (95% CI [0.866, 0.976], p-value <0.001). The disagreement among paramedics is larger possibly because of lack of extensive neurological training. The high Kappa score of the residents indicates that the rating among residents are more consistent and reliable. This may be related to the fact that the certified neurologists also trained the neurology trainees.

Model interpretability Here we show that the single frame estimator can capture the pathological facial weakness. Fig. 3.13 shows the results of projecting the shape and appearance-based features of training (in lighter color) and testing samples (in darker color) onto a smaller pLDA subspace. One observation is that the between-class discriminatory information still maintains, while the dimension of HoG features and landmarks features is significantly reduced. To quantitatively assess the discriminatory information of the shape based features and appearance based features, we perform a statistical test (McNemar's test [125, 126]) and show that the classification performance difference between the shape-based features and appearance-based features is statistically significant (p-value < 0.001) using a 5-fold cross validation scheme. To further show that meaningful pathological features can be captured by the proposed single-frame estimator, we visualize 5 modes of variation for appearance and shape based features along the most discriminant pLDA direction for the near mouth region. Fig.3.14 is produced by computing the value of $\mu + \alpha\sigma$ and projecting it back to the original feature space, where μ is the mean features, σ is the features variation along first pLDA direction, and α is the coefficient that specifies the degree of variation. The sign of α contains the class-related information (normal, left, and right) while the amplitude of α implies the degree of deformation. Specifically, In our case, we set α equals to -3, -1.5, 0, 1.5, and 3. In Fig. 3.14, the middle column is the mean features ($\mu = 0$ and $\alpha = 0$), which corresponds to the normal subject. With the increase of α ($\alpha = 1.5$ and 3), the mean features deform into the features corresponding to the right facial weakness. On the other hand, when α decreases (e.g., equals to -1.5 and -3), the features become the pattern corresponding to the left facial weakness. Another observation is that the amplitude of α implies the degree of deformation, a larger α means a more severe facial weakness symptom. Together, Fig. 3.13 and 3.14 show that the single-frame estimator is able to classify and capture the clinically meaningful facial asymmetry using both shape and appearance-based features. By visualizing the modes of discrimination in the image dataset, we can derive that the sign of α contains the class-related information (normal, left, and right) while its amplitude indicates the degree of deformation. Therefore, given the shape and appearance-based features of a test image sample, we can project it onto the pLDA subspace and compute the value of α. Then using the corresponding value of α one can determine which class it belongs to. Another benefit of our analysis is that projecting the high-dimensional features onto a low-dimensional subspace reduces the computation complexity and increases the computation efficiency.

To further increase the interpretability and transparency of the proposed method, we examine the geometric structure of Bi-LSTM hidden states h_t^2 at time t, by projecting it onto a 2-dimensional PCA subspace [127]. To be precise, we perform PCA analysis on the Bi-LSTM hidden states $h_{t=1}^2$

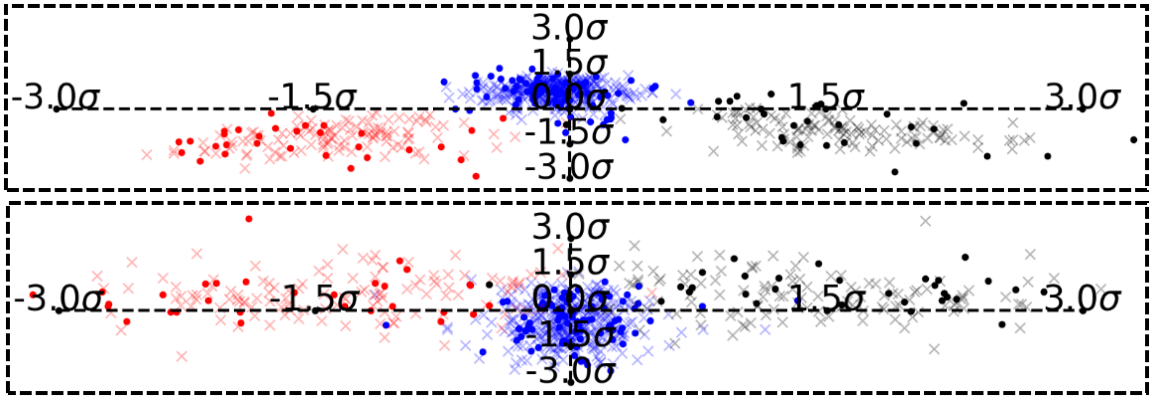

Figure 3.13: The distribution of projection of appearance-based features and shape-based features onto the pLDA subspace for image dataset: normal (blue), right (black), and left (red).

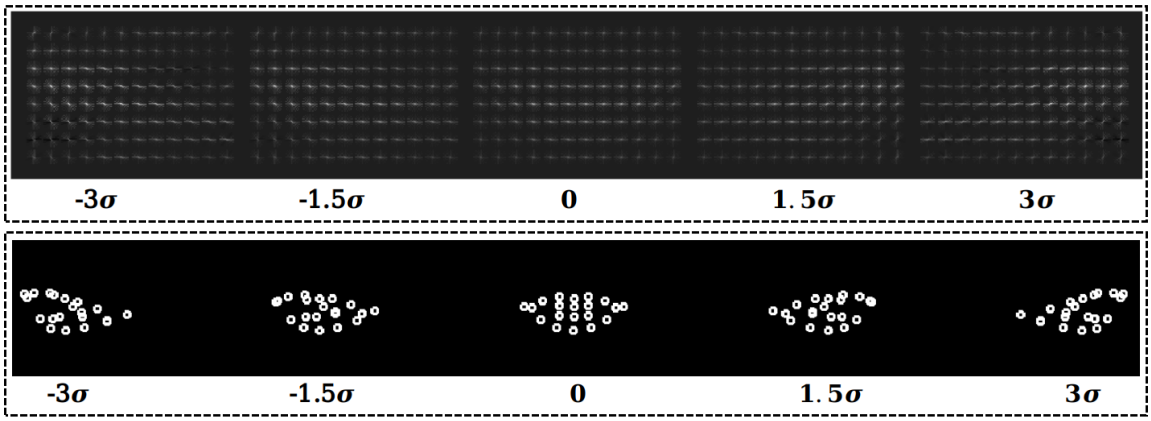

Figure 3.14: Modes of variation along first pLDA direction in HoG features space and in landmark features space.

for all the training samples. Then the projection of the hidden states $h^2_{t=1}$ for all training samples and testing samples onto the top two principle components is shown in left panel of Fig. 3.15. The right side of Fig. 3.15 shows the projection of Bi-LSTM hidden states $h^2_{t=20}$ at time T (last frame) using the same PCA setup. Overall, Fig. 3.15 illustrates that the hidden states of Bi-LSTM evolve alone in a low-dimensional subspace and become more separable when Bi-LSTM continues to take the input.

The window length T is a hyper-parameter that determines the number of consecutive frames with maximum muscle activation for the temporal modeling component of the proposed examination. We evaluate different values of T (corresponding to various time duration in seconds) and provide the averaged 5-fold test accuracy as shown in Fig. 3.16. As shown in the figure the proposed system is relatively insensitive to various values of T. Because of the heterogeneity of the dataset collected

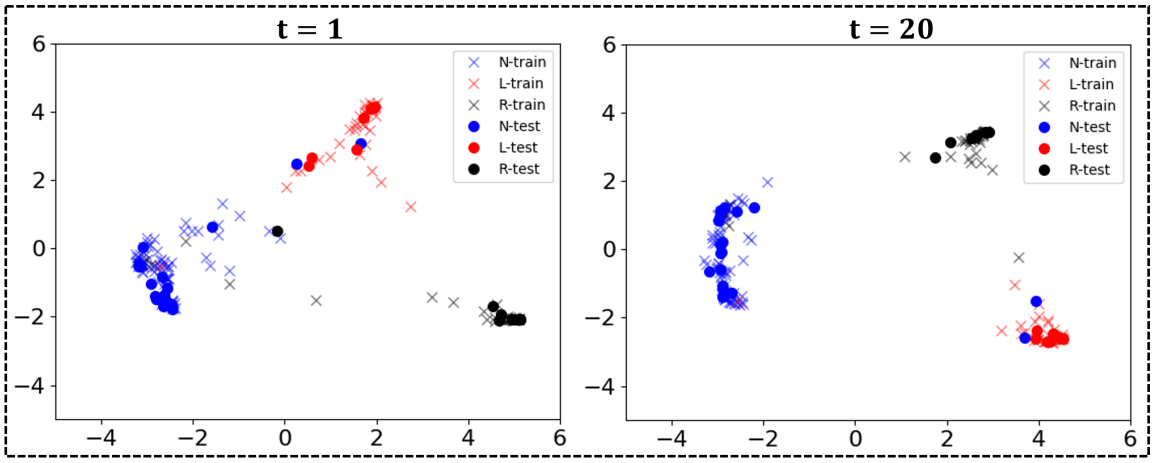

Figure 3.15: Evolution of hidden state of BiLSTM at time $t = 1$ and $t = 20$.

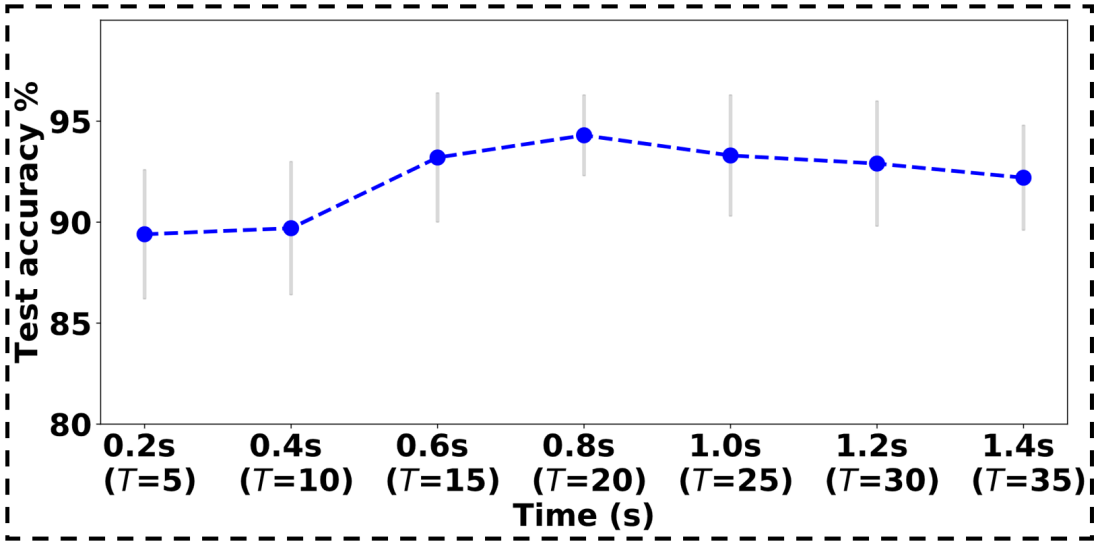

Figure 3.16: The test classification accuracy (blue marker) with standard deviation (gray line segment) for different T values.

in the wild, the duration of maximum muscle activation expressed by different subjects varies. Thus in our study, we choose T to be 0.83 seconds (20 frames). A smaller T is undesired because it would not be physiologically meaningful. In contrast, if T is too long, this increases the likelihood of added unwanted variations (e.g. oscillations due to adherence to instructions, noise) could lead to reduced predictive accuracy. Fig. 3.16 displays how the test accuracy changes with different window duration in seconds.

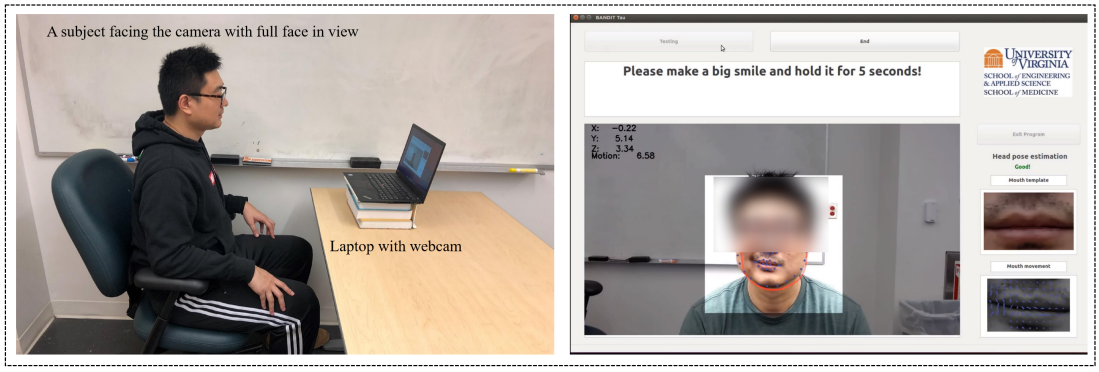

Figure 3.17: A user scenario (left), and a prototype implementation of the proposed system (right).

3.5 Prototype development

We develop a prototype that integrates the proposed algorithm and a GUI to allow users to interact with the framework in real-time. The prototype is running on a regular laptop PC with a Intel Dual-Core i7 processor, 16GB of RAM, 512GB of space, and an integrated graphics card. The webcam is a Logitech C920 HD PRO webcam [128]. The software is coded in Python 3.5 with the standard libraries such as PyQt for GUI interface design. To be specific, a typical user scenario is the one where a suspected patient follows the standard clinical examine instructions presented on the screen of a camera-enabled smart device whose camera records the examination as shown in Fig. 3.17. The instructions require the patient to maintain a smile and hold for a few seconds. Then, accurate, computationally efficient, and robust invariant feature extraction is performed followed by analytics which quantitatively assess the risk of facial weakness. The final assessment is then utilized by healthcare professionals to better support treatment decisions.

3.6 Summary

To sum up, the proposed approach is a proof-of-concept study showing that it can be beneficial to assist the paramedics to identify the facial weakness in the field or, more importantly, whenever expertise in neurology is not available either for emergency patient triage (e.g., pre-hospital stroke care) or chronic disease management (e.g., Bell's palsy rehabilitation screen), leading to increased coverage and earlier treatment for prehosptial stroke care.

Chapter 4

Optimal transport based illumination-invariant representation for face analysis

Illumination variation is an important issue for many computer vision tasks [129], including medical computer vision [13, 22]. Especially, when deploying systems in real-world clinical settings, it requires to take different illumination conditions into account, because poor illumination conditions are able to greatly decrease system performance [2]. We encounter the illumination issues in the patient dataset that are collected in outpatient clinical rooms or inpatient wards at hospitals. Specifically, poor illumination conditions such as low lighting and partial lighting degrade image quality significantly. To address these issues, this chapter presents a novel local sliced-Wasserstein feature set for illumination-invariant face analysis [75], which is beyond the scope of facial weakness detection. The proposed approach makes use of lighting insensitive measures, image gradient information, to construct a local low-level image descriptor feature set based on optimal transport metrics. To be specific, the method depends on mathematical modeling of local gradient distributions using the Radon Cumulative Distribution Transform (R-CDT) [130]. We demonstrate that lighting variations cause certain types of deformations for local image gradient distributions which, when expressed in R-CDT domain, can be modeled as a subspace. Face recognition is then performed using a nearest subspace in R-CDT domain of local gradient distributions. Experiment results demonstrate that the

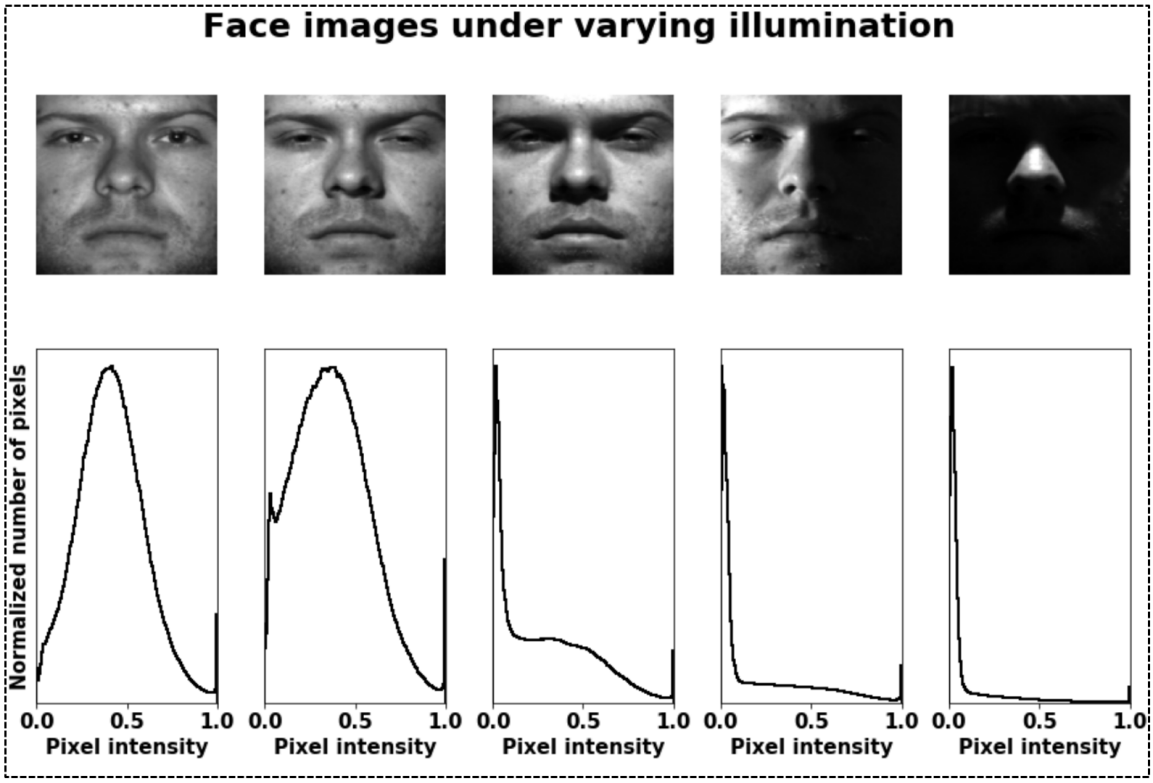

Figure 4.1: Images of a subject under varying illumination conditions and the corresponding histogram of pixel intensities.

proposed method outperforms other alternatives in several face recognition tasks with challenging illumination conditions.

4.1 Overview

Automated face recognition is a necessary task for many machine-human interaction applications. Illumination variations can cause significant appearance changes for the same person and significantly affect recognition accuracy. Several pioneer studies observed that variations among images of the same person owing to variable lighting can appear to be larger than those owing to change in identity [131, 132]. One example is illustrated by the well-known Yale B Extended Face database [129], as shown in Fig. 4.1. The top row of Fig. 4.1 shows images of the same face acquired with different lighting conditions from the face dataset [129]. The bottom row shows the corresponding histograms of the pixel intensities, which change dramatically due to varying illumination conditions. Clearly, identifying a person when illumination changes are drastic can be challenging. To resolve illumination issues for face recognition under varying lighting conditions, researchers have investigated

many approaches [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 129, 143, 144, 145]. However, the formulation of these methods are either empirical or often lacking of theoretical understanding with respect to illumination effects. In addition, these methods require many training images with different lighting conditions to be effective.

To address these issues, we propose a novel transport-based local patch-wise feature set for illumination-invariant face recognition. It leverages local patch-wise image gradient measurements to form a new image descriptive feature set based on three key ideas: (1) it is a local image gradient-based feature set; (2) we mathematically shows that lighting variations can cause deformations of 2D gradient distributions; (3) certain deformations (e.g., translation and scaling) of 2D gradient distributions can be represented as subspaces in transport (sliced-Wasserstein) R-CDT domain. To be specific, local image feature sets are widely used for face recognition. Numerous methods following this concept have been developed such as SIFT [146], HoG [147], and LBP [148]. By dividing images into small local patches and computing corresponding representations regarding local gradient information within each local image patch, local low-level gradient-based descriptive feature sets are favored, because the spatial differentiation operation naturally eliminates additive constants (i.e. brightness changes), and local patches provide more tolerance regarding slight misalignment [146, 147]. In addition, the illumination variation assumption assumes that lighting change is smooth in neighboring regions [149, 133, 134, 135, 136]. Thus local low-level feature sets are capable of providing robustness to illumination variations [150]. Secondly, we illustrate that varying illumination conditions lead to certain types of deformations of local 2D discrete distributions within an image patch. Taking this knowledge into account, we model the local image patch with different illumination conditions as a subspace in R-CDT domain [130, 151]. Using certain "convexifying" properties of the R-CDT transform, we are able to build classifiers that are invariant to certain deformations of the gradient distribution caused by changing illumination conditions. Experiment evaluations demonstrate that the proposed method achieves competitive performance among comparable approaches in three face dataset with illumination variations.

4.2 Related work

Illumination-invariant face recognition methods can be classified into three main types of approaches: (1) illumination invariant feature extraction; (2) 3D face modeling; (3) deep learning approaches with data augmentation.

Notations	
Ω, Ω_k	image coordinate domain, coordinate domain for patch k
Ω_k^N	a set $\{\mathbf{x}_1^k, ..., \mathbf{x}_N^k\}$ of N pixel locations on Ω_k
I_k	$I\|_{\Omega_k}$:k_{th} patch of image I
C, L, K	total number of subjects, image samples, and patches
δ_z	Dirac measure centered at z
$P_{\nabla I_k}$	2D discrete gradient distribution of I_k
$P_{\nabla_\theta I_k}$	1D discrete distribution of directional derivative of I_k alone angle θ
$\mathbb{H}$	a subset of bijection deformations from $\mathbb{R}^2$ to $\mathbb{R}^2$
$\mathbb{H}_0$	set of all possible compositions of translation and scaling diffeomorphisms
$\mathbf{w}_\theta$	directional vector $[\cos\theta, \sin\theta]^T$
$P^h_{\nabla I_k}$	2D discrete gradient distribution deformed (push-forwarded) from $P_{\nabla I_k}$ via $h \in \mathbb{H}$
$\mathcal{F}$	Discrete CDT Transform for 1D discrete distributions
$\mathcal{F}^*$	Discrete R-CDT transform for 2D discrete distributions
$\widehat{P}_{\nabla I_k}$	$\widehat{P}_{\nabla I_k} := \mathcal{F}^*\big(P_{\nabla I_k}\big)$
$\mathbb{S}^c$	a generative model for images of subject c under illumination variations
$\mathbb{V}_k^c$	the subspace in the Discrete R-CDT domain corresponding to k_{th} patch of subject c
$d(\cdot,\cdot)$	the discrete sliced Wasserstein distance between two 2D distributions

Most lie on the spectrum of illumination invariant feature extraction. Illumination-invariant feature extraction approaches are capable of eliminating lighting variations by utilizing holistic decomposition [134], quotient models [135, 136], or logarithm difference [133]. These approaches perform illumination normalization on images in such a way that they have robust appearances under varying illumination. One straightforward approach is to employ the log transform for illumination normalization [152], followed by lighting invariant feature extraction [134, 142], because log transform is able to normalize the contrast, by mapping the narrow range of low intensity values in the input into a wide range of output levels and expanding the value of dark pixels [152]. More specifically, Zhu *et. al.* first apply the log transform and then compute HoG features for face recognition [142]. Similarly, Chen *et. al.* use the log transform and then decompose the image into high-frequency and low-frequency components keeping only high-frequency components for face recognition [134]. Lai *et. al.* calculate the difference between neighboring pixel values in log transformation domain, formulating the logarithm-difference edge map. Then based on the size of the neighborhood, different scales of edge maps are computed and aggregated to represent each face [133]. Other approaches such as WebberFace [136] and GradientFace [135] are quotient-based models. To be precise, quotient models seek to represent images in such way that the current pixel value in the new representation is the ratio between the difference of current pixel and its neighboring pixel to the current pixel value in original image. However, although these methods achieve excellent results on some datasets, they are lack of mathematical understanding regarding illumination variations and often not effective for illumination variations that include shadows.

Another research direction is to use a set of images acquired under varying lighting conditions to build a 3D face model that can render all possible illumination variations. Several researchers

studied the properties of learned subspace models such as convexity and dimension [153, 129, 154]. In addition, deep learning-based methods such as VGGface [137] and others [155, 138, 139, 140] have become increasingly popular, because of the increased classification results in standard face recognition tasks. However, this type of methods typically requires a large amount of data for training, which need to be under various illumination conditions. In addition, to overcome the issue of limited samples, many data augmentation strategies are employed. However, even with data augmentation techniques [156], deep learning-based methods tend to have relatively poor generalization properties [157].

4.3 The proposed approach

Summary We propose a novel method for illumination invariant face representation that are based on the following ideas: (1) first, we patchify images into multiple image patches and calculate the 2D discrete gradient distributions for each individual image patch; (2) secondly, we mathematically show that illumination variations cause transport-type perturbation for local 2D discrete gradient distributions; (3) thirdly, we define a sliced-Wasserstein representation for local 2D discrete gradient distributions, which has several properties to address the transport-like variations and convexity the classification problem in R-CDT domain; (4) built upon this new representation, we construct a nearest subspace model to perform face recognition under varying illuminations. The rationales behind each step is detailed in the subsequent sections.

4.3.1 Effects of varying illumination conditions on local 2D gradient distributions

First, we introduce several notations, and then mathematically demonstrate that illumination variations cause deformations of the local 2D discrete gradient distributions in a small region.

Notation We first provide the definition of an image I, its k_{th} patch, and 2D discrete gradient distribution $P_{\nabla I_k}$. Namely, an image $I : \Omega \to \mathbb{R}_+$ can be thought as a mapping from the unit square $\Omega = [0, 1] \times [0, 1]$ to the set $\mathbb{R}_+$ of non-negative real numbers. Let $\Omega_k \subset \Omega$ refers to a set of pixel coordinates in the k^{th} neighborhood in Ω and $\Omega_k^N = \{\mathbf{x}_1^k, ..., \mathbf{x}_N^k\} \subset \Omega$ is the set of N pixel locations for patch k.

Given an image patch $I_k : \Omega_k \to \mathbb{R}_+$, the corresponding 2D discrete gradient distribution $P_{\nabla I_k}$ for the k_{th} patch is defined as

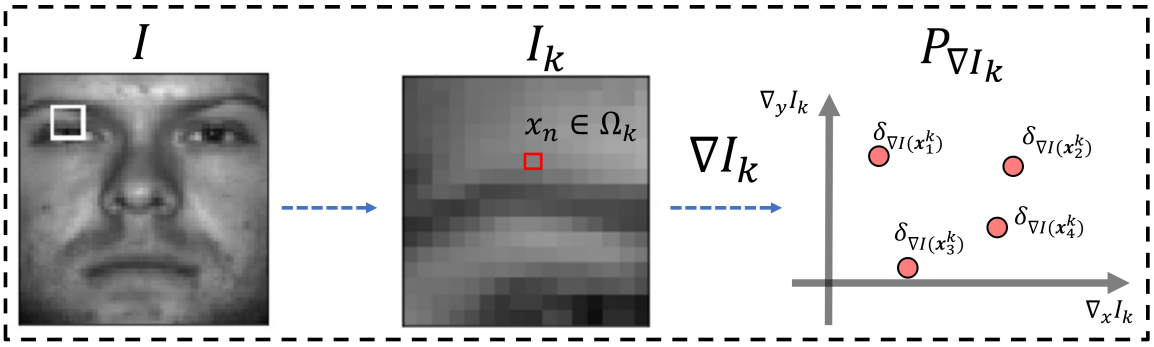

Figure 4.2: From image space to local 2d discrete gradient distribution space.

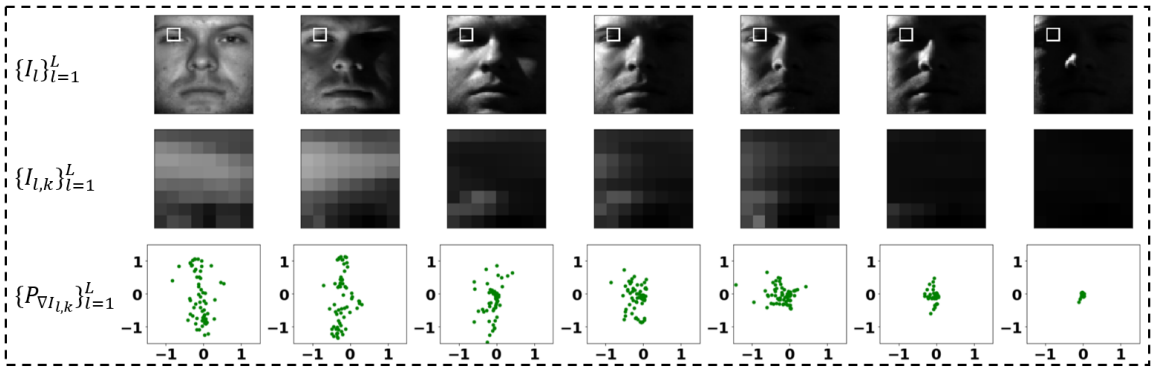

Figure 4.3: Top row shows face images under various real-world illumination conditions, middle row shows the corresponding patches of face images, bottom row shows the local 2D discrete gradient distributions.

$$P_{\nabla I_k} := \frac{1}{N} \sum_{\mathbf{x} \in \Omega_k^N} \delta_{\nabla I(\mathbf{x})}. \tag{4.1}$$

An example image I, image patch I_k, and corresponding distribution $P_{\nabla I_k}$ are shown in Fig. 4.2. Note that our use of measure theoretic notation δ_z for a delta mass positioned at z. A Dirac measure $\delta_{\mathbf{z}}$ on $\mathbb{R}^n$ is a measure with mass concentrated at $\mathbf{z} \in \mathbb{R}^n$ $(n \geq 1)$ such that for any (measurable) set $A \subseteq \mathbb{R}^n$:

$$\delta_{\mathbf{z}}(A) = \begin{cases} 1, & \mathbf{z} \in A \\ 0, & \mathbf{z} \notin A. \end{cases} \tag{4.2}$$

In other words, a set has measure 1 if it contains the point z and measure zero otherwise. In particular, $\delta_{\mathbf{z}}(\{\mathbf{z}\})$ =1 and $\delta_{\mathbf{z}}(\{\mathbf{z}'\})$ =0 for any $\mathbf{z}' \neq \mathbf{z}$.

Since it is interesting to model how $P_{\nabla I_k}$ deforms under different illumination variations, we define a bijective transformation h that is able to transform $P_{\nabla I_k}$. One can think that the transformation

function h is corresponding to one certain lighting condition. More formally, given the local gradient distribution $P_{\nabla I_k}$ of a subject observed under a certain illumination condition, the corresponding patch gradient distribution of the same subject after some illumination changes can be modeled by a bijective transformation h as:

$$P^h_{\nabla I_k} := h_\sharp P_{\nabla I_k} = \frac{1}{N} \sum_{\mathbf{x} \in \Omega_k^N} \delta_{h(\nabla I(\mathbf{x}))}, \quad h \in \mathbb{H} \tag{4.3}$$

where $\mathbb{H}$ is a set of bijective deformations from $\mathbb{R}^2$ to $\mathbb{R}^2$ that are particular to certain illumination assumptions. We denote the push forward of distribution $P_{\nabla I_k}$ as $h_\sharp P_{\nabla I_k}$.

Effects of illumination changes on 2D gradient distributions An example of one image patch observed under different illumination conditions is shown in Fig. 4.3. The example shows that the gradient distribution $P_{\nabla I_k}$ undergoes deformations that may include translation, scaling, skewing, rotation, and nonrigid deformations according to the reflectance properties of the object being imaged, its three dimensional configuration, and specific illumination conditions. This means that in the local image patch k, one can model the 2D gradient discrete distribution as an instance of a template 2D discrete gradient distribution observed under some deformations or confound h. In other words, the 2D discrete distribution for the set of image patch of an person could be a fixed pattern, but observed under affine deformation effects in the 2D gradient space, such as translations and scaling effects which lead brightness and contrast variations in image space. Other unknown deformations could be present as well, as shown in Fig. 4.3.

Here we propose a transport-based way of modeling such deformations. The set of gradient distributions of a subject image patch I_k under various illumination variations with the following template-deformation based generative model:

$$\mathbb{P}_{\mathbb{H},k} = \{P^h_{\nabla I_k} \mid h \in \mathbb{H}\}. \tag{4.4}$$

A particular model for set $\mathbb{H}$ that describes illumination effects is proposed below. Here $\mathbb{P}_{\mathbb{H},k}$ refers to the set of all possible observations of $P^h_{\nabla I_k}$. We will utilize geometric properties of $\mathbb{P}_{\mathbb{H},k}$ in the problem statement and solution described below.

Modeling $\mathbb{H}$: local illumination changes and patch generative model The set $\mathbb{H}$ of deformations corresponding to the set of illumination variations within a subject class can be hard to

specify in general. However, we can make some reasonable assumptions and propose an approximate model that can enable robust classification. First, using the smoothness assumptions that are widely used for illumination-invariant face recognition [149, 133, 134, 135, 136], for small enough neighborhood size $|\Omega_k|$ (i.e., the area of neighborhood Ω_k), we postulate that,

$$h(\mathbf{z}) \sim \alpha\mathbf{z} + \mathbf{b} \tag{4.5}$$

where $\mathbf{z}$ is a gradient coordinate (i.e. $\mathbf{z} = \nabla I_k(x), x \in \Omega_k$), and where $\alpha \in \mathbb{R}$ is an unknown scaling function, and $\mathbf{b} \in \mathbb{R}^2$ is an unknown translation vector. The illumination model above can be derived from the assumption that within a local patch, illumination variations can be expressed as:

$$\alpha I_k(\mathbf{x}) + \beta + \mathbf{b}^T\mathbf{x}, \mathbf{x} \in \Omega_k. \tag{4.6}$$

Under the assumption that $|\Omega_k|$ is small, in the equation above, $\alpha > 0$ is known as contrast, β illumination intensity, and $\mathbf{b}$ is a linear gradient (caused by illumination at an angle, or potentially shadows) superimposed on the image. Under the assumption of small neighborhood $|\Omega_k|$, the model we propose in equation (4.5) can be understood as the gradient of (4.6), and thus can be understood in terms of illumination intensity, contrast, and linear illumination gradient.

In other words, varying parameters α and $\mathbf{b}$ of the illumination model, defined in equation (4.6), can lead to changes of contrast, brightness, and shadow in image space. Here Fig.4.4 and Fig.4.5 provide visualizations of such variations. Top and middle panels show simulated illumination conditions in the image space using equation (4.6) where $\beta = 0$ and $\mathbf{b} = \mathbf{0}$. The bottom panel demonstrates the corresponding scaling effects caused by different α values in 2D discrete distribution space. Note that equation (4.6) is applied to the entire image domain for the illustration purposes. Fig.4.4 shows that changing α not only contributes to contrast change (top two rows) in image space but also causes scaling effects for the corresponding local 2D discrete distributions (bottom row). Likewise in Fig.4.5, the original face image is shown on the left, we add the linear illumination gradient, specified by a constant vector $\mathbf{b}$, to simulate the lighting coming from the side, as shown one in right part of Fig.4.5. Simulated illumination conditions in the image space using equation (4.6) where $\alpha = 1$, $\beta = 0$, and $\mathbf{b}$ is a constant vector. Adding a linear gradient is able to simulate lighting coming from the side, therefore, resulting in a translation effect in 2D discrete distribution space. Correspondingly, the local 2D discrete distribution of the simulated image experiences a translation effect.

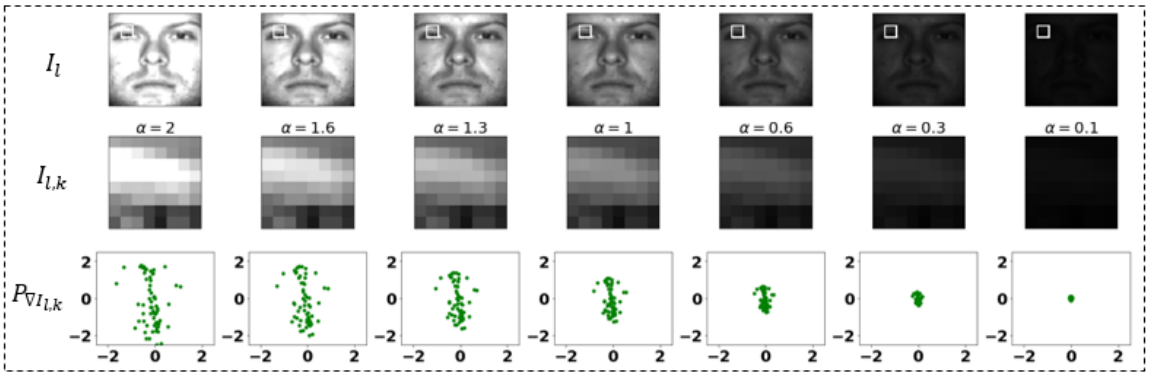

Figure 4.4: Simulated contrast effects.

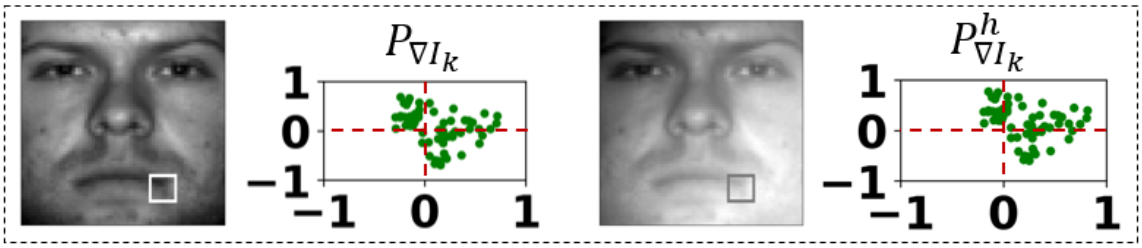

Figure 4.5: Simulated partial lighting effects.

In addition, in equation (4.6), β is responsible for controlling image brightness. However, it is eliminated automatically in derivative operation when calculating image gradients. Please note that there are several existing works that aimed to address the contrast variations (α in equation (4.6)) by normalization. For instance, HoG features perform block-wise L_2-norm normalization within a local region [141]. Furthermore, applying the log transform is also able to normalize the contrast, by expanding the value of dark pixels [152]. Though these approaches improve performance to certain degree but are ineffective for challenging illumination conditions as demonstrated in experimental results.

Leveraging the knowledge present in equation (4.6) and equation (4.5), we propose the following patch-wise affine generative model and classification problem using gradient distributions for face images under varying illumination conditions. Specifically, based on (4.5) above, we propose a specific transport-based model to approximate the set of bijections $\mathbb{H}$ that cause gradient deformations (pushforward) of a given gradient distribution:

$$\mathbb{H}_0 = \{h(\mathbf{z}) = \alpha\mathbf{z} + \mathbf{b} \mid a > 0, \mathbf{b} \in \mathbb{R}^2\}. \tag{4.7}$$

Finally, using the notation established earlier, we express the set of gradient distributions observed under an unknown illumination function $h \in \mathbb{H}_0$ as

$$P^h_{\nabla I_k} := \frac{1}{N} \sum_{\mathbf{x} \in \Omega^N_k} \delta_{h(\nabla I(\mathbf{x}))} = \frac{1}{N} \sum_{\mathbf{x} \in \Omega^N_k} \delta_{\alpha \nabla I(\mathbf{x}) + \mathbf{b}}. \tag{4.8}$$

and the set of all possible observations as

$$\mathbb{P}_{\mathbb{H}_0,k} = \{P^h_{\nabla I_k} = h_\sharp P_{\nabla I_k} \mid h \in \mathbb{H}_0\}. \tag{4.9}$$

We denote the set defined in equation (4.9) as a model for the gradient distribution of patch k under illumination model $\mathbb{H}_0$. Specifically $\mathbb{H}_0$ is capable of isotropically scaling and translating a given distribution $P_{\nabla I_k}$.

To here, equipped with the relationship between illumination variations and 2D gradient distributions as well as the proposed local deformation set H_0, we present the problem statement using the local illumination-based generative model for local pixel intensities and local 2D gradient distributions as below:

Problem statement: Let the local illumination-based generative model for images pertaining to subject (class) $c = 1, \cdots, C$ be defined as:

$$\mathbb{S}^c = \left\{ I^{c,j} \;\middle|\; \begin{array}{l} I^{c,j}_k(\mathbf{x}) = \alpha^j_k I^c_k(\mathbf{x}) + \beta^j_k + \mathbf{b}^j_k \cdot \mathbf{x}, \\ \mathbf{x} \in \Omega_k, \alpha^j_k > 0, \beta^j_k \in \mathbb{R}, \mathbf{b}^j_k \in \mathbb{R}^2, \\ k = 1, ..., K \end{array} \right\}, \tag{4.10}$$

where I^c refers to an (unknown) template image for subject (class) c. This model defines an infinite set, whose elements $I^{c,j}$ can be generated by applying the illumination model (4.6) on each patch k of I^c independently. In other words, the generative model is flexible to allow each patch k to contain its own contrast, brightness, and gradient vector $(\alpha^j_k, \beta^j_k, \mathbf{b}^j_k)$ parameters. These illumination parameters are unknown for any given photograph. Given L training images $\{I^{c,l}\}^L_{l=1} \subseteq \mathbb{S}^c$ from each class $c = 1, ..., C$ where $\mathbb{S}^c \cap \mathbb{S}^{c'} = \emptyset$ for all $c \neq c'$, the goal is to determine the class of an unknown image I^t obtained from the same generative model.

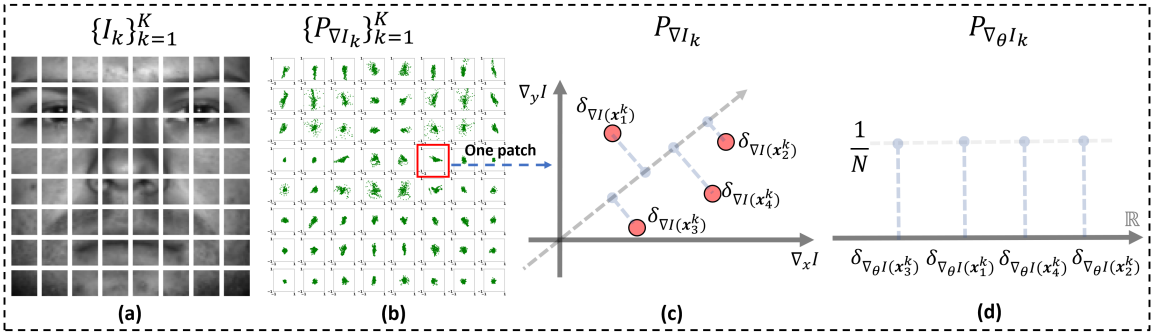

Figure 4.6: An illustration example of computing a sliced one-dimensional discrete distribution for one image patch.

It is not hard to see that $\nabla I_k^{c,j} = h_k^j \circ \nabla I_k^c$ where $h_k^j(\mathbf{z}) = \alpha_k^j \mathbf{z} + \mathbf{b}_k^j$. In other words, the patch gradient distributions for each class c satisfy the affine generative model stated in (4.9), i.e.,

$$\{P_{\nabla I_k^{c,j}} \mid I^{c,j} \in \mathbb{S}^c\} = \mathbb{P}^c_{\mathbb{H}_0,k} := \{P^h_{\nabla I_k^c} \mid h \in \mathbb{H}_0\}. \tag{4.11}$$

Proposed solution To address the problem, we propose a straightforward, non iterative, solution to the classification problem stated above. The solution is inspired on prior work on classification of distributions [158, 130, 151] and utilizes the fact that the gradient distribution $P_{\nabla I_k^t}$ for patch k is an element of $\mathbb{P}^c_{\mathbb{H}_0,k}$, with c unknown. In other words $P_{\nabla I_k^t} \in \mathbb{P}^c_{\mathbb{H}_0,k}$, with $\mathbb{P}^c_{\mathbb{H}_0,k} = \{h_\sharp P_{\nabla I_k^c} \mid h \in \mathbb{H}_0\}$, following the definition in equation (4.9), for some unknown c, we use a distance function $d(P_{\nabla I_k^t}, \mathbb{P}^c_{\mathbb{H}_0,k})$ that measures the sliced-Wasserstein distance [159, 160] between $P_{\nabla I_k^t}$ and the nearest point in set $\mathbb{P}^c_{\mathbb{H}_0,k}$ to compute the solution of the classification problem stated above as:

$$c^* = \arg\min_c \sum_{k=1}^K d^2(P_{\nabla I_k^t}, \mathbb{P}^c_{\mathbb{H}_0,k}). \tag{4.12}$$

It is easy to show that the minimization above obtains the correct solution to the problem statement above, provided that for at least least one k we have that $\mathbb{P}^c_{\mathbb{H}_0,k} \cap \mathbb{P}^{c'}_{\mathbb{H}_0,k} = \emptyset$ whenever $c \neq c'$. Below we show how we can estimate $d^2(P_{\nabla I_k^t}, \mathbb{P}^c_{\mathbb{H}_0,k})$ with the aid of a newly introduced operation which we will call the Discrete Radon Cumulative Distribution transform (Discrete R-CDT).

4.3.2 Sliced-Wasserstein representation of local 2D discrete distribution

Discrete CDT Transform

Definition Given an image I where the patch-wise representation is $\{I_k\}_{k=1}^K$, we compute its local gradient distribution $P_{\nabla I_k}$ as defined in equation (4.1) and use a modified version of the Radon Cumulative Distribution Transform (R-CDT) to represent $P_{\nabla I_k}$. In other words, a set of one-dimensional discrete distributions is used to represent $P_{\nabla I_k}$ by the slicing operation, which projects each $\nabla I(\mathbf{x})$ onto a weight vector $\mathbf{w}_\theta = (\cos\theta, \sin\theta)^T$ along a slicing angle $\theta \in [0, \pi)$ as $\nabla_\theta I(x) = \nabla I(\mathbf{x}) \cdot \theta$. Thus for a certain θ, one can obtain an one-dimensional discrete distribution as $P_{\nabla_\theta I_k}$. Fig. 4.6 illustrates such a concept.

Mathematically, the 1D distribution $P_{\nabla_\theta I_k^c}$ of projected gradients can be computed via

$$P_{\nabla_\theta I_k} = \frac{1}{N} \sum_{\mathbf{x} \in \Omega_k^N} \delta_{\nabla I(\mathbf{x}) \cdot \mathbf{w}_\theta}, \tag{4.13}$$

where $\mathbf{w}_\theta = (\cos\theta, \sin\theta)^T$ is a unit vector in the direction of θ.

Inspired on earlier work on the CDT [158] and R-CDT [130], we define a new transformation, the Discrete CDT Transform, for one-dimensional discrete probability distributions.

Definition 3.1 (Discrete Cumulative Distribution Transform): Let $\mathbb{P}_N(\mathbb{R}) := \{P_Z = \frac{1}{N}\sum_{i=1}^N \delta_{z_i} \mid Z = \{z_i\}_{i=1}^N \subset \mathbb{R}\}$ be the set of discrete probability distributions concentrated on N points on $\mathbb{R}$. The Discrete R-CDT transform $\mathcal{F} : \mathbb{P}_N(\mathbb{R}) \to \mathbb{R}^N$ is defined

$$\mathcal{F}(P_Z) = \mathcal{P}\big[z_1, ..., z_N\big]^T = \big[\widetilde{z}_1, ..., \widetilde{z}_N\big]^T, \tag{4.14}$$

where $\mathcal{P}$ is a permutation matrix such that $\widetilde{z}_1 \leq \cdots \leq \widetilde{z}_N$. In other words, the Discrete CDT Transform $\mathcal{F}$ takes the one-dimensional discrete distribution $P_{\nabla_\theta I_k}$ as input and outputs a vector which are concentration locations of the discrete distribution in an increasing order.

One interesting propriety of the transform $\mathcal{F}$ in eq. (4.14) is the composition property, which helps to address the translation and scaling effects, thus rendering classes convex and simplifying the classification task. Specifically, the general composition property of the transform $\mathcal{F}$ is defined as:

Composition property Let $T : \mathbb{R} \to \mathbb{R}$ be a strictly increasing function and $P_Z \in \mathbb{P}_N(\mathbb{R})$, then

$$\mathcal{F}\big(P_Z^T\big) = T \circ \mathcal{F}\big(P_Z\big), \tag{4.15}$$

where the composition is operated entry-wise and $P_Z^T = \frac{1}{N}\sum_{i=1}^N \delta_{T(z_i))}$ is the push-forward [1] distribution of P_Z by T. The property implies that for any strictly increasing T, applying the transform $\mathcal{F}$ to the push-forward distribution P_Z^T equates to compose T with the transform $\mathcal{F}(P_Z)$.

Proof: Let $P_Z = \frac{1}{N}\sum_{i=1}^N \delta_{z_i}$ with $\mathcal{F}(P_Z) = [\widetilde{z}_1, \widetilde{z}_2, \cdots \widetilde{z}_N]^T$ where $\widetilde{z}_1 \le \widetilde{z}_2 \le \cdots \widetilde{z}_N$ is reordered from $[z_1, ..., z_N]$. Observing that $P_Z^T = \frac{1}{N}\sum_{i=1}^N \delta_{T(z_i)}$ and by the definition of Discrete CDT Transform, $\mathcal{F}(P_Z^T)$ is the vector of which the entries are $T(z_i)$'s in an increasing order, i.e,

$$\mathcal{F}(P_Z^T) = [\widetilde{T(z_1)}, \widetilde{T(z_2)}, \cdots \widetilde{T(z_N)}]^T \tag{4.16}$$

where $\widetilde{T(z_1)} \le \widetilde{T(z_2)} \le \cdots \le \widetilde{T(z_N)}$. On the other hand, since T is strictly increasing, we have that $T(\widetilde{z}_1) \le T(\widetilde{z}_2) \le \cdots \le T(\widetilde{z}_N)$, which is also an reordering of $T(z_1), ..., T(z_N)$. Hence we have that

$$\mathcal{F}(P_Z^T) = \begin{bmatrix} T(\widetilde{z}_1) \\ T(\widetilde{z}_2) \\ \vdots \\ T(\widetilde{z}_N) \end{bmatrix} = T \circ \mathcal{F}(P_Z). \tag{4.17}$$

In addition, there are several other properties for Discrete CDT that one can exploit and leverage to render non-linear problems linearly separable in the transformed space. To be precise, we note the following interesting cases where T is a translation or a scaling diffeomorphism.

Translation property Let $T : \mathbb{R} \to \mathbb{R}$ be the translation function where $T(x) = x + a$ for some $a \in \mathbb{R}$. By the composition property, we have that

$$\mathcal{F}(P_Z^T) = \mathcal{F}(P_Z) + a, \tag{4.18}$$

where the addition on the RHS is operated entry-wise.

[1] In general the push-forward measure $T_\sharp\mu$ of a measure μ under $T : X \to Y$ is defined by the property that $T_\sharp\mu(U) = \mu(T^{-1}(U))$ for any measurable set $U \subseteq Y$. In particular, for any one-dimensional discrete distribution P_Z, given any Lebesgue measurable function $T : \mathbb{R} \to \mathbb{R}$, it can be shown that $T_\sharp(P_Z) = \frac{1}{N}\sum_{i=1}^N \delta_{T(z_i)}$.

Scaling property Let $T : \mathbb{R} \to \mathbb{R}$ be a scaling where $T(x) = cx$ for some $c > 0$. By the composition property, we have that

$$\mathcal{F}(P_Z^T) = c\mathcal{F}(P_Z). \tag{4.19}$$

Convexity property Using the composition property, we can derive the following convexity property of the transform $\mathcal{F}$. Let $\mathbb{G}_1 \subseteq \mathbb{H}_1$ where $\mathbb{H}_1 = \{T : \mathbb{R} \to \mathbb{R} : T \text{ is a strictly increasing diffeomorphism}\}$. Then given a one dimensional discrete distribution P_Z for some $Z = \{z_i\}_{i=1}^N$, the set of transforms $\mathcal{F}\big(\mathbb{P}_Z^{\mathbb{G}_1}\big) := \{\mathcal{F}(P_Z^T) \mid T \in \mathbb{G}_1\}$ is convex if $\mathbb{G}_1$ is convex.

Proof: Let $T_1, T_2 \in \mathbb{G}$ and $\lambda \in [0, 1]$. Then by the definition of the Discrete CDT Transform and using the composition property of Discrete CDT Transform, we have that

$$\lambda\mathcal{F}\Big(P_Z^{T_1}\Big) + (1-\lambda)\mathcal{F}\Big(P_Z^{T_2}\Big) \tag{4.20}$$

$$= \lambda\Big(T_1 \circ \mathcal{F}\big(P_Z\big)\Big) + (1-\lambda)\Big(T_2 \circ \mathcal{F}\big(P_Z\big)\Big) \tag{4.21}$$

$$= \big(\lambda T_1 + (1-\lambda)T_2\big) \circ \mathcal{F}\big(P_Z\big) \tag{4.22}$$

$$= \mathcal{F}\Big(P_Z^{\lambda T_1 + (1-\lambda)T_2}\Big) \in \mathcal{F}\big(\mathbb{P}_Z^{\mathbb{G}_1}\big) \tag{4.23}$$

where the inclusion in (4.23) is due to the fact that $\mathbb{G}_1$ is convex and in particular, $\lambda T_1 + (1-\lambda)T_2) \in \mathbb{G}_1$. Hence $\mathcal{F}\big(\mathbb{P}_Z^{\mathbb{G}_1}\big)$ is convex. This convex propriety implies that a convex combination of translated and scaled one-dimensional discrete distributions in the transformed $\mathcal{F}$ domain is still convex, thus rendering classes convex and simplifying the classification task.

Connection to the Wasserstein distance It is easy to show that $\mathcal{F}$ is an isometric embedding from the 1D discrete probability distribution space with the Wasserstein metric to the transform space with the Euclidean distance [159, 161]. To summarize, for two discrete distributions $P_{Z^{(1)}}$ and $P_{Z^{(2)}}$ in $\mathbb{P}(\mathbb{R})$, the Wasserstein distance between $P_{Z^{(1)}}$ and $P_{Z^{(2)}}$ is computed via

$$W_2(P_{Z^{(1)}}, P_{Z^{(2)}}) = \sqrt{\frac{1}{N}||\mathcal{F}(P_{Z^{(1)}}) - \mathcal{F}(P_{Z^{(2)}})||^2} \tag{4.24}$$

where $||\cdot||$ denotes the Euclidean distance on $\mathbb{R}^N$. In particular, for a directional derivative distribution $P_{\nabla_\theta I_k} \in \mathbb{P}_N(\mathbb{R})$, it has

$$\mathcal{F}(P_{\nabla_\theta I_k}) := \mathcal{P}_\theta \begin{bmatrix} \nabla_\theta I(\mathbf{x}_1^k) \\ \nabla_\theta I(\mathbf{x}_2^k) \\ \vdots \\ \nabla_\theta I(\mathbf{x}_N^k) \end{bmatrix} = \begin{bmatrix} \widetilde{\nabla_\theta I(\mathbf{x}_1^k)} \\ \widetilde{\nabla_\theta I(\mathbf{x}_2^k)} \\ \vdots \\ \widetilde{\nabla_\theta I(\mathbf{x}_N^k)} \end{bmatrix}, \tag{4.25}$$

where $\nabla_\theta I(\mathbf{x}_i^k) = \nabla I(\mathbf{x}_i^k) \cdot \mathbf{w}_\theta$ and $\mathcal{P}_\theta$ is a permutation matrix such that $\widetilde{\nabla_\theta I(\mathbf{x}_1{}^k)} \leq \widetilde{\nabla_\theta I(\mathbf{x}_2{}^k)} \leq \cdots \leq \widetilde{\nabla_\theta I(\mathbf{x}_N{}^k)}$.

Furthermore, in a more general setting, for any two 1D discrete distributions $P_{Z^{(1)}}$ and $P_{Z^{(2)}}$ in $\mathbb{P}(\mathbb{R})$,

$$W_2(P_{Z^{(1)}}, P_{Z^{(2)}}) = \sqrt{\frac{1}{N}||\mathcal{F}(P_{Z^{(1)}}) - \mathcal{F}(P_{Z^{(1)}})||^2}, \tag{4.26}$$

where $||\cdot||$ denotes the Euclidean distance on $\mathbb{R}^N$. Indeed, for two discrete measures $P_{Z^{(1)}} = \frac{1}{N}\sum_{i=1}^N \delta_{z_i^{(1)}}$ and $P_{Z^{(2)}} = \frac{1}{N}\sum_{i=1}^N \delta_{z_i^{(2)}}$, the 2-Wasserstein distance between the two measures is the same as the Euclidean distance between the mass location vectors sorted in an increasing order, i.e., $W_2(P_{Z^{(1)}}, P_{Z^{(2)}}) = \sqrt{\frac{1}{N}\sum_{i=1}^N (\widetilde{z}_i^{(1)} - \widetilde{z}_i^{(2)})^2}$, where $\widetilde{z}_1^1 \leq \cdots \leq \widetilde{z}_N^1$ and $\widetilde{z}_1^2 \leq \cdots \leq \widetilde{z}_N^2$ are sorted versions of $z_1^{(1)}, ..., z_N^{(1)}$ and $z_1^{(2)}, ..., z_N^{(2)}$ respectively. This can be seen as a special case of Proposition 2.17 in [161].

Discrete R-CDT transform

To here, we illustrate the optimal transport representation for one-dimensional discrete distribution P_Z and its several interesting properties, which could be extended to 2D discrete distributions. In this section, we will present the Discrete R-CDT transform for 2D discrete gradient distributions. To be precise, leveraging the "slicing" idea shown in (4.13) and the Discrete CDT Transform defined in (4.14), the Discrete R-CDT transform for 2D discrete distributions is defined as:

Definition 3.2 (Discrete R-CDT transform): Let $\mathbb{P}_N(\mathbb{R}^2) := \{P_\mathbf{Z} = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{z}_i} \mid \mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^N \subset \mathbb{R}^2\}$ be the set of discrete probability distributions concentrated on N points on $\mathbb{R}^2$. The Discrete R-CDT transform $\mathcal{F}^* : \mathbb{P}_N(\mathbb{R}^2) \to \left(\mathbb{R}^N\right)^{[0,\pi)} := \{v : [0,\pi) \to \mathbb{R}^N\}$, denoted as $\widehat{P}_\mathbf{Z} := \mathcal{F}^*(P_\mathbf{Z})$, is defined such that for each $\theta \in [0,\pi)$:

$$\Big(\mathcal{F}^*(P_\mathbf{Z})\Big)(\theta) = \mathcal{F}\Big(P_{\mathbf{Z}_\theta}\Big), \tag{4.27}$$

where $P_{\mathbf{Z}_\theta} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\mathbf{z}_i \cdot \mathbf{w}_\theta}$ is a one-dimensional distribution concentrated on the projected values of Z onto the directional vector $\mathbf{w}_\theta$. This definition is an extension for 2D discrete distribution, meaning that applying $\mathcal{F}^*$ on the $P_{\mathbf{Z}}$ equates to applying $\mathcal{F}$ on a sequence of projected one-dimensional representations $P_{\mathbf{Z}_\theta}$[2] of $P_{\mathbf{Z}}$ by weight vector $\mathbf{w}_\theta = (\cos\theta, \sin\theta)^T$, indexed by an angle $\theta \in [0, \pi)$.

Convexity property of Discrete R-CDT transform The convex property can also be extended to $\mathcal{F}^*(\cdot)$ in 2D case, if the deformation function h is an element of $\mathbb{P}_{\mathbb{H}_0,k}$. Namely, we will show that the set $\mathcal{F}^*\big(\mathbb{P}_{\mathbb{H}_0,k}\big)$ is convex.

Proof: Let $P^{h_1}_{\nabla I_k^c}, P^{h_2}_{\nabla I_k^c} \in \mathbb{P}_{\mathbb{H}_0,k}$ where $h_1(\mathbf{z}) = a_1\mathbf{z} + \mathbf{b}_1, h_2(\mathbf{z}) = a_2\mathbf{z} + \mathbf{b}_2 \in \mathbb{H}_0$. By definition, for any $\theta \in [0, \pi)$

$$\begin{aligned}
\mathcal{F}^*\big(P^{h_1}_{\nabla I_k^c}\big)(\theta) &= a_1\mathcal{F}^*\big(P(\nabla I_k^c)\big) + \mathbf{b}_1 \cdot \mathbf{w}_\theta \\
\mathcal{F}^*\big(P^{h_2}_{\nabla I_k^c}\big)(\theta) &= a_2\mathcal{F}^*\big(P(\nabla I_k^c)\big) + \mathbf{b}_2 \cdot \mathbf{w}_\theta
\end{aligned}$$

Given $\lambda \in [0, 1]$, we have that

$$\begin{aligned}
&\left(\lambda\mathcal{F}^*\big(P^{h_1}_{\nabla I_k}\big) + (1-\lambda)\mathcal{F}^*\big(P^{h_2}_{\nabla I_k}\big)(\theta)\right)(\theta) \\
=&\big(\lambda a_1 + (1-\lambda)a_2\big)\mathcal{F}^*\big(P(\nabla I_k)\big) + \big(\lambda\mathbf{b}_1 + (1-\lambda)\mathbf{b}_2\big) \cdot \mathbf{w}_\theta, \\
=&\mathcal{F}^*\big(P^{h_\lambda}_{\nabla I_k}\big)(\theta),
\end{aligned}$$

where $h_\lambda(\mathbf{z}) = \big(\lambda a_1 + (1-\lambda)a_2\big)\mathbf{z} + \lambda\mathbf{b}_1 + (1-\lambda)\mathbf{b}_2 \in \mathbb{H}_0$. Hence $\lambda\mathcal{F}^*\big(P^{h_1}_{\nabla I_k}\big) + (1-\lambda)\mathcal{F}^*\big(P^{h_2}_{\nabla I_k}\big) = \mathcal{F}^*\big(P^{h_\lambda}_{\nabla I_k}\big) \in \mathcal{F}^*\big(\mathbb{P}_{\mathbb{H}_0,k}\big)$.

With (4.24) in mind, we define the Discrete Sliced Wasserstein Distance [160]

$$\begin{aligned}
d\big(P_{\mathbf{Z}^{(1)}}, P_{\mathbf{Z}^{(2)}}\big) :&= ||\widehat{P}_{\mathbf{Z}^{(1)}} - \widehat{P}_{\mathbf{Z}^{(2)}}||_{L^2([0,\pi),\mathbb{R}^N)} \\
&= ||\mathcal{F}^*\big(P_{\mathbf{Z}^{(1)}}\big) - \mathcal{F}^*\big(P_{\mathbf{Z}^{(2)}}\big)||_{L^2([0,\pi),\mathbb{R}^N)} \\
&= \sqrt{\int_0^\pi ||\mathcal{F}\big(P_{\mathbf{Z}^{(1)}_\theta}\big) - \mathcal{F}\big(P_{\mathbf{Z}^{(2)}_\theta}\big)||^2 d\theta}
\end{aligned} \tag{4.28}$$

where $||\cdot||$ denotes the Euclidean norm in $\mathbb{R}^N$. In particular, when $P_{\mathbf{Z}}$ is a gradient distribution, say $P_{\mathbf{Z}} = P_{\nabla I_k}$, then

$$\widehat{P}_{\nabla I_k}(\theta) = \Big(\mathcal{F}^*(P_{\nabla I_k})\Big)(\theta) = \mathcal{F}(P_{\nabla_\theta I_k}). \tag{4.29}$$

[2]This projection to $\mathbf{w}_\theta$ direction for every θ is similar to the sliced projections in Radon transform, and this is why we have "Radon" in the name of this new transform.

In other words, applying the Discrete R-CDT (operator $\mathcal{F}^*$) on the local 2D discrete distribution (i.e, $P_{\nabla I^c_{1,k}}$) equates to applying Discrete CDT Transform(i.e., $\mathcal{F}$) on a collection of projected one-dimensional representations[3] of (e.g., $P_{\nabla_\theta I_k}$), indexed by $\theta \in [0, \pi)$. The corresponding Discrete Sliced Wasserstein distance between two patch gradient distributions $P_{\nabla I^t_k}$ and $P_{\nabla I^{c,j}_k}$ is

$$\begin{aligned} d(P_{\nabla I^t_k}, P_{\nabla I^{c,j}_k}) &= ||\widehat{P}_{\nabla I^t_k} - \widehat{P}_{\nabla I^{c,j}_k}||_{L^2([0,\pi),\mathbb{R}^N)} \\ &= \sqrt{\int_0^\pi ||\mathcal{F}\big(P_{\nabla_\theta I^t_k}\big) - \mathcal{F}\big(P_{\nabla_\theta I^{c,j}_k}\big)||^2 d\theta}. \end{aligned} \tag{4.30}$$

The minimization problem (4.12) is hence equivalent to

$$\begin{aligned} c^* &= \arg\min_c \sum_{k=1}^K d(P_{\nabla I^t_k}, \mathbb{P}^c_{\mathbb{H}_0,k}) \\ &= \arg\min_c \sum_{k=1}^K \min_{h_k \in \mathbb{H}_0} ||\widehat{P}_{\nabla I^t_k} - \widehat{P}^{h_k}_{\nabla I^c_k}||_{L^2([0,\pi),\mathbb{R}^N)}. \end{aligned} \tag{4.31}$$

Remark 3.1: In practice, we take θ from a finite set $\{\theta_1, ..., \theta_m\}$ for some positive integer m, $\widehat{P}_{\nabla I_k} = \mathcal{F}^*(P_{\nabla I_k})$ can be represented by a matrix of size $N \times m$ and reshaped as a long vector of length $m * N$.

In summary, the Discrete R-CDT transform takes 2D discrete distribution as input and outputs a sequence of vectors indexed by θ in some finite set.

4.3.3 Nearest subspace learning

Next we leverage the generative model as in (4.10) together with the Discrete R-CDT transform $\mathcal{F}^*$ to form a nearest subspace classification method to facilitate the classification strategy in (4.12). It is not hard to see that $\mathcal{F}^*\big(\mathbb{P}^c_{\mathbb{H}_0,k}\big)$ is convex, meaning that $\lambda\mathcal{F}^*\big(P^{h_1}_{\nabla I^c_k}\big) + (1-\lambda)\mathcal{F}^*\big(P^{h_2}_{\nabla I^c_k}\big)$ lies in $\mathcal{F}^*\big(\mathbb{P}_{\mathbb{H}_0,k}\big)$ for all $\lambda \in [0,1]$ and $h_1, h_2 \in \mathbb{H}_0$. Indeed, the deformations $\mathbb{H}_0$ in 2D discrete distribution space also cause the corresponding translation and scaling effects in Discrete R-CDT transform

[3]This projection is similar to the sliced projections in the Radon transform, and this is why we have "Radon" (R) in the name of this new transform.

space:

$$\begin{aligned}\mathcal{F}^*\big(P_{(\alpha\nabla I_k+\mathbf{b})}\big)(\theta) &= \mathcal{F}\big(P_{(\alpha\nabla_\theta I_k+\mathbf{b}\cdot\mathbf{w}_\theta)}\big)\\ &= \alpha\mathcal{F}(P_{\nabla_\theta I_k}) + \mathbf{b}\cdot\mathbf{w}_\theta\\ &= \alpha\mathcal{F}^*(P_{\nabla I_k})(\theta) + \mathbf{b}\cdot\mathbf{w}_\theta, \end{aligned} \tag{4.32}$$

where the second equation follows from the composition property of the Discrete CDT transform (please see Section 4.3.2) and the addition on the RHS is operated entry-wise. We then expand the convex set $\mathcal{F}^*\big(\mathbb{P}_{\mathbb{H}_0,k}\big)$ to form a subspace $\mathbb{V}_k^c = \text{span}\big(\mathcal{F}^*(\mathbb{P}^c_{\mathbb{H}_0,k})\big)$ and further assume that when $c \neq c'$, there exists a patch k such that $\mathbb{P}^c_{\mathbb{H}_0,k} \cap \mathbb{V}_k^{c'} = \emptyset$, which is consistent with the assumption that the image class of subject c will not overlap with images of a different subject c' under all possible illumination variations. With the above considerations in mind, the constrained minimization problem in (4.12) or (4.31) can be modified to a simple subspace projection problem (4.33), which can be solved by basic linear algebra techniques in transform domain as shown in sections below:

$$\begin{aligned} c^* &= \arg\min_c \min_{\widehat{P}\in\mathbb{V}_k^c} \sum_{k=1}^K ||\widehat{P}_{\nabla I_k^t} - \widehat{P}||_{L^2([0,\pi),\mathbb{R}^N)}\\ &= \arg\min_c \sum_{k=1}^K d\big(\widehat{P}_{\nabla I_k^t}, \mathbb{V}_k^c\big)\\ &= \arg\min_c \sum_{k=1}^K d_k^c \end{aligned} \tag{4.33}$$

where $d_k^c := d\big(\widehat{P}_{\nabla I_k^t}, \mathbb{V}_k^c\big)$ is the $d(\cdot,\cdot)$ distance of $\widehat{P}_{\nabla I_k^t}$ to the subspace $\mathbb{V}_k^c$, which can be computed in a convenient form as a least squares projection as shown below. In summary, the class is determined by the smallest distance $d^c = \sum_{k=1}^K d_k^c$.

We prove that using the smallest d^c in equation (4.33) solves the classification problem.

Proposition: Assume that for any $c \neq c'$, there exists a k_0 (possibly depending on c, c') such that $\mathbb{P}^c_{\mathbb{H}_0,k_0} \cap \mathbb{V}_{k_0}^{c'} = \emptyset$. Then given a test image $I^t \in \mathbb{S}^c$,

$$d^c = \sum_{k=1}^K d_k^c = \sum_{k=1}^k d(P_{\nabla I_k^t}, \mathbb{V}_k^c) = 0, \tag{4.34}$$

while

$$d^{c'} = \sum_{k=1}^K d_k^{c'} = \sum_{k=1}^k d(P_{\nabla I_k^t}, \mathbb{V}_k^{c'}) > 0, \tag{4.35}$$

if $\mathbb{V}_{k_0}^{c'}$ is a closed subspace[4].

Proof: Generally speaking, given a closed subspace $\mathbb{V}$ of a metric space with distance metric d, $d(v, \mathbb{V}) > 0$ if and only if $v \notin \mathbb{V}$. To show (4.34), it suffices to show that $P_{\nabla I_k^t} \in \mathbb{P}_{\mathbb{H}_0,k}^c \subseteq V_k^c$ for $k = 1, ...K$, which follows from the definition of the generative model $\mathbb{S}^c$ and the fact that $I^t \in \mathbb{S}^c$.

On the other hand, to show (4.35), it suffices to show that there exists a k_0 such that $P_{\nabla I_{k_0}^t} \notin V_k^{c'}$, which follows from the assumption that $\mathbb{P}_{\mathbb{H}_0,k_0}^c \cap \mathbb{V}_{k_0}^{c'} = \emptyset$ and the fact that $P_{\nabla I_{k_0}^t} \in \mathbb{P}_{\mathbb{H}_0,k_0}^c$. Hence we have that $d_{k_0}^{c'} = d(P_{\nabla I_{k_0}^t}, \mathbb{V}_{k_0}^{c'}) > 0$.

In particular, if a test image I^t belongs to subject c, $\sum_{k=1}^K d_k^c < \sum_{k=1}^K d_k^{c'}$ for any $c' \neq c$.

Implementation

We present implementation details regarding the proposed algorithm for the training and testing phases. Specifically, the training phase is to build a subspace for each image patch of each subject using all training samples. In the testing phase, firstly, for each image patch, the distance between the testing sample and the subspace of each image patch of each subject is computed. Then, the distance from image patch is aggregated. Finally, the classification result is the subject who has the shortest aggregated distance. The detailed training and testing algorithms are described below.

Training Given a total of $\{I_l^c\}_{l=1}^L$ L training images for c_{th} subject. Each image I_l is partitioned into K image patches $\{I_{l,k}^c\}_{k=1}^K$. We approximate the subspace $\mathbb{V}_k^c$ using L training images $\{I_l^c\}_{l=1}^L$ by

$$\mathbb{V}_k^c = \text{span}\Big(\{\widehat{P}_{\nabla I_{1,k}^c}, \cdots, \widehat{P}_{\nabla I_{L,k}^c}\} \cup \mathbb{U}_T\Big), \tag{4.36}$$

where $\mathbb{U}_T = \{\mu_1(n,\theta), \mu_2(n,\theta)\}$ with $\mu_1(n,\theta) = \cos\theta$, $\mu_2(n,\theta) = \sin\theta$ can be used to automatically model translation and scaling within a subject gradient distribution class by observing equation (4.32)[151, 162].

For each class c and each patch k,

1. Compute the transforms $\widehat{P}_{\nabla I_{1,k}^c}, \cdots, \widehat{P}_{\nabla I_{L,k}^c}$ corresponding to the training images
2. Use Principal Component Analysis (PCA) [163], keeping enough components to retain 99% of the training data variance, to orthogonalize $\{\widehat{P}_{\nabla I_{1,k}^c}, \cdots, \widehat{P}_{\nabla I_{L,k}^c}\} \cup U_T$ to obtain a set of

[4]In practice, $\mathbb{V}_{k_0}^{c'}$ is a finite dimensional space and is hence closed.

orthonormal basis vectors $\{v^c_{1,k}, v^c_{2,k}, \cdots\}$ and form a matrix B^c_k with $\{v^c_{1,k}, v^c_{2,k}, \cdots\}$ as its columns:

$$B^c_k = [v^c_{1,k}, v^c_{2,k}, \cdots]. \tag{4.37}$$

When enough training data is available, we split the training data into training and validation sets, and choose the smallest number of components that allow for highest classification accuracy on the validation set.

It is worth noting that if the deformation strictly follows the set $\mathbb{H}_0$, only taking span of one transformed training example with $\mathbb{U}_T$ is necessary. However, in reality it is often the case that more complicated illumination effects than defined in $\mathbb{H}_0$ are present. We can enhance $\mathbb{H}_0$ by using any available training images using eq. (4.36) to take span of multiple transformed training examples with $\mathbb{U}_T$. This technique allows the proposed method to learn from more complicated lighting variations.

Testing Given a testing image I^t, the first step is to segment I^t into K image patches $\{I^t_k\}^K_{k=1}$ and calculate the Discrete R-CDT transform representation $\{\widehat{P}_{\nabla I^t_k}\}^K_{k=1}$. Then, for each image patch k, we compute the distance

$$d(\widehat{P}_{\nabla I^t_k}, \mathbb{V}^c_k) = ||\widehat{P}_{\nabla I^t_k} - B^c_k (B^c_k)^T \widehat{P}_{\nabla I^t_k})||, \tag{4.38}$$

where $(B^c_k)^T$ is the transpose of matrix B^c_k and $||\cdot||$ denote the Euclidean norm.

Finally, we compute d^c by summing the distance contribution d^c_k and search for the nearest subspace as the classification result via:

$$\arg\min_c d^c = \sum_{k=1}^{K} d^c_k. \tag{4.39}$$

4.4 Experiment

In order to evaluate the proposed algorithm, we compare it with multiple illumination-invariant face recognition algorithms [134, 133, 135, 136, 142, 147] and several deep learning based alternatives [139, 138, 140] with illumination data augmentation strategy on three face dataset with challenging illumination conditions [129, 164, 165]. We first present the experiment setup, discuss the experimental evaluation results, and provide a hyperparameter study regarding parameter selection at the end.

Table 4.1: Image acquisition setup for Extended Yale Face Database B

	Number of Images	Azi. & Ele. angles
Training set	11	-10 $\leq$ Azi. $\leq$ 10 and -20$\leq$ Ele. $\leq$ 20
Test subset 1	10	-25 $\leq$ Azi. <-10 or 10<Azi. $\leq$ 25
Test subset 2	18	-60 $\leq$ Azi. <-25 or 25< Azi. $\leq$ 60; Azi.=0 and Ele. = 35 or = 45
Test subset 3	12	-95$\leq$ Azi. < -60 or 60 < Azi. $\leq$ 95
Test subset 4	13	Azi. < -95 or Azi. > 95; Azi.=0 and Ele.=90

4.4.1 Experiment setup

We evaluate the proposed method on three different face recognition datasets with illumination variations: Extended Yale Face Database B [129], CAS-PEAL dataset [164], and AR Face Dataset [165]. The Extended Yale Face Database B [129] has 38 different subjects, each of which has 64 images under different illumination conditions. It is worth noting that several subjects do not have 64 images due to corrupted images during the acquisition phase as explained in [129]. Face images are collected by changing angles between light source direction and the camera axis, as indicated in Table 4.1, thus having various illumination effects. Based on the degree of lighting conditions, the dataset is divided into five subsets as commonly done [135, 136, 133]. Specifically, the *training subset*, *test subset 1*, *test subset 2*, *test subset 3*, and *test subset 4* contain 11, 10, 18, 12, and 13 images, respectively, as shown in Table 4.1. Each image has a distinct lighting condition. Fig. 4.7 and Fig. 4.8 exemplify face images from the *training subset* and *testing subset*. It is easy to observe that images from *testing subset* has a high degree illumination variations, including the most extreme illumination conditions in *testing set 4* where subjects are hardly visible in dark environments. CAS-PEAL dataset [164] has 233 subjects, each of which has more than 9 images under different lighting conditions. However, unlike the Extended Yale Face Database B, each subject has only one "clean" image that is acquired under standard illumination. Others are the images with varying lighting effects. Based on the azimuth angle (e.g., $\{-90, -45, 0, 45, 90\}$) and elevation angle (e.g., $\{-45, 0, 45\}$), the face images with various levels of illumination are categorized into three different subsets as well. The azimuth angle for *Test subset 1*, *Test subset 2*, and *Test subset 3* are 0 degree, -45 or 45 degree, and -90 or 90 degree, respectively. Fig. 4.7 and 4.8 illustrate sample images from the CAS-PEAL-R1 dataset, including the "clean" images and images with changing lighting conditions. Compared with the previous two datasets, the AR face dataset has less illumination variations as shown in Fig. 4.9. Each subject in AR face dataset has two sets of images that are

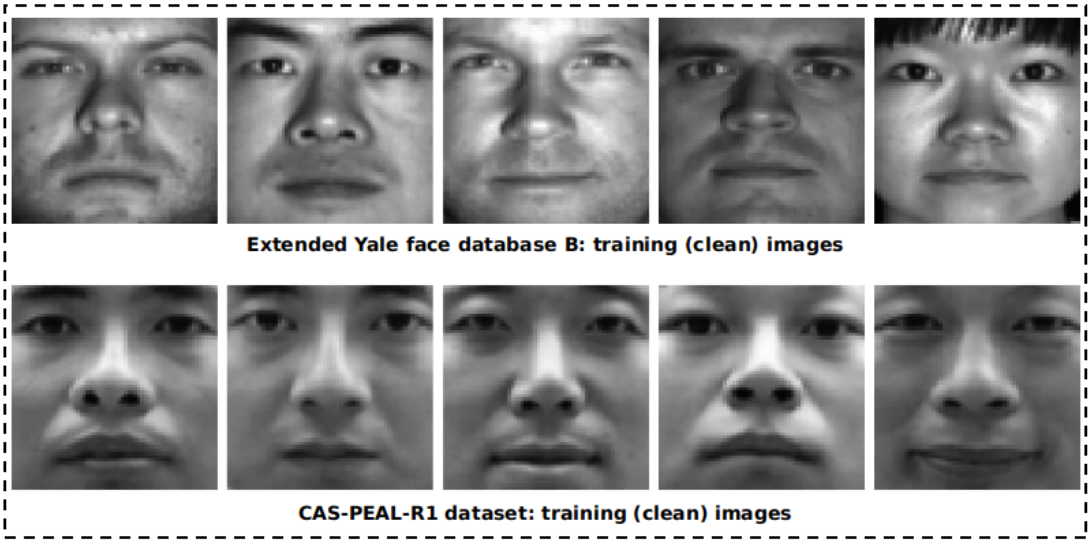

Figure 4.7: Clean images from the Extended Yale face database B and CAS-PEAL-R1 dataset.

taken in two separated sessions. Each set contains one "clean" image and 3 images with different levels of illumination. In total, each subject has 2 clean images and 6 images with different lighting variations. Accordingly, we construct the *test subset 1* and *test subset 2* for testing. The images with left and right illumination are included in *test subset 1*. The images with side lighting on are contained in *test subset 2*. The clean images serve as the training samples.

The experiment evaluation focuses on comparisons between the proposed method and the illumination-invariant feature extraction based approaches, plus the deep learning based approaches with illumination data augmentation, because illumination-invariant feature extraction methods, which aim to eliminate lighting effects by holistic decomposition [134], quotient models [135, 136], or logarithm difference [133], are the mainstream algorithms for illumination-invariant face analysis, while deep learning based approaches with illumination augmentation are the most popular methods. Specifically, we consider three state-of-the-art deep learning models: VGGFace, ResNet-50 and DenseNet-121. We use 90% and 10% of the original training data for training and validation, respectively. Validation is performed every ten iterations, the final test accuracy is based on the model checkpoint that has the best validation accuracy. When there is only one training sample available, each sample is augmented 5 times using the illumination model stated in equation 4.6. For all the experiments we use an Adam optimizer [166] with a learning rate of 0.001.

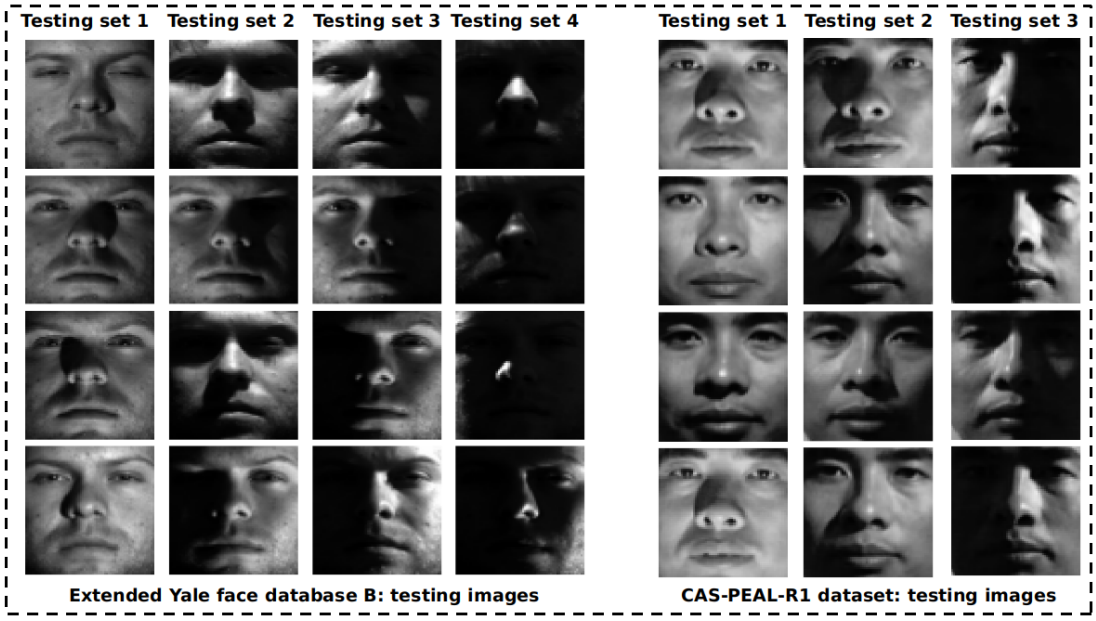

Figure 4.8: Image examples with changing illumination effects from one subject in Extended Yale face database B and CAS-PEAL-R1 dataset.

Figure 4.9: Image examples from AR face dataset.

4.4.2 Evaluation results

Two experiments, denoted as *Test 1* and *Test 2*, were conducted to evaluate the proposed approach using different training strategies on the Extended Yale Face Database B. *Test 1* trains the proposed model by utilizing all images in the *training set*, and performs evaluations on *Test subsets 1, 2, 3, & 4.* respectively. We further remark that in *Test 1*, since there are multiple training images available, we set aside a validation set to choose the number of components in the orthogonalization step (please see Method Section) using PCA. Specifically, We use a 66% and 33% of original training data for training and validation. We use the validation set to define the smallest number of components that maximize classification accuracy. *Test 2* trains the proposed model using one single sample, which is randomly selected from the *training set*, then it follows the same testing procedure as discussed above in *Test 1.* Table 4.2 presents experimental evaluation results of *Test 1* and *Test 2.* Clearly, the proposed method achieves the top and most robust performance across different test subsets in both of *Test 1* and *Test 2.* Specifically, in *Test subset 1&2* of *Test 1*, most of the approaches achieve satisfactory performance, meaning that they are able to address mild illumination effects for face recognition tasks. With increasing levels of illumination conditions in *Test subset 2,3,&4*, such as when uneven and large areas of shadow are present, other methods experienced significant decreases in performance. A similar trend can be observed in *Test 2* as well. Of the comparison methods, the WebberFace, GradientFace, and MSLDE approaches consistently achieve better performance than their deep learning counterparts, indicating that they are able to eliminate some illumination effects after the "normalization" stage, especially in the test subsets (e.g., *Test subset 1&2*) with low-level of illumination variations. In addition, incorporating the log transform into HoG + SVM pipeline also improves performance. With regards to the deep learning-based methods, some achieve excellent performance in *Test subset 1&2* of *Test 1* and *Test 2*, where the minimum illumination effects exist. Nevertheless, for *Test subset 3&4*, their classification accuracy decrease significantly. This may be due to the fact that only a limited number of training samples are available. As commonly done in the machine learning literature, we employed a data augmentation strategy to address the shortage of training data. Specifically, we adopt the model stated in equation (4.6) to randomly augment training samples with different levels of illumination (e.g., to simulate various contrast and brightness changes)[5]. Each parameter in equation (4.6) is randomly configured as $\alpha \in [0.1, 3]$, $\beta \in [1, 30]$, $\mathbf{b} \in [0.1, 3]$). We observe that for deep learning approaches illumination data augmentation strategy improves classification performance to a degree in cases where illumination

[5]Please note in here equation (4.6) is applied to the entire image domain rather than a single patch.

Table 4.2: Classification accuracy for Yale face dataset

	Extended Yale face database (Number of classes = 38)							
	Test 1				*Test 2*			
	Test subset 1	*Test subset 2*	*Test subset 3*	*Test subset 4*	*Test subset 1*	*Test subset 2*	*Test subset 3*	*Test subset 4*
WebberFace + SVM	**100**%	95.1%	94.2%	90.6%	92.1%	78.8%	70.1%	74.4%
MSLDE + SVM	**100**%	93.3%	82.0%	79.1%	87.6%	64.9%	60.8%	64.0%
GradientFace + SVM	**100**%	91.5%	88.4%	85.1%	80.5%	58.6%	65.2%	60.0%
Log + DCTface + SVM	97.8%	80.5%	70.7%	43.6%	72.8%	41.3%	39.8%	24.6%
HoG + SVM	**100**%	82.4%	49.1%	65.9%	68.6%	47.2%	30.7%	42.0%
Log + HoG + SVM	**100**%	91.1%	70.5%	80.6%	73.1%	51.9%	44.2%	51.0%
VGGface	67.8%	14.6%	3.7%	2.4%	15.0%	5.7%	2.4%	2.6%
VGG face with data aug.	98.6%	77.1%	55.5%	25.3%	66.0%	33.5%	10.6%	7.7%
ResNet-50	96.5%	50.3%	9.5%	4.0%	63.9%	14.9%	3.5%	2.2%
ResNet-50 with data aug.	97.8%	73.3%	42.0%	26.9%	91.0%	49.3%	23.8%	14.4%
DenseNet-121	85.5%	36.4%	5.9%	3.8%	57.1%	15.5%	4.6%	2.0%
DenseNet-121 with data aug.	98.1%	58.7%	23.4%	14.8%	78.4%	24.8%	9.2%	2.0%
Discrete R-CDT + NS(ours)	**100**%	**98.8**%	**96.2**%	**94.4**%	**98.4**%	**95.5**%	**92.4**%	**91.8**%

changes are small (e.g., in *Test subset 1&2*). However, when illumination effects become severe, e.g., in *Test subset 3&4*, data augmentation become less effective. Though using data augmentation is a valid option to increase performance, it is still extremely difficulty to prescribe how to augment the data, in addition to other issues such as computational complexity and out of distribution performance issues [162]. Finally, we note that the performance of deep learning-based approaches can vary significantly due to the network architecture.

In this experiment with respect to CAS-PEAL-R1, the model only utilizes the single clean image of each subject for training. Evaluations are performed on three testing subsets: *Test subset 1, 2, & 3*. Table 4.3 summarizes the experiment results. Overall, the proposed method outperforms other alternatives by a large margin, which is a consistent trend across the three testing subsets. With respect to the comparison methods, GradientFace and MSLDE achieve the best performance. Deep learning-based approaches are not able to perform well perhaps due to the limited number of training samples, even illumination data augmentation strategies are used.

Finally, we evaluate and compare the proposed method on AR face dataset which has 100 subjects. It is worth noting that the framework Log + HoG [142] method achieves the best performance while the method proposed here obtains comparable performance. Other methods such as MSLDE and GradientFace also have high accuracies. It is worth noting that illumination data augmentation are most effective for the AR face dataset, which allows deep learning approaches to gain significant performance improvement, implying that employing data augmentation is an effective route to address mild illumination variations.

Simulated mouth dataset for facial weakness analysis We also evaluate the proposed method on a neurologist-verified mouth dataset [108]. The mouth dataset contains 361 mouth images including 76 mouth images with left facial weakness, 72 mouth images with right facial weakness,

Table 4.3: Classification accuracy for CAS-PEAL dataset

	CAS-PEAL face dataset (Number of classes = 233)		
	Test subset 1	*Test subset 2*	*Test subset 3*
WebberFace + SVM	16.2%	19.5%	27.5%
MSLDE + SVM	18.0%	17.9%	40.0%
GradientFace + SVM	21.1%	20.0%	38.4%
Log + DCTface + SVM	11.4%	8.5%	4.0%
HoG + SVM	19.3%	13.2%	19.0%
Log + HoG + SVM	22.8%	17.2%	24.3%
VGGface	2.3%	1.7%	0.7%
VGG face with data aug.	5.5%	3.5%	2.6%
ResNet-50	3.8%	1.8%	1.0%
ResNet-50 with data aug.	6.9%	4.8%	5.1%
DenseNet-121	1.4%	1.1%	0.8%
DenseNet-121 with data aug.	3.3%	2.0%	1.0%
Discrete R-CDT + NS(ours)	**42.2**%	**50.2**%	**49.7**%

Table 4.4: Classification accuracy for AR face dataset

	ARFace dataset (Number of classes = 100)	
	Test subset 1	*Test subset 2*
WebberFace + SVM	93.7%	87.0%
MSLDE + SVM	97.5%	94.5%
GradientFace + SVM	97.5%	93.0%
Log + DCTface + SVM	87.5%	73.0%
HoG + SVM	99.7%	83.5%
Log + HoG + SVM	**100.0**%	**95.5**%
VGGface	19.7%	11.5%
VGG face with data aug.	80.2%	58.4%
ResNet-50	59.5%	7.5%
ResNet-50 with data aug.	92.7%	80.0%
DenseNet-121	54.2%	9.0%
DenseNet-121 with data aug.	89.9%	82.4%
Discrete R-CDT + NS(ours)	99.5%	94.5%

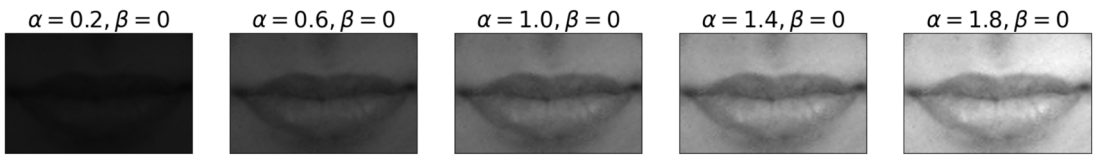

Figure 4.10: The perturbed mouth dataset.

and 213 mouth images without facial weakness (normal). We first perform the standard 5-fold cross-validation testing to show the effectiveness of all methods on the unperturbed mouth dataset. Then we perform a out-of-distribution test to illustrate that the proposed method is robust to simulated lighting perturbations as shown in Fig 4.10. To be precise, during the test phase of out-of-distribution test, test images are transformed via eq.(4.6) while the training images were the original images. We report the accuracy averaged over these 5 times in Table. 4.6. The second column of Table. 4.6 demonstrates that all methods achieve similar performance to discriminate three classes. We use two illumination perturbation models: $J(\mathbf{x}) = c_1 I(\mathbf{x}) + c_0$ ($c_1 \in [0.1, 3]$, $c_0 \in [1, 30]$) to simulate contrast and brightness variations and $J(\mathbf{x}) = c_1 I(\mathbf{x}) + c_0 + \mathbf{b}^T \mathbf{x}$ ($c_1 \in [0.1, 3]$, $c_0 \in [1, 30]$, $b_x \in [0.1, 3]$, $b_y \in [0.1, 3]$) to simulate contrast and brightness variations plus a smooth linear lighting gradient. It is easy to observe that the proposed method outperforms other comparison methods while performance of state-of-the-art deep learning models suffer significantly under simulated illumination variations. Especially, the linear gradient item $\mathbf{b}$ causes notable performance decrease for all methods except the proposed one.

Hyper-parameter study The experimental results presented above show that the face recognition approach we proposed is more robust to variations in illumination conditions than a variety of existing methods. The method is based on splitting face images into finite support neighborhoods Ω_k, and using a Discrete R-CDT representation for the gradient distribution within each neighborhood. As such, certain parametric choices (number of projections in R-CDT, neighborhood size, and neighborhood overlap) have to be made. To study the best choices for these parameters, we followed the same training and testing strategy used in *Test 2* for the Yale Face dataset and focus on the *Test subset 4*, which has the most challenging illumination conditions. Table 4.5 provides performance comparison using different parameter configurations such as varying cell size, overlap size, and number of projections. When varying different patch sizes, the overlap size is set as 0. Overall, results show that utilizing smaller patch improves the performance. We postulate this may be because the illumination model described in equation (4.6) is more accurate for small neighbor-

Table 4.5: Hyper-parameter study

	Number of projections					
	2	3	4	8	20	45
cell size = 16	35.1%	40.0%	38.7%	39.7%	39.5%	40.0
cell size = 8	62.4%	65.3%	63.2%	69.7%	69.4%	69.5
cell size = 4	78.5%	86.5%	87.3%	89.3%	89.3%	89.3
Overlap size						
	0 cell size = 8	2 cell size = 8	4 cell size = 8	0 cell size = 4	1 cell size = 4	2 cell size = 4
Accuracy	63.2%	73.2%	76.9%	87.3%	91.8%	93.4

Table 4.6: Classification accuracy for the perturbed mouth dataset

	Cross validation	Perturbation model: $J(\mathbf{x}) = c_1 I(\mathbf{x}) + c_0$	Perturbation model: $J(\mathbf{x}) = c_1 I(\mathbf{x}) + c_0 + \mathbf{b}^T \mathbf{x}$
HoG + SVM [141]	92.5 1.4	88.6 3.0	64.8 5.9
ResNet-18 [138]	96.4 1.8	77.0 4.0	42.4 9.8
ResNet-50 [138]	94.7 2.3	66.2 3.6	44.9 7.5
VGG-16 [139]	95.0 2.2	83.3 5.6	57.1 2.7
DenseNet-121 [140]	94.7 5.0	71.7 2.8	38.5 8.0
Discrete R-CDT + NS (ours)	**95.0 1.6**	**92.2 1.3**	**81.1 4.4**

hoods. Secondly, when the neighborhood size is fixed, increasing number of projections and overlap size are able to provide performance improvement to some extend.

Chapter 5

Discussion and future work

5.1 Vision-based facial weakness detection

We study the topics of vision-based facial weakness detection and quantification from static images or videos on several datasets labeled and verified by board-certified neurologists. The proposed study first presents an automatic pathological facial weakness detection tool for static images using a supervised learning method, showing that shape-based (e.g, landmarks) features are informative for classifying normal v.s. pathological facial weakness. The second study illustrates that landmark-based approaches suffer from inaccurate localization issues, even these landmarks are extracted from several state-of-the-art landmark extraction algorithms. To address this issue, a straightforward solution is to utilize a combination of shape-based features and landmarks-based features. In addition, we employ a linear SVM classifier to show that the shape-based features and texture-based features contain clinically meaningful information for identifying facial weakness. Utilizing the knowledge learned from two previous studies, we aim to examine facial weakness from video data. The proposed framework first leverages a subspace learning model, which is derived from the previous two studies, to extract the most discriminant shape and appearance information regarding facial weakness from each individual frame. Then a recurrent neural network models the temporal dynamics through a Bi-LSTM network and generate the prediction. Experimental evaluation shows that our solution is able to outperform other state-of-the-art alternatives, achieving the equal performance to paramedics. Most importantly, it is able to provide visualizable and interpretable results regarding facial weakness, which greatly increases model transparency and interpretability. Furthermore, a live and real-time prototype with interactive GUI is implemented on a regular laptop. The proof-

of-concept prototype is ready to be validated in the real-world clinical settings. The important implication of this study is that the proposed method opens a new opportunity of providing clinical assistance to non-neurologist providers such as paramedics to increase the coverage of standard neurological care in the prehospistal setting.

A substantial amount of experiences and lessons is gained from our study. First, building such an "in-the-wild" dataset not only time-consuming, labor-intensive but also expensive in terms of neurology expertise. The researchers spend numerous hours to collect and curate the image and video dataset based on the data acquisition protocol. It is also extremely costly to obtain the ratings from the board-certified neurologists for the verification process as well as other clinical raters such as paramedics and neurology residents. Therefore, collecting a real-patient dataset will be much challenging due to the prevalence of the disease, heterogeneity of symptom manifestations, and patient privacy standards. However, we believe the experience gained from assembling the "in-the-wild" dataset streamlines the data acquisition pipeline and enables us to avoid potential pitfalls, thus laying a solid foundation to enroll patients in the real-clinical setting. For the prototype implementation, one important issue is the head pose, because it is important to ensure that the subject faces the camera directly with a full view. Therefore, a head pose estimation model is embedded into the prototype, in order to check the subject's head. In the meanwhile, we also notice that multiple videos experience facial landmark detection failures due to the lighting and appearance variations. Therefore, in order to improve the landmark extraction accuracy, future work will also include to train a dedicated facial landmark extractor for facial weakness patients. In terms of ground truth labelling, given the fact that the data was collected from open access repositories and the in-person examinations of the same person were not available, it is still possible for all neurology experts to make the same mistakes for ground truth labeling. Therefore, to minimize the possible mislabeling issues, three board-certified experienced neurologists rated the image and video dataset independently. Only the image and video data with the concordant ratings are used our study in a such way that the ground truth labeling issues are minimized.

To sum up, to arrive at a full-fledged video solution for facial weakness detection and quantification that could be used in the real-world clinical setting, multiple future research efforts are required, such as collecting images and videos of real patients with in-person examinations results (e.g., the clinical diagnosis, imaging findings, and electronic health records), development of a dedicated facial landmark extractor, and testing the proposed facial weakness analysis framework on patients in the real-world clinical settings to improve the model generalizability and avoid overfitting issues.

5.2 Illumination-invariant face recognition

We propose a novel method for illumination invariant face representation that are based on the following ideas: (1) first, we patchify images into multiple image patches and calculate the 2d discrete gradient distributions for each individual image patch; (2) secondly, we shows that illumination variations cause optimal transport type of perturbation for local 2D discrete gradient distributions; (3) thirdly, we define a sliced optimal transport based representation to represent local 2D discrete gradient distributions, which has several properties to address the transport-like variations in 2D discrete gradient space and convexity the classification problem; (4) built upon this new representation, we construct a nearest subspace model to perform face recognition under varying illuminations. The experimental evaluations demonstrate the advantages of the proposed approaches when compared with other alternatives.

However, there are multiple interesting perspectives worth further exploring as future work. Firstly, in here, we empirically demonstrate that incorporating log transform into existing pipeline is able to increase classification accuracy. Specifically, as a widely used preprocessing step for multiple illumination-invariant face analysis work [134, 142, 133], the log transformation [152] is able to circumvent illumination effects. Thus, we also applied the log transform as a preproccessing step in the proposed approach. Table 5.1 demonstrates classification results as well as the performance improvement (as highlighted using an upward arrow) for three face datasets used in the experiment. Overall, leveraging log transform enhances the classification accuracy in most of testing cases, especially for the dataset that has large areas of illumination effects such as *test subset 4* in the Yale Face dataset. However, because using log transform changes the equation (4.6), a careful mathematical analysis of incorporating the log transform in our gradient distribution representation will be the subject of future work. Secondly, besides the translation and scaling effects, there are multiple other interesting effects such as rotation, anisotropic scaling, and shearing that could be future studied to improve the system's performance. In addition, in existing implementation, the proposed approach uses an equal weight coefficient strategy when aggregating the distance from each image patch. However, it is clear that several image patches (e.g., patches include eyes, noses, and mouths) contain more informative information, it is worth to investigate a learning-based approach to derive such weight coefficients for potential performance improvements. Last but not the least, equation (4.36) means that taking the span of multiple transformed training examples allows the proposed model to capture more complicated lighting variations than the ones captured by the set $\mathbb{H}_0$. Table 4.2 is able to support this claim, by demonstrating that with an increasing number of training images, the

Table 5.1: Classification accuracy using log transform

	Accuracy			
	Test subset 1	*Test subset 2*	*Test subset 3*	*Test subset 4*
Yale Face database (*Test 1*)	100%	98.9% (0.1% ↑)	96.6% (0.4% ↑)	95.7%(1.3% ↑)
Yale Face database (*Test 2*)	98.9%(0.5% ↑)	95.5%	95.5%(3.1% ↑)	96.3%(4.5% ↑)
	Test subset 1	*Test subset 2*	*Test subset 3*	-
CAS-PEAL-R1 dataset	45.2%(3.0% ↑)	54.3%(4.1% ↑)	52.4%(2.7% ↑)	-
	Test subset 1	*Test subset 2*	-	-
ARFace dataset	99.5%	95.5%(↑ 1.0%)	-	-

classification accuracy improves. The same trend is also applied to the CAS-PEAL-R1 face dataset. When the *test subset 1* and *test subset 2* are included into the training set, the classification accuracy increases from 49.7% to 82.51% accordingly.

To sum up, future work for illumination-invariant face recognition will include: (1) studying log transform; (2) modeling rotation, anisotropic scaling, and shearing effects of 2D discrete gradient distribution in discrete R-CDT domain; (3) obtaining weight coefficients for image patches in the nearest subspace learning setting.

Chapter 6

Conclusion

Facial weakness is a common neurological deficit that associated to lack of facial muscle control due to neurological injury. Thus, the configuration and movement of the face can indicate the presence or absence of several neurological diseases. In order to detect and quantify facial weakness from static images and videos, we develop an automated video-based system that can detect and quantify facial weakness using medical imaging analysis and computer vision. Our studies start from the image-based facial weakness analysis by showing that (1) shape-based features and texture-based features are effective for facial weakness detection and quantification, and (2) clinically meaningful pathology can be detected by the proposed approach. In addition, we provide a fully-automated video solution for facial weakness detection and quantification, which not only achieves equivalent performance to paramedics, but also provides the visualizable and interpretable results to illustrate how shape and appearance-based features are used. Lastly, a prototype is implemented on a regular laptop to demonstrate the feasibility of our study as a proof-of-concept.

In the meantime, in our studies, illumination variability is able to affect performance of the proposed algorithm. To address lighting variations issues, inspired by the fact that patch-wise texture-based features compare favorably to other alternatives in our case, we present a new sliced-Wasserstein distance based representation for face analysis that is beyond the scope of facial weakness analysis. The proposed approach makes use of the lighting-insensitive measure, image gradients, to construct a low-level patch-wise image feature set based on optimal transport metric. To be specific, first we mathematically demonstrate that lighting variation causes a transport-type perturbation for the 2D discrete distributions of the image gradient. The transport-type perturbation can be alleviated using optimal transport-based metric in R-CDT domain, which convexifies the associated

face analysis tasks. After being convexified, a nearest subspace learning approach is utilized to classify these convex sets in discrete R-CDT domain. The experiment results demonstrate the superiority of the proposed method compared to other approaches.

References

[1] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *NPJ digital medicine*, 3(1):1–8, 2020.

[2] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12, 2020.

[3] Shi Cheng, Lakshman S Tamil, and Benjamin Levine. A mobile health system to identify the onset of paroxysmal atrial fibrillation. In *2015 International Conference on Healthcare Informatics*, pages 189–192. IEEE, 2015.

[4] Alex Mariakakis, Megan A Banks, Lauren Phillipi, Lei Yu, James Taylor, and Shwetak N Patel. Biliscreen: smartphone-based scleral jaundice monitoring for liver and pancreatic disorders. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):20, 2017.

[5] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, and Jorge Cuadros. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[6] Thomas Martini Jørgensen, Andreas Tycho, Mette Mogensen, Peter Bjerring, and Gregor BE Jemec. Machine-learning classification of non-melanoma skin cancers from image features obtained by optical coherence tomography. *Skin Research and Technology*, 14(3):364–369, 2008.

[7] Jacob Levman, Tony Leung, Petrina Causer, Don Plewes, and Anne L Martel. Classification of dynamic contrast-enhanced magnetic resonance breast lesions by support vector machines. *IEEE Transactions on Medical Imaging*, 27(5):688–696, 2008.

[8] Titus J Brinker, Achim Hekler, Alexander H Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Tim Holland-Letz, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, 113:47–54, 2019.

[9] RR Rawat, I Ortega, P Roy, F Sha, D Shibata, et al. Deep learned tissue "fingerprints" classify breast cancers by er/pr/her2 status from h&e images. sci rep 10: 7275, 2020.

[10] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.

[11] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567, 2018.

[12] Y.n LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[13] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):1–9, 2021.

[14] Albert Haque, Arnold Milstein, and Li Fei-Fei. Illuminating the dark spaces of healthcare with ambient intelligence. *Nature*, 585(7824):193–202, 2020.

[15] Jedrek Wosik, Marat Fudim, Blake Cameron, Ziad F Gellad, Alex Cho, Donna Phinney, Simon Curtis, Matthew Roman, Eric G Poon, Jeffrey Ferranti, et al. Telehealth transformation: Covid-19 and the rise of virtual care. *Journal of the American Medical Informatics Association*, 27(6):957–962, 2020.

[16] W. Wang, A. den Brinker, et al. Discriminative signatures for remote-ppg. *IEEE Transactions on Biomedical Engineering*, 67(5):1462–1473, 2019.

[17] S. Yeung, F. Rinaldo, et al. A computer vision system for deep learning-based detection of patient mobilization activities in the icu. *NPJ digital medicine*, 2(1):1–5, 2019.

[18] F. Deng, J. Dong, et al. Design and implementation of a noncontact sleep monitoring system using infrared cameras and motion sensor. *IEEE Transactions on Instrumentation and Measurement*, 67(7):1555–1563, 2018.

[19] Z. Huang, Y. Liu, et al. Video-based fall detection for seniors with human pose estimation. In *2018 4th International Conference on Universal Village (UV)*, pages 1–4. IEEE, 2018.

[20] J. Thevenot, M. López, and A. Hadid. A survey on computer vision for assistive medical diagnosis from faces. *IEEE journal of biomedical and health informatics*, 22(5):1497–1511, 2017.

[21] Jeffrey De Fauw, Joseph R Ledsam, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

[22] Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE transactions on medical imaging*, 35(5):1153–1159, 2016.

[23] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[24] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 2021.

[25] Etienne Bennequin, Victor Bouvier, Myriam Tami, Antoine Toubhans, and Céline Hudelot. Bridging few-shot learning and adaptation: New challenges of support-query shift. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 554–569. Springer, 2021.

[26] Paul Kruszka, Antonio R Porras, Andrew K Sobering, Felicia A Ikolo, Samantha La Qua, Vorasuk Shotelersuk, Brian HY Chung, Gary TK Mok, Annette Uwineza, and Mutesa. Down syndrome in diverse populations. *American Journal of Medical Genetics Part A*, 173(1):42–53, 2017.

[27] Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

[28] Hananel Hazan, Dan Hilu, Larry Manevitz, Lorraine O Ramig, and Shimon Sapir. Early diagnosis of parkinson's disease via machine learning on speech data. In *Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of*, pages 1–4. IEEE, 2012.

[29] Peter Garrard, Vassiliki Rentoumi, Benno Gesierich, Bruce Miller, and Maria Luisa Gorno-Tempini. Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. *Cortex*, 55:122–129, 2014.

[30] Shyamal Patel, Konrad Lorincz, Richard Hughes, Nancy Huggins, John Growdon, David Standaert, Metin Akay, Jennifer Dy, Matt Welsh, and Paolo Bonato. Monitoring motor fluctuations in patients with parkinson's disease using wearable sensors. *IEEE transactions on information technology in biomedicine*, 13(6):864–873, 2009.

[31] Thibaud Sénéchal, Jay Turcot, and Rana El Kaliouby. Smile or smirk? automatic detection of spontaneous asymmetric smiles to understand viewer experience. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[32] Lex Fridman, Joonbum Lee, Bryan Reimer, and Trent Victor. 'owl'and 'lizard': patterns of head pose and eye pose in driver gaze classification. *IET Computer Vision*, 10(4):308–314, 2016.

[33] Jaime M Hatcher-Martin, Jamie Lynn Adams, Eric R Anderson, Riley Bove, Tamika M Burrus, Mahan Chehrenama, Mary Dolan O'Brien, Dawn S Eliashiv, Deniz Erten-Lyons, Barbara S Giesser, et al. Telemedicine in neurology: telemedicine work group of the american academy of neurology update. *Neurology*, 94(1):30–38, 2020.

[34] Amy K Guzik and Jeffrey A Switzer. Teleneurology is neurology. *Neurology*, 2020.

[35] Ralph L Sacco. Neurology: Challenges, opportunities, and the way forward. *Neurology*, 93(21):911–918, 2019.

[36] Philip B Gorelick. The global burden of stroke: persistent and disabling. *The Lancet Neurology*, 18(5):417–418, 2019.

[37] Center for Disease Control DHDSP. Stroke facts, sep 2020.

[38] Valery L Feigin, Gregory A Roth, Mohsen Naghavi, Priya Parmar, Rita Krishnamurthi, Sumeet Chugh, George A Mensah, Bo Norrving, Ivy Shiue, Marie Ng, et al. Global burden of stroke and risk factors in 188 countries, during 1990–2013: a systematic analysis for the global burden of disease study 2013. *The Lancet Neurology*, 15(9):913–924, 2016.

[39] The Lancet Neurology. The shared burden of stroke and dementia. *The Lancet Neurology*, 15(9):891, 2016.

[40] Jeffrey L Saver, Gregg C Fonarow, Eric E Smith, Mathew J Reeves, Maria V Grau-Sepulveda, Wenqin Pan, DaiWai M Olson, Adrian F Hernandez, Eric D Peterson, and Lee H Schwamm. Time to treatment with intravenous tissue plasminogen activator and outcome from acute ischemic stroke. *Jama*, 309(23):2480–2488, 2013.

[41] Sunil A Sheth, Reza Jahan, Jan Gralla, Vitor M Pereira, Raul G Nogueira, Elad I Levy, Osama O Zaidat, Jeffrey L Saver, and SWIFT-STAR Trialists. Time to endovascular reperfusion and degree of disability in acute stroke. *Annals of neurology*, 78(4):584–593, 2015.

[42] David S Liebeskind, Reza Jahan, Raul G Nogueira, Tudor G Jovin, Helmi L Lutsep, and Jeffrey L Saver. Early arrival at the emergency department is associated with better collaterals, smaller established infarcts and better clinical outcomes with endovascular stroke therapy: Swift study. *Journal of neurointerventional surgery*, 8(6):553–558, 2016.

[43] Abdul R Abdullah, Eric E Smith, Paul D Biddinger, Deidre Kalenderian, and Lee H Schwamm. Advance hospital notification by ems in acute stroke is associated with shorter door-to-computed tomography time andincreased likelihood of administration of tissue-plasminogen activator. *Prehospital emergency care*, 12(4):426–431, 2008.

[44] James S McKinney, Krishna Mylavarapu, Judith Lane, Virginia Roberts, Pamela Ohman-Strickland, and Mark A Merlin. Hospital prenotification of stroke patients by emergency medical services improves stroke time targets. *Journal of Stroke and Cerebrovascular Diseases*, 22(2):113–118, 2013.

[45] Ian Mosley, Marcus Nicol, Geoffrey Donnan, Ian Patrick, Fergus Kerr, and Helen Dewey. The impact of ambulance practice on acute stroke care. *Stroke*, 38(10):2765–2770, 2007.

[46] David J Gladstone, Lance H Rodan, Demetrios J Sahlas, Liesly Lee, Brian J Murray, Jon E Ween, James R Perry, Jordan Chenkin, Laurie J Morrison, Shann Beck, et al. A citywide prehospital protocol increases access to stroke thrombolysis in toronto. *Stroke*, 40(12):3841–3844, 2009.

[47] Wade S Smith, Marshal Isaacs, and Megan D Corry. Accuracy of paramedic identification of stroke and transient ischemic attack in the field. *Prehospital Emergency Care*, 2(3):170–175, 1998.

[48] Kaiz S Asif, Marc A Lazzaro, and Osama Zaidat. Identifying delays to mechanical thrombectomy for acute stroke: onset to door and door to clot times. *Journal of neurointerventional surgery*, 6(7):505–510, 2014.

[49] M Hansen, SH Sindrup, PB Christensen, NK Olsen, O Kristensen, and ML Friis. Interobserver variation in the evaluation of neurological signs: observer dependent factors. *Acta neurologica scandinavica*, 90(3):145–149, 1994.

[50] A Mohd Nor, C McAllister, SJ Louw, AG Dyker, M Davis, D Jenkinson, and GA Ford. Agreement between ambulance paramedic-and physician-recorded neurological signs with face arm speech test (fast) in acute stroke patients. *Stroke*, 35(6):1355–1359, 2004.

[51] J Adam Oostema, John Konen, Todd Chassee, Mojdeh Nasiri, and Mathew J Reeves. Clinical predictors of accurate prehospital stroke recognition. *Stroke*, 46(6):1513–1517, 2015.

[52] Craig W Brown and Mary J Macleod. The positive predictive value of an ambulance prealert for stroke and transient ischaemic attack. *European Journal of Emergency Medicine*, 25(6):411–415, 2018.

[53] Allison E Arch, David C Weisman, Steven Coca, Karin V Nystrom, Charles R Wira III, and Joseph L Schindler. Missed ischemic stroke diagnosis in the emergency department by emergency medicine and neurology services. *Stroke*, 47(3):668–673, 2016.

[54] Noreen Kamal, Shubin Sheng, Ying Xian, Roland Matsouaka, Michael D Hill, Deepak L Bhatt, Jeffrey L Saver, Mathew J Reeves, Gregg C Fonarow, Lee H Schwamm, et al. Delays in door-to-needle times and their impact on treatment time and outcomes in get with the guidelines-stroke. *Stroke*, 48(4):946–954, 2017.

[55] Sergio Gonzales, Michael T Mullen, Lesli Skolarus, Dylan P Thibault, Uduak Udoeyo, and Allison W Willis. Progressive rural–urban disparity in acute stroke care. *Neurology*, 88(5):441–448, 2017.

[56] Thanh N Nguyen, Mohamad Abdalkader, Tudor G Jovin, Raul G Nogueira, Ashutosh P Jadhav, Diogo C Haussen, Ameer E Hassan, Roberta Novakovic, Sunil A Sheth, Santiago Ortega-Gutierrez, et al. Mechanical thrombectomy in the era of the covid-19 pandemic: emergency preparedness for neuroscience teams: a guidance statement from the society of vascular and interventional neurology. *Stroke*, 51(6):1896–1901, 2020.

[57] Thomas J Oxley, J Mocco, Shahram Majidi, Christopher P Kellner, Hazem Shoirah, I Paul Singh, Reade A De Leacy, Tomoyoshi Shigematsu, Travis R Ladner, Kurt A Yaeger, et al. Large-vessel stroke as a presenting feature of covid-19 in the young. *New England Journal of Medicine*, 382(20):e60, 2020.

[58] Jason M Lippman, Sherita N Chapman Smith, Timothy L McMurry, Zachary G Sutton, Brian S Gunnell, Jack Cote, Debra G Perina, David C Cattell-Gordon, Karen S Rheuban, and Nina J Solenski. Mobile telestroke during ambulance transport is feasible in a rural ems setting: the itreat study. *Telemedicine and e-Health*, 22(6):507–513, 2016.

[59] Sherita N Chapman Smith, Prasanthi Govindarajan, Matthew M Padrick, Jason M Lippman, Timothy L McMurry, Brian L Resler, Kevin Keenan, Brian S Gunnell, Prachi Mehndiratta, and Christina Y Chee. A low-cost, tablet-based option for prehospital neurologic assessment the itreat study. *Neurology*, 87(1):19–26, 2016.

[60] Tzu-Ching Wu, Stephanie A Parker, Amanda Jagolino, Jose-Miguel Yamal, Ritvij Bowry, Abraham Thomas, Amy Yu, and James C Grotta. Telemedicine can replace the neurologist on a mobile stroke unit. *Stroke*, 48(2):493–496, 2017.

[61] Enrique C Leira, Brian Kaskie, Michael T Froehler, and Harold P Adams Jr. The growing shortage of vascular neurologists in the era of health reform: planning is brain! *Stroke*, 44(3):822–827, 2013.

[62] Timothy M Dall, Michael V Storm, Ritashree Chakrabarti, Oksana Drogan, Christopher M Keran, Peter D Donofrio, Victor W Henderson, Henry J Kaminski, James C Stevens, and Thomas R Vidic. Supply and demand analysis of the current and future us neurology workforce. *Neurology*, 81(5):470–478, 2013.

[63] Donna C Bergen. Training and distribution of neurologists worldwide. *Journal of the neurological sciences*, 198(1):3–7, 2002.

[64] Nancy K Glober, Karl A Sporer, Kama Z Guluma, John P Serra, Joe A Barger, John F Brown, Gregory H Gilbert, Kristi L Koenig, Eric M Rudnick, and Angelo A Salvucci. Acute stroke: current evidence-based recommendations for prehospital care. *Western Journal of Emergency Medicine*, 17(2):104, 2016.

[65] Rashmi U Kothari, Arthur Pancioli, Tiepu Liu, Thomas Brott, and Joseph Broderick. Cincinnati prehospital stroke scale: reproducibility and validity. *Annals of emergency medicine*, 33(4):373–378, 1999.

[66] Ethan S Brandler, Mohit Sharma, Richard H Sinert, and Steven R Levine. Prehospital stroke scales in urban environments a systematic review. *Neurology*, 82(24):2241–2249, 2014.

[67] Y. Zhuang, O. Uribe, M. Mcdonald, I. Lin, D. Arteaga, W. Dalrymple, B. Worrall, A. Southerland, and G. Rohde. Pathological facial weakness detection using computational image analysis. In *Biomedical Imaging (ISBI), 2018 IEEE 15th International Symposium on.*

[68] Y. Zhuang et al. Facial weakness analysis and quantification of static images. *IEEE Journal of Biomedical and Health Informatics*, 2020.

[69] Y. Zhuang et al. F-dit-v: An automated video classification tool for facial weakness detection. *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 1–4, 2019.

[70] Yan Zhuang, Mark M McDonald, Chad M Aldridge, Mohamed Abul Hassan, Omar Uribe, Daniel Arteaga, Andrew M Southerland, and Gustavo K Rohde. Video-based facial weakness analysis. *IEEE Transactions on Biomedical Engineering*, 68(9):2698–2705, 2021.

[71] Mohamed Abul Hassan, Xuwang Yin, Yan Zhuang, Chad M Aldridge, Timothy McMurry, Andrew M Southerland, and Gustavo K Rohde. A pilot study on video-based eye movement assessment of the neuroeye examination. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE, 2021.

[72] Mohamed Abul Hassan Ameen, Chad M Aldridge, Yan Zhuang, Xuwang Yin, Timothy McMurry, Gustavo K Rohde, and Andrew M Southerland. Approach to quantify eye movements to augment stroke diagnosis with a non-calibrated eye-tracker. submitted to IEEE TBME.

[73] Andrew Southerland, Mohamed Hassan, Chad Aldridge, Yan Zhuang, Timothy McMurry, and Gustavo Rohde. Comparison of calibration vs non-calibration techniques in the automated capture of eye movement data: Initial validation of the roadie device (4324), 2021.

[74] Yan Zhuang, Mark Mcdonald, Chad Aldridge, Mohamed Abul Hassan, Omar Uribe, Daniel Arteaga, Andrew M Southerland, and Gustavo Rohde. Facial weakness detection demo. `https://youtu.be/4KKaNqFGRwk`, 2021. Online; accessed 29 January 2021.

[75] Yan Zhuang, Shiying Li, Xuwang Yin, Abu Hasnat Mohammad Rubaiyat, Gustavo K Rohde, et al. Local sliced-wasserstein feature sets for illumination-invariant face recognition. *arXiv preprint arXiv:2202.10642*, 2022.

[76] A. Song, G Xu, X. Ding, J. Song, G Xu, and W. Zhang. Assessment for facial nerve paralysis based on facial asymmetry. *Australasian physical & engineering sciences in medicine*, 40(4):851–860, 2017.

[77] A. Gaber, M. Taher, and M. Wahed. Quantifying facial paralysis using the kinect v2. In *Conference proceedings:... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, volume 2015, page 2497, 2015.

[78] H. Kim, S. Kim, Y. Kim, and K. Park. A smartphone-based automatic diagnosis system for facial nerve palsy. *Sensors*, 15(10):26756–26768, 2015.

[79] T. Wang, S. Zhang, J. Dong, L. Liu, and H. Yu. Automatic evaluation of the degree of facial nerve paralysis. *Multimedia Tools and Applications*, 75(19):11893–11908, 2016.

[80] J. Thevenot, M. López, and A. Hadid. A survey on computer vision for assistive medical diagnosis from faces. *IEEE journal of biomedical and health informatics*, 22(5):1497–1511, 2018.

[81] Z. Guo, G. Dan, J. Xiang, J. Wang, W. Yang, H. Ding, O. Deussen, and Y. Zhou. An unobtrusive computerized assessment framework for unilateral peripheral facial paralysis. *IEEE journal of biomedical and health informatics*, 22(3):835–841, 2018.

[82] S. He, J. Soraghan, B. O'Reilly, and D. Xing. Quantitative analysis of facial paralysis using local binary patterns in biomedical videos. *IEEE Transactions on Biomedical Engineering*, 56(7):1864–1870, 2009.

[83] S. He, J. Soraghan, and B. O'Reilly. Biomedical image sequence analysis with application to automatic quantitative assessment of facial paralysis. *Journal on Image and Video Processing*, 2007(3):2, 2007.

[84] P. Li et al. A two-stage method for assessing facial paralysis severity by fusing multiple classifiers. In *Chinese Conference on Biometric Recognition*, pages 231–239. Springer, 2019.

[85] Z. Guo, M. Shen, L. Duan, Y. Zhou, J. Xiang, H. Ding, S. Chen, O. Deussen, and G. Dan. Deep assessment process: Objective assessment process for unilateral peripheral facial paralysis via deep convolutional neural network. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 135–138. IEEE, 2017.

[86] René Handschu, Mateusz Scibor, Martin Nückel, Dirk Asshoff, Barbara Willaczek, Frank Erbguth, Stefan Schwab, and Frank Daumann. Teleneurology in stroke management: costs of service in different organizational models. *Journal of neurology*, 261(10):2003–2008, 2014.

[87] Nolwenn Riou-Comte, Gioia Mione, Lisa Humbertjean, Arielle Brunner, Arnaud Vezain, Karine Lavandier, Sophie Marchal, Serge Bracard, Marc Debouverie, and Sébastien Richard. implementation and evaluation of an economic model for telestroke: experience from virtuall, france. *Frontiers in neurology*, 8:613, 2017.

[88] Jeffrey A Switzer, Bart M Demaerschalk, Jipan Xie, Liangyi Fan, Kathleen F Villa, and Eric Q Wu. Cost-effectiveness of hub-and-spoke telestroke networks for the management of acute ischemic stroke from the hospitals' perspectives. *Circulation: Cardiovascular Quality and Outcomes*, 6(1):18–26, 2013.

[89] R. Almutiry et. al. Facial behaviour analysis in parkinson's disease. In *International Conference on Medical Imaging and Virtual Reality*, pages 329–339. Springer, 2016.

[90] P. Kruszka et al. 22q11. 2 deletion syndrome in diverse populations. *American Journal of Medical Genetics Part A*, 173(4):879–888, 2017.

[91] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[92] H. Yu et. al. A direct lda algorithm for high-dimensional data—with application to face recognition. *Pattern recognition*, 34(10):2067–2070, 2001.

[93] W. Wang et. al. Penalized fisher discriminant analysis and its application to image-based morphometry. *Pattern recognition letters*, 32(15):2128–2135, 2011.

[94] F. Pedregosa et. al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[95] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[96] E. Alabort-i Medina, J.and Antonakos, J. Booth, P. Snape, and S. Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 679–682. ACM, 2014.

[97] S. Zhu, C. Li, C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015.

[98] M. Najibi, P. Samangouei, R. Chellappa, and L. Davis. Ssh: Single stage headless face detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4875–4884, 2017.

[99] J. Yang, Q. Liu, and K. Zhang. Stacked hourglass network for robust facial landmark localisation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 79–87, 2017.

[100] G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou. A comprehensive performance evaluation of deformable face tracking "in-the-wild". *International Journal of Computer Vision*, 126(2-4):198–232, 2018.

[101] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, and S. Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *International Journal of Computer Vision*, 127(6-7):599–624, 2019.

[102] C. Hsu and C. Lin. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425, 2002.

[103] Y Tang. Challenges in representation learning: Facial expression recognition challenge implementation. *University of Toronto*, 2013.

[104] L Vieira. Flat-facial landmarks annotation tool, 2017.

[105] H. Bristow and S. Lucey. Why do linear svms trained on hog features perform so well? *arXiv preprint arXiv:1406.2419*, 2014.

[106] Y. Chang and C. Lin. Feature ranking using linear svm. In *Causation and Prediction Challenge*, pages 53–64, 2008.

[107] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[108] Yan Zhuang, Mark McDonald, Omar Uribe, Xuwang Yin, Dhyey Parikh, Andrew M Southerland, and Gustavo Rohde. Facial weakness analysis and quantification of static images. *IEEE Journal of Biomedical and Health Informatics*, 2020.

[109] D. Haase et al. Automated and objective action coding of facial expressions in patients with acute facial palsy. *European Archives of Oto-Rhino-Laryngology*, 272(5):1259–1267, 2015.

[110] L. Modersohn and J. Denzler. Facial paresis index prediction by exploiting active appearance models for compact discriminative features. In *VISIGRAPP (4: VISAPP)*, pages 271–278, 2016.

[111] P. Xu et al. Automatic evaluation of facial nerve paralysis by dual-path lstm with deep differentiated network. *Neurocomputing*, 2020.

[112] G. Storey et al. 3dpalsynet: a facial palsy grading and motion recognition framework using fully 3d convolutional neural networks. *IEEE access*, 7:121655–121664, 2019.

[113] A. Bandini et al. Automatic detection of amyotrophic lateral sclerosis (als) from video-based analysis of facial movements: speech and non-speech tasks. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 150–157. IEEE, 2018.

[114] M. Alagha et al. Reproducibility of the dynamics of facial expressions in unilateral facial palsy. *International journal of oral and maxillofacial surgery*, 47(2):268–275, 2018.

[115] P. Desrosiers et al. Analyzing of facial paralysis by shape analysis of 3d face sequences. *Image and Vision Computing*, 67:67–88, 2017.

[116] Y. Zhuang, O. Uribe, M. McDonald, X. Yin, D. Parikh, A. Southerland, and G. Rohde. F-dit-v: An automated video classification tool for facial weakness detection. In *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 1–4, May 2019.

[117] J. Bouguet. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel Corporation*, 5(1-10):4, 2001.

[118] H. Abdi and L. Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[119] I. Goodfellow et al. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[120] J. Donahue et al. Long-term recurrent convolutional networks for visual recognition and description. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[121] C. Ma et al. Ts-lstm and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Processing: Image Communication*, 71:76–87, 2019.

[122] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[123] S. Nag et al. Facial micro-expression spotting and recognition using time contrasted feature with visual memory. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2022–2026. IEEE, 2019.

[124] H. Adams et al. Baseline nih stroke scale score strongly predicts outcome after stroke: a report of the trial of org 10172 in acute stroke treatment (toast). *Neurology*, 53(1):126–126, 1999.

[125] T. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[126] S. Raschka. Mlxtend: providing machine learning and data science utilities and extensions to python's scientific computing stack. *Journal of open source software*, 3(24):638, 2018.

[127] N. Maheswaranathan et al. Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. In *Advances in Neural Information Processing Systems*, pages 15696–15705, 2019.

[128] Logitech. Pro webcam c920. `https://www.logitech.com/en-us/products/webcams/c920-pro-hd-webcam.960-000764.html`.

[129] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE transactions on pattern analysis and machine intelligence*, 23(6):643–660, 2001.

[130] Soheil Kolouri, Se Rim Park, and Gustavo K Rohde. The radon cumulative distribution transform and its application to image classification. *IEEE transactions on image processing*, 25(2):920–934, 2015.

[131] Peter N. Belhumeur, João P Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):711–720, 1997.

[132] Yael Adini, Yael Moses, and Shimon Ullman. Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):721–732, 1997.

[133] Zhao-Rong Lai, Dao-Qing Dai, Chuan-Xian Ren, and Ke-Kun Huang. Multiscale logarithm difference edgemaps for face recognition against varying lighting conditions. *IEEE transactions on image processing*, 24(6):1735–1747, 2015.

[134] Weilong Chen, Meng Joo Er, and Shiqian Wu. Illumination compensation and normalization using logarithm and discrete cosine transform. In *ICARCV 2004 8th Control, Automation, Robotics and Vision Conference, 2004.*, volume 1, pages 380–385. IEEE, 2004.

[135] Taiping Zhang, Yuan Yan Tang, Bin Fang, Zhaowei Shang, and Xiaoyu Liu. Face recognition under varying illumination using gradientfaces. *IEEE Transactions on image processing*, 18(11):2599–2606, 2009.

[136] Biao Wang, Weifeng Li, Wenming Yang, and Qingmin Liao. Illumination normalization based on weber's law with application to face recognition. *IEEE Signal Processing Letters*, 18(8):462–465, 2011.

[137] Zakariya Qawaqneh, Arafat Abu Mallouh, and Buket D Barkana. Deep convolutional neural network for age estimation based on vgg-face model. *arXiv preprint arXiv:1709.01664*, 2017.

[138] 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.

[139] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[140] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[141] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[142] Jun-Yong Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Logarithm gradient histogram: A general illumination invariant descriptor for face recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2013.

[143] Ronen Basri and David W Jacobs. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence*, 25(2):218–233, 2003.

[144] Jeffrey Ho and David Kriegman. On the effect of illumination and face recognition. *Face Processing: Advanced Modeling and Methods*, 2005.

[145] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2008.

[146] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[147] P. Felzenszwalb et al. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009.

[148] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):2037–2041, 2006.

[149] Jie Chen, Shiguang Shan, Chu He, Guoying Zhao, Matti Pietikäinen, Xilin Chen, and Wen Gao. Wld: A robust local image descriptor. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1705–1720, 2009.

[150] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133, 2014.

[151] Mohammad Shifat-E-Rabbi, Xuwang Yin, Abu Hasnat Mohammad Rubaiyat, Shiying Li, Soheil Kolouri, Akram Aldroubi, Jonathan M Nichols, Gustavo K Rohde, et al. Radon cumulative distribution transform subspace modeling for image classification. *J Math Imaging Vis*, 63:1185–1203, 2021.

[152] Rafael C Gonzalez, Richard E Woods, et al. Digital image processing, 2002.

[153] Russell Epstein, Peter Hallinan, and Alan Yuille. 5±2 eigenimages suffice: An empirical investigation of low-dimensional lighting models. In *IEEE Workshop on Physics-Based Vision*, pages 108–116, 1995.

[154] Ravi Ramamoorthi. Analytic pca construction for theoretical analysis of lighting variability in images of a lambertian object. *IEEE transactions on pattern analysis and machine intelligence*, 24(10):1322–1333, 2002.

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

[156] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[157] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018.

[158] Se Rim Park, Soheil Kolouri, Shinjini Kundu, and Gustavo K Rohde. The cumulative distribution transform and linear pattern classification. *Applied and computational harmonic analysis*, 45(3):616–641, 2018.

[159] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59, 2017.

[160] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

[161] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[162] Mohammad Shifat-E-Rabbi, Yan Zhuang, Shiying Li, Abu Hasnat Mohammad Rubaiyat, Xuwang Yin, Gustavo K Rohde, et al. Invariance encoding in sliced-wasserstein space for image classification with limited training data. *arXiv preprint arXiv:2201.02980*, 2022.

[163] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[164] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. The cas-peal large-scale chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(1):149–161, 2007.

[165] Aleix Martinez and Robert Benavente. The ar face database: Cvc technical report, 24. Technical report, The Ohio State University, 1998.

[166] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.