AUTOMATIC RECOGNITION OF HANDWRITTEN MEDICAL FORMS FOR SEARCH ENGINES

By

Robert Jay Milewski

September 2006

A DISSERTATION SUBMITTED TO THE

FACULTY OF THE GRADUATE SCHOOL OF THE STATE

UNIVERSITY OF NEW YORK AT BUFFALO

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Committee

Chairman:	Dr. Venugopal Govindaraju Computer Science and Engineering, University at Buffalo
Members:	Dr. William Rapaport Computer Science and Engineering, University at Buffalo
	Dr. Jinhui Xu Computer Science and Engineering, University at Buffalo
	Dr. Liana Lorigo Computer Science and Engineering, University at Buffalo
Outside Reader:	Dr. Daniel Lopresti Computer Science and Engineering, Lehigh University

Acknowledgments

• Center for Unified Biometrics and Sensors (CUBS) has provided important technical discussion and resources.

• National Science Foundation (NSF) has funded this research:
Grant Award Abstract #0429358
Grant Title: Emergency Medicine, Disease Surveillance, and Informatics.

• New York State Department of Health has written letters of support.

• Western Regional Emergency Medical Services (WREMS) has supplied the PCR images necessary for the research.

• Special thanks to Casey Czamara, Michael Gerfin, Julie Milewski, Ranga Setlur, Eugenia Smith, Andrea Kurpiewski, Robert Dennison, Susan Katz and Jeffrey Gerfin for information access, editing, discussion, data entry, and/or scheduling.

Dedications

This research is dedicated to the victims, families and friends of those affected by the September 11, 2001 attacks. In addition, all professionals in the Emergency Medical Services, Hazardous Materials, Fire Department Rescue, and Law Enforcement Units who are always there... as both a former Emergency Medical Technician and citizen of the United States of America, I thank you and hope that this research assists future rescue efforts.

Quotation

Few are those who see with their own eyes and feel with their own hearts.

-Albert Einstein

Abstract

Handwriting recognition (HR) is a challenging problem that is made tractable only by the contextual constraints offered by specific applications. The population of a national emergency medical service database from the collection of the New York State (NYS) Pre-hospital Care Report (PCR) calls for handwriting recognition. Such a database can enable emergency preparedness, response, and homeland security. We address several research challenges presented by the task of reading hand-filled PCR's in particular and medical forms in general. Written text on such forms has poor legibility due to insufficient size of writing areas (e.g. compressed text or text curved along margin), vehicle motion, writing with gloves, and the immediacy of the emergency environment. Challenges include: (i) written matter often spilling beyond the form boundaries, (ii) diverse lexicons in the medical domain, and (iii) low recognition performance due to poor legibility of text. A fourth challenge is that modern search engines expect to operate on known text and not on handwriting. In order to address these issues, we have developed the following: (i) the first text extraction technique which operates on carbon paper, (ii) a lexicon reduction strategy which maps partial recognition information to medical topic categories, and (iii) an information retrieval system capable of searching forms using handwriting recognition results.

While the emphasis of this research is on medical forms, the ideas extend to any domain in which there is at least one sentence of text that can be classified under high level topic categories. In the application of medical forms, it is shown that the words written by health care professionals involved in all aspects of patient assessment can be organized within the context of anatomical positions. Conceivably, if a patient with a broken leg is

rescued, then the handwriting will be related to the identification and rescue efforts involving the anatomical position of *legs*.

The primary issue is how the category is determined if the handwritten words are unknown. Since both the lexicon reduction step, aimed at improving the recognition performance, as well as the search engine require the recognized words, a new paradigm must be developed to solve the problem. The algorithm described in this research automatically learns the salient relationships between characters from correlated words, and maps such related characters to categories. The quantity of initially recognized characters is restricted to two per word since the recognition engines cannot successfully extract all the characters; two is empirically determined as the maximum reasonable quantity of characters recognized with high confidence. This raises the issue of collisions between words that have the same uni-gram or bi-gram. To address this situation, two steps are performed: (i) the distance information between character information is encoded, and (ii) the usage of uni/bi-gram cohesive phrases, instead of independent words, is mapped under the category. At this stage, a list of spatially encoded uni/bi-grams under a category exists. However, the notion of a collision now also extends to the category level. For example, the category *arms* and *legs* may both contain the phrase *blood loss*. To handle such ambiguities, it is necessary to determine which uni/bi-gram phrases most uniquely defines each category. A series of steps is used to extract and weight the most relevant uni/bi-gram phrases (a.k.a. *terms*) against the categories with which they are associated. Given a new form, the characters are extracted, the category is automatically determined, the lexicon is reduced, the handwriting recognition is performed, and query matches are returned. This results in recognition improvements between 4.50%-7.25% after binarization and post-processing, a handwriting recognition improvement of 7.42% with a reduction in error rate of 10.88% and an increase of effective queries by 50%.

The hybridization of handwriting recognition, natural language processing, contextual knowledge representation, and information retrieval is novel. We show that it is possible to automatically determine a high level category and use it for both recognition and retrieval even with several levels of ambiguity and recognition errors. Medical forms were chosen due to the high level of complexity inherent in a large heterogeneous corpus of medical, pharmacological and English texts. Forms were written by multiple writers in complex emergency environments. The emergencies reported involve such situations as extreme temperature and weather, fire, vehicle accidents, ambulance movement, and hazardous materials. By addressing these problems in the most extreme circumstances, we were able to gain information in two areas: (i) algorithm effectiveness in the worst case scenario, and (ii) insight into human cognitive interpretations.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Research in recognition of medical forms builds on the knowledge acquired in several branches of artificial intelligence during the last half century: handwriting recognition, information tetrieval, image processing, forms processing, natural language processing, and computational linguistics. Medical forms are used for several reasons: (i) No pattern recognition group has previously attempted the recognition of handwriting on medical forms (to the best of our knowledge), (ii) the nature of these forms presents complex PR problems, (iii) medical printed text has a history of successful research involving semantic analysis, and (iv) there is high demand for epidemiological and health surveillance data.

This research analyzes handwriting in the context of medical forms. There are two goals of this research. *Recognition* is the task of using an image containing handwriting as an input and producing the ASCII conversion as output. The *retrieval* component involves the search of medical forms based on a human query. Each goal can be broken down into several component sub-goals, each with its own history of algorithmic approaches, capabilities, and limitations.

This research operates under several assumptions: (i) the locations of words on the medical form have been previously segmented, (ii) at least 50% of the words on a form are readable by a person in order to be considered in any training or test deck, (iii) layout of the forms is known and consistent, and (iv) at least a single sentence of handwritten text exists on the form.

1.1 Motivation

The National Strategy for Homeland Security released by the White House in July 2002 lists *Emergency Preparedness and Response* as one of the six critical mission areas for Homeland Security [131]. Despite the publicity surrounding the issue of preparedness to guard against possible bio-terrorist attacks, there have not been many short to middle-term practical technological solutions proposed to address the problem. There is a strong demand in the New York State Department of Health to greatly improve the speed of collecting NYS Pre-Hospital Care Reports (PCR) [135] data to populate a national Emergency Medical Service (EMS) database, thereby providing emergency medical service providers and health care administrators with a wealth of data that can be used in epidemiological analysis, counter bio-terrorism and Mass Casualty Incidents. These forms, which are a few years behind in storage and analysis, are also used for legal documentation and EMS quality assurance.

Seven years of health care experience in the capacities of NYS Certified First Responder and NYS Emergency Medical Technician, specialization in the Artificial Intelligence track of the Computer Science program, and work with the Center of Excellence for Document Analysis and Recognition (CEDAR) led the author to initiate and carry out this research. The intent is to explore algorithms for the recognition and search of handwritten medical documents, toward the larger goal of a real-time searchable database for improved knowledge, treatment and rescue efforts in the medical field.

1.2 Contributions

This dissertation makes the following contributions:

• The first application of recognition of handwritten medical forms.

• The first search engine using handwritten forms.

• The first binarization and post-processing strategy on carbon forms.

• The first binarization algorithm using sinusoidal waves.

• A paradigm showing a mapping between character encodings to a topic categorization used for lexicon reduction. This strategy is reusable for other lexicon driven handwriting recognizers that are based on character segmentation.

• New metrics for measuring the performance of lexicon reduction systems.

• Construction of the first data set of actual handwritten emergency medical documents for use in document analysis research.

• Compatibility with standard information protocols used by Health Level 7 (HL7) [55] and the Center for Disease Control (CDC) [23].

• A framework for an automated, centralized, and secure health surveillance network.

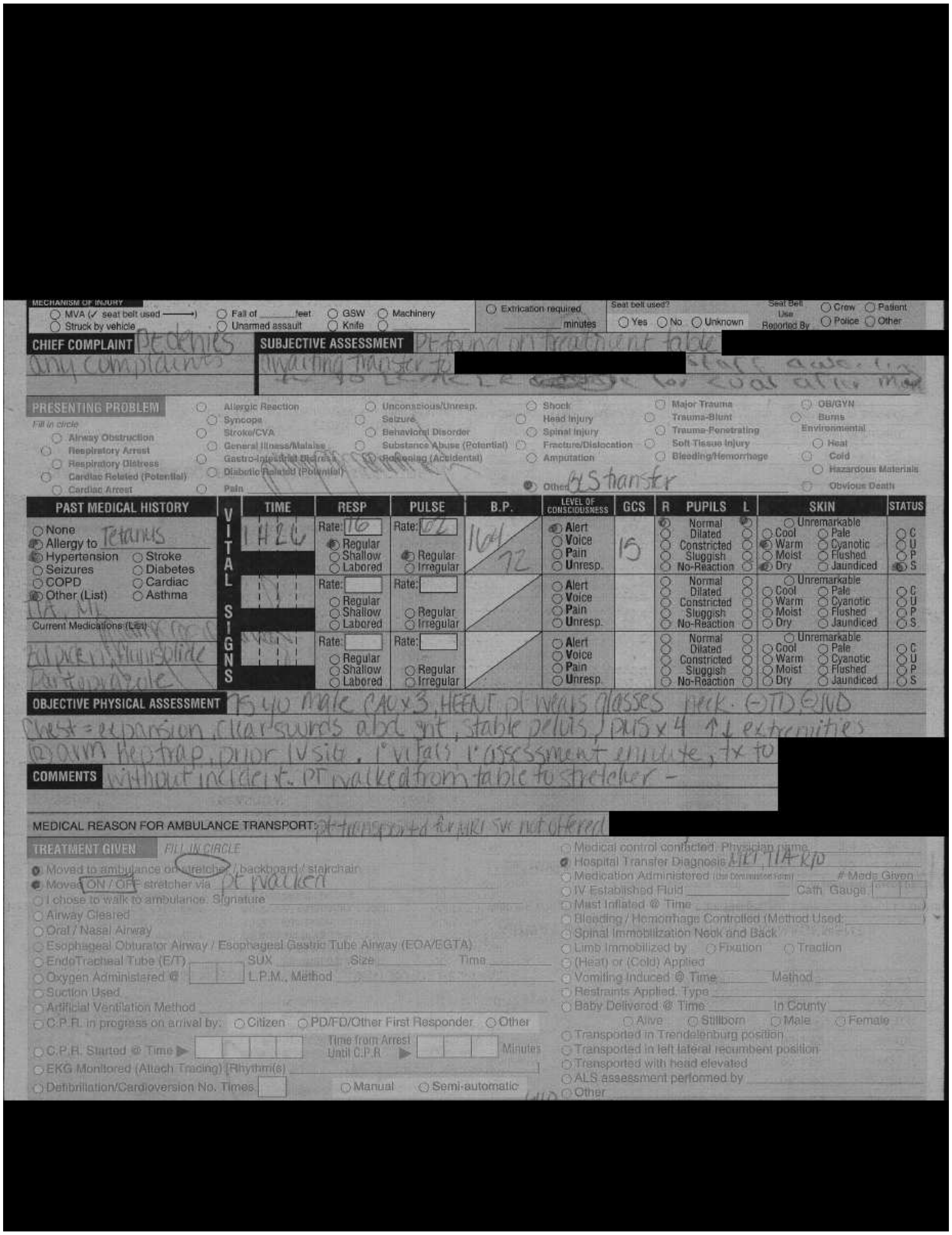
• An advanced software system with diverse visual interfaces and command-line execution modes.

1.3 PCR Background

In the United States, any pre-hospital emergency medical care provided has to be rigorously documented. Departments of Health of each state prescribe a standard medical form to be used in documenting all information on the patient's status and treatment from the moment the rescue effort begins until he or she is transported to the hospital. State laws require emergency personnel to completely fill out this form for each patient prior to admission into the hospital for care.

Data for this research, in the form of actual research copies of the PCR forms [135] (see Figure 1.1), have been obtained under an agreement with the Western Regional Emergency Medical Services (WREMS) [135] division of the New York State (NYS) Department of Health. Each PCR is stored as a 300 DPI (dots/inch) color image. Computations are only performed on zones containing the relevant medical information; more specifically, no computations are performed on zones containing patient identifying information or on PCRs involving patients with behavioral disorders. The PCR is a form used to gather vital patient information that is used by health care administrators as a resource to identify trends through macro-analysis. Currently, PCRs are mostly paper forms, and the process of keying this data into a database that can be processed and mined for trend information can take up to several years in many states. A nationwide database of PCR data would be invaluable for a public health syndromic surveillance system.

There are five major zones on the PCR containing the handwritten information of interest (ordered from top to bottom): Chief Complaint (Figure 1.2 Location 8), Subject Assessment (Figure 1.2 Location 9), Past Medical History (Figure 1.2 Location 11), Objective Physical Assessment (Figure 1.2 Location 13), Comments (Figure 1.2 Location 14). These handwritten areas contain numbers (e.g. 84), symbols (e.g. $\uparrow$ = increase), abbreviations (see Appendix C for examples), anatomical descriptions (e.g. thoracic), medical


[redacted]

MECHANISM OF INJURY
MVA (✓ seat belt used) Struck by vehicle Fall of ___ feet Unarmed assault GSW Knife Machinery
Extrication required ___ minutes
Seat belt used? Yes No Unknown
Seat Belt Use Reported By Crew Patient Police Other

CHIEF COMPLAINT Pt denies any complaints
SUBJECTIVE ASSESSMENT Pt found on treatment table [redacted] awaiting transfer to [redacted]

PRESENTING PROBLEM
Fill in circle
Airway Obstruction, Respiratory Arrest, Respiratory Distress, Cardiac Related (Potential), Cardiac Arrest, Allergic Reaction, Syncope, Stroke/CVA, General Illness/Malaise, Gastro-Intestinal Distress, Diabetic Related (Potential), Pain, Unconscious/Unresp., Seizure, Behavioral Disorder, Substance Abuse (Potential), Poisoning (Accidental), Shock, Head Injury, Spinal Injury, Fracture/Dislocation, Amputation, Other BLS transfer, Major Trauma, Trauma-Blunt, Trauma-Penetrating, Soft Tissue Injury, Bleeding/Hemorrhage, OB/GYN, Burns, Environmental, Heat, Cold, Hazardous Materials, Obvious Death

PAST MEDICAL HISTORY
None
Allergy to Tetanus
Hypertension, Stroke, Seizures, Diabetes, COPD, Cardiac, Other (List), Asthma
TIA, MI
Current Medications (List)

VITAL SIGNS
TIME	RESP	PULSE	B.P.	LEVEL OF CONSCIOUSNESS	GCS	PUPILS	SKIN	STATUS
1426	Rate: 16 Regular	Rate: 102 Regular	164/72	Alert	15	Normal	Warm, Dry	S

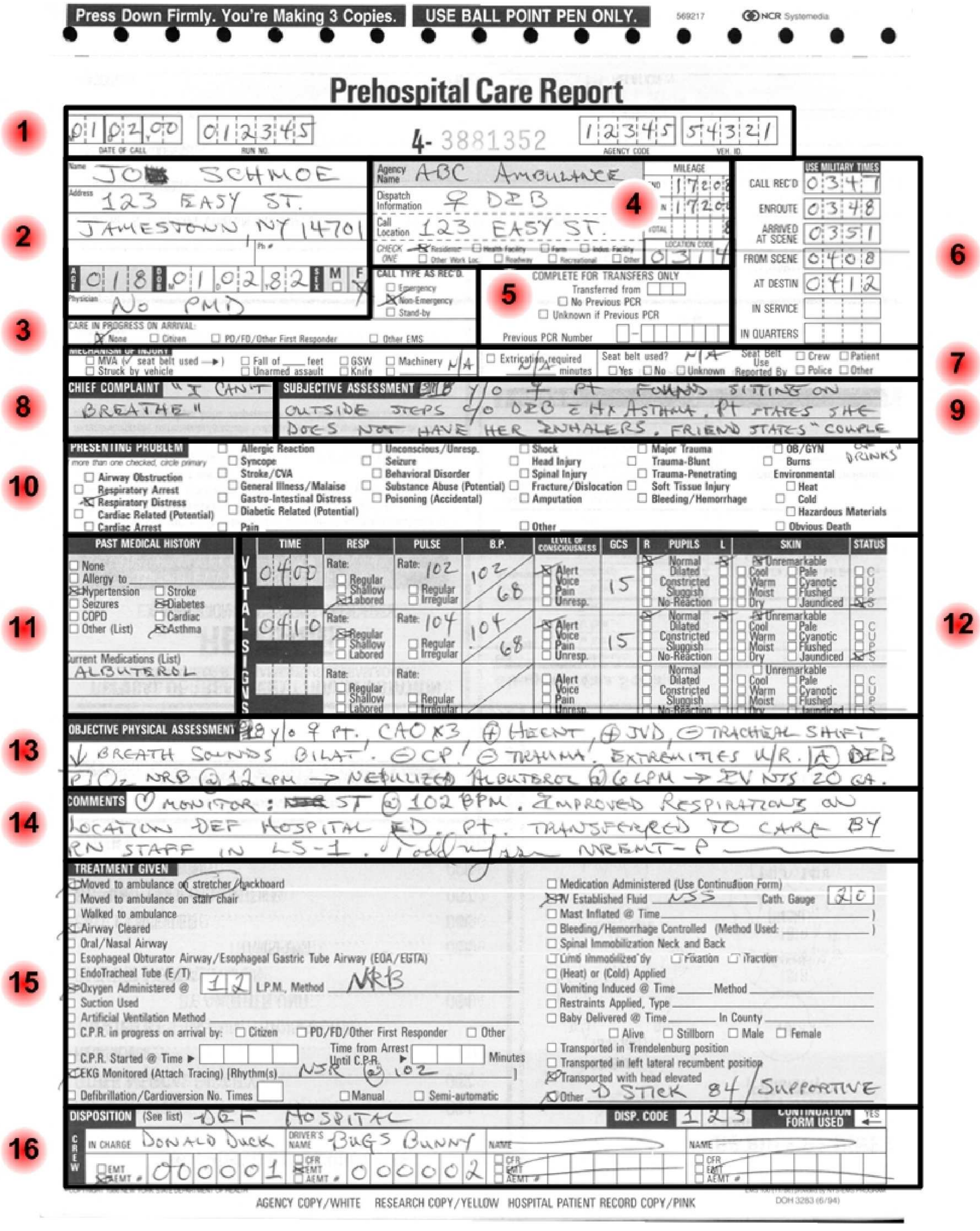
OBJECTIVE PHYSICAL ASSESSMENT 75 yo male CAO x 3, HEENT pt wears glasses, neck (-) TD, (-) JVD
Chest = expansion, clear sounds abd. snt, stable pelvis, PMS x 4 ↑↓ extremities
(R) arm hep trap, prior lvsig. ↑ vitals ↑ assessment enroute, tx to [redacted]

COMMENTS without incident. Pt walked from table to stretcher –

MEDICAL REASON FOR AMBULANCE TRANSPORT: pt transported for MRI svc not offered [redacted]

TREATMENT GIVEN FILL IN CIRCLE
Moved to ambulance on stretcher / backboard / stairchair
Moved ON / OFF stretcher via pt walked
I chose to walk to ambulance. Signature
Airway Cleared
Oral / Nasal Airway
Esophageal Obturator Airway / Esophageal Gastric Tube Airway (EOA/EGTA)
EndoTracheal Tube (E/T) SUX Size Time
Oxygen Administered @ L.P.M., Method
Suction Used
Artificial Ventilation Method
C.P.R. in progress on arrival by: Citizen PD/FD/Other First Responder Other
C.P.R. Started @ Time Time from Arrest Until C.P.R. Minutes
EKG Monitored (Attach Tracing) [Rhythm(s)]
Defibrillation/Cardioversion No. Times Manual Semi-automatic
Medical control contacted. Physician name
Hospital Transfer Diagnosis MRI TIA R/O
Medication Administered # Meds Given
IV Established Fluid Cath. Gauge
Mast Inflated @ Time
Bleeding / Hemorrhage Controlled (Method Used:
Spinal Immobilization Neck and Back
Limb Immobilized by Fixation Traction
(Heat) or (Cold) Applied
Vomiting Induced @ Time Method
Restraints Applied, Type
Baby Delivered @ Time In County Alive Stillborn Male Female
Transported in Trendelenburg position
Transported in left lateral recumbent position
Transported with head elevated
ALS assessment performed by
Other
[redacted]

Figure 1.1: Pre-Hospital Care Report (PCR) example (identifying information hidden)



Press Down Firmly. You're Making 3 Copies. USE BALL POINT PEN ONLY. 569217 NCR Systemedia

Prehospital Care Report

1 — DATE OF CALL: 01 02 00 — RUN NO.: 012345 — 4-3881352 — AGENCY CODE: 12345 — VEH. ID: 54321

2 — Name: JO[redacted] SCHMOE — Address: 123 EASY ST. JAMESTOWN, NY 14701 — Ph #

AGE: 018 — DOB: 01 02 82 — SEX: M F [X] — Physician: NO PMD

4 — Agency Name: ABC AMBULANCE — MILEAGE: END 17208, BEGIN 17200, TOTAL — Dispatch Information: ♀ DEB — Call Location: 123 EASY ST. — CHECK ONE: [X] Residence ☐ Health Facility ☐ Farm ☐ Indus. Facility ☐ Other Work Loc. ☐ Roadway ☐ Recreational ☐ Other — LOCATION CODE: 0314

6 — USE MILITARY TIMES: CALL REC'D 0347 — ENROUTE 0348 — ARRIVED AT SCENE 0351 — FROM SCENE 0408 — AT DESTIN 0412 — IN SERVICE — IN QUARTERS

CALL TYPE AS REC'D: ☐ Emergency [X] Non-Emergency ☐ Stand-by

5 — COMPLETE FOR TRANSFERS ONLY: Transferred from ☐ No Previous PCR ☐ Unknown if Previous PCR — Previous PCR Number

3 — CARE IN PROGRESS ON ARRIVAL: [X] None ☐ Citizen ☐ PD/FD/Other First Responder ☐ Other EMS

7 — MECHANISM OF INJURY: ☐ MVA (✓ seat belt used →) ☐ Struck by vehicle ☐ Fall of ___ feet ☐ Unarmed assault ☐ GSW ☐ Knife ☐ Machinery N/A ☐ Extrication required N/A minutes — Seat belt used? N/A ☐ Yes ☐ No ☐ Unknown — Seat Belt Use Reported By ☐ Crew ☐ Patient ☐ Police ☐ Other

8 — CHIEF COMPLAINT: "I CAN'T BREATHE"

9 — SUBJECTIVE ASSESSMENT: 18 Y/O ♀ PT FOUND SITTING ON OUTSIDE STEPS C/O DIB ē Hx ASTHMA. PT STATES SHE DOES NOT HAVE HER INHALERS. FRIEND STATES "COUPLE DRINKS"

10 — PRESENTING PROBLEM (more than one checked, circle primary): ☐ Airway Obstruction ☐ Respiratory Arrest [X] Respiratory Distress ☐ Cardiac Related (Potential) ☐ Cardiac Arrest ☐ Allergic Reaction ☐ Syncope ☐ Stroke/CVA ☐ General Illness/Malaise ☐ Gastro-Intestinal Distress ☐ Diabetic Related (Potential) ☐ Pain ☐ Unconscious/Unresp. ☐ Seizure ☐ Behavioral Disorder ☐ Substance Abuse (Potential) ☐ Poisoning (Accidental) ☐ Shock ☐ Head Injury ☐ Spinal Injury ☐ Fracture/Dislocation ☐ Amputation ☐ Other ☐ Major Trauma ☐ Trauma-Blunt ☐ Trauma-Penetrating ☐ Soft Tissue Injury ☐ Bleeding/Hemorrhage ☐ OB/GYN ☐ Burns Environmental ☐ Heat ☐ Cold ☐ Hazardous Materials ☐ Obvious Death

11 — PAST MEDICAL HISTORY: ☐ None ☐ Allergy to [X] Hypertension ☐ Stroke ☐ Seizures [X] Diabetes ☐ COPD ☐ Cardiac ☐ Other (List) [X] Asthma — Current Medications (List): ALBUTEROL

12 — VITAL SIGNS

TIME	RESP	PULSE	B.P.	LEVEL OF CONSCIOUSNESS	GCS	PUPILS	SKIN	STATUS
0400	Rate: [X] Labored	Rate: 102 Regular	102/68	[X] Alert	15	Normal	[X] Unremarkable	S
0410	Rate: [X] Regular	Rate: 104	104/68	[X] Alert	15	Normal	[X] Unremarkable	S
	Rate:	Rate:						

13 — OBJECTIVE PHYSICAL ASSESSMENT: 18 Y/O ♀ PT. CAO X3, ⊕ HEENT, ⊖ JVD, ⊖ TRACHEAL SHIFT. ↓ BREATH SOUNDS BILAT. ⊖ CP! ⊖ TRAUMA. EXTREMITIES W/R. [A] DEB PT O2 NRB @ 12 LPM → NEBULIZED ALBUTEROL @ 6 LPM → IV NTS 20 GA.

14 — COMMENTS: ♡ MONITOR: ~~NSR~~ ST @ 102 BPM. IMPROVED RESPIRATIONS ON LOCATION DEF HOSPITAL ED. PT. TRANSFERRED TO CARE BY RN STAFF IN LS-1. [signature] NREMT-P

15 — TREATMENT GIVEN: [X] Moved to ambulance on stretcher/backboard ☐ Moved to ambulance on stair chair ☐ Walked to ambulance [X] Airway Cleared ☐ Oral/Nasal Airway ☐ Esophageal Obturator Airway/Esophageal Gastric Tube Airway (EOA/EGTA) ☐ EndoTracheal Tube (E/T) [X] Oxygen Administered @ 12 L.P.M., Method NRB ☐ Suction Used ☐ Artificial Ventilation Method ☐ C.P.R. in progress on arrival by: ☐ Citizen ☐ PD/FD/Other First Responder ☐ Other ☐ C.P.R. Started @ Time ▶ Time from Arrest Until C.P.R. ▶ Minutes [X] EKG Monitored (Attach Tracing) [Rhythm(s) NSR @ 102] ☐ Defibrillation/Cardioversion No. Times ☐ Manual ☐ Semi-automatic ☐ Medication Administered (Use Continuation Form) [X] IV Established Fluid NSS Cath. Gauge 20 ☐ Mast Inflated @ Time ☐ Bleeding/Hemorrhage Controlled (Method Used:) ☐ Spinal Immobilization Neck and Back ☐ Limb Immobilized by ☐ Fixation ☐ Traction ☐ (Heat) or (Cold) Applied ☐ Vomiting Induced @ Time ___ Method ☐ Restraints Applied, Type ☐ Baby Delivered @ Time ___ In County ☐ Alive ☐ Stillborn ☐ Male ☐ Female ☐ Transported in Trendelenburg position ☐ Transported in left lateral recumbent position [X] Transported with head elevated [X] Other D STICK 84 / SUPPORTIVE

16 — DISPOSITION (See list): DEF HOSPITAL — DISP. CODE: 123 — CONTINUATION FORM USED: YES

CREW: IN CHARGE DONALD DUCK — EMT # 000001 — DRIVER'S NAME BUGS BUNNY — EMT # 000002 — NAME — NAME

AGENCY COPY/WHITE RESEARCH COPY/YELLOW HOSPITAL PATIENT RECORD COPY/PINK DOH 3283 (6/94)

Figure 1.2: Pre-Hospital Care Report (PCR) labeled (simulated information)

conditions (e.g. pneumothorax), pharmacological words (e.g. codeine) and common English. The handwritten zones can contain data from a large heterogeneous lexicon, and the text often does not fit perfectly within form boundaries. The text zones, therefore, present a highly challenging recognition problem. The rest of the form consists of elements such as check-boxes, segmented character locations, and segmented digit-only locations. The recognition of such elements are less challenging than handwriting recognition. However, elements are not as complete or as verbose as handwriting.

The form available for processing and data mining is a carbon copy. Figure 1.2 is a top copy which is held by the hospital and is not available for research. A top copy is the most cleanly written since the data loss issues are not present (as with the carbon copies). However, the carbon copies are still used in the medical system and are still backlogged. The carbon mesh residue in various locations on the form, and broken/unnatural handwriting due to ambulance movement and emergency environments add further complexity to the document. There is also an extension form that allows health-care providers to continue writing if there is no room in the Comments PCR region. The use of this form is, however, rare.

While other work has been performed in the area of unconstrained handwriting, it has been limited to a large lexicon of only English [130]. There has also been little prior work on poor handwriting [18], especially in the environment of this carbon paper and in the emergency environment. Each section of this research will discuss these issues in further detail.

1.4 Taxonomy

In this section, the most commonly encountered handwriting styles are listed. Any combination of these styles can be found in the emergency environment. This provides some

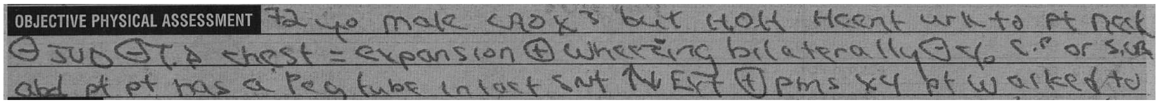


Figure 1.3: Good handwriting pressure

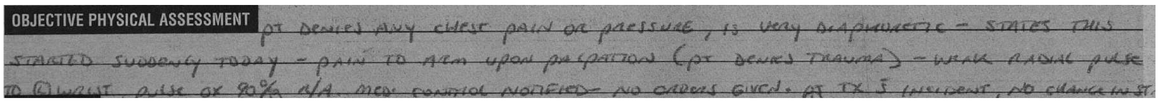


Figure 1.4: Small print

visual insight into the challenges of medical handwriting recognition. The diversity of characters and pen-to-paper handwriting pressure lead to recognition errors (see Figures 1.3 - 1.14). Note that PCRs containing an additional "continuation form," which may be attached in rare occurences, are omitted from this research. PCRs are sometimes accompanied by a printed electrocardiogram sample. The integration of such information is outside the scope of this research.

1.5 PCR Training and Test Deck Construction

Medical form training and test sets have been created manually. A software data entry system has been developed that allows humans, known as truthers, to manually segment all PCR form zones and words, and to provide a human interpretation for the word, denoted as the truth. The use of *truth* indicates that the human classification of text is always the correct interpretation.

The process of data entry, known as truthing, has two phases: (i) the digital transcription of medical form text, and (ii) the classification of forms into topic categories. The

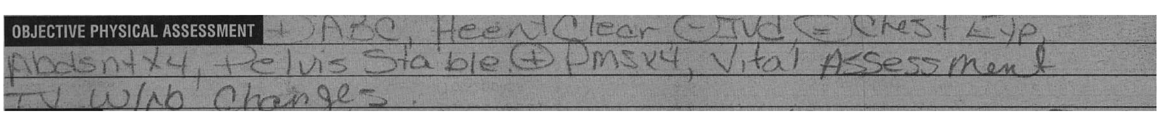


Figure 1.5: Full line height

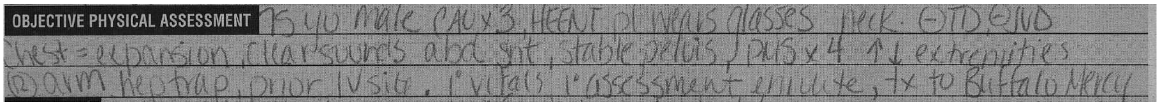


Figure 1.6: Narrow width

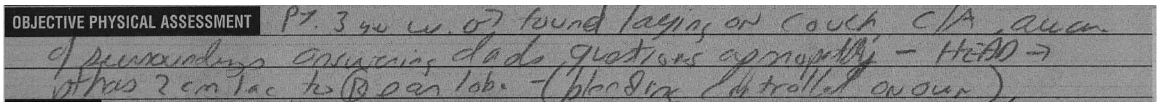


Figure 1.7: Mixed print and cursive types

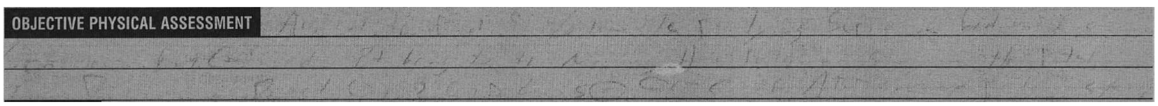


Figure 1.8: No handwriting pressure on carbon copy

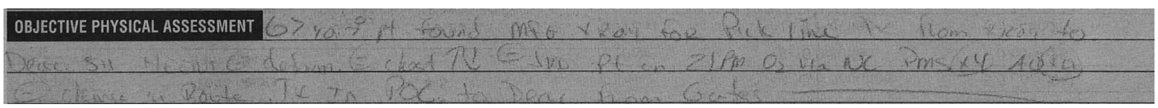


Figure 1.9: Reduced/Degraded pressure on carbon copy

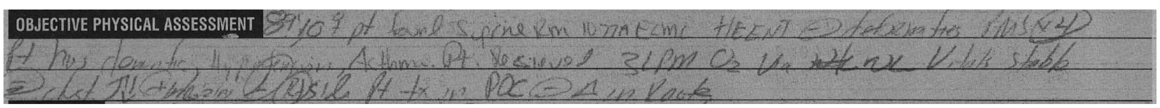


Figure 1.10: Mixed pressure censitivity print type

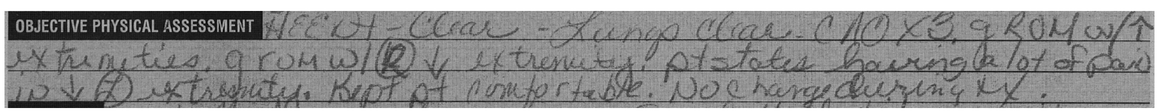


Figure 1.11: Mixed pressure cursive

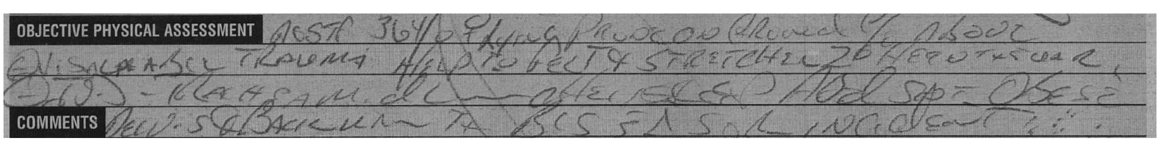


Figure 1.12: Linear line violation

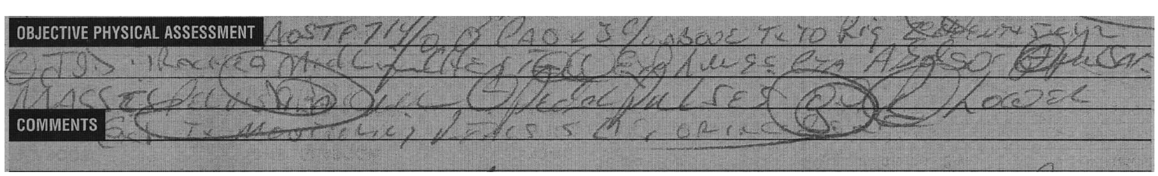


Figure 1.13: Nonlinear line violation

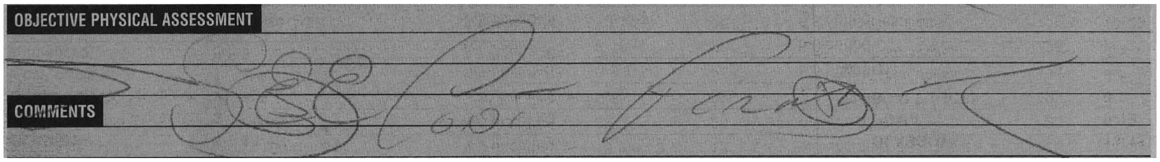


Figure 1.14: Reference to another document such as a continuation form

distribution of PCR forms under each category is approximately equal in both the training and test decks. The task has been supervised and performed by a health care professional with seven years of field emergency medical services (EMS) experience. This corpus of handwritten emergency medical forms is the first of its kind.

1.6 Outline of Dissertation

Figure 1.15 graphically illustrates the layout of this dissertation. Chapter 1 introduces the reader with the problem statement, motivation, and introduces the medical forms that are used in this research. Chapter 2 then discusses prior work in form recognition, handwriting recognition, the need for contextual models in handwriting analysis, information retrieval, latent semantic analysis and some cognitive insight into the approach of this work. Chapter 3 identifies all related pre-processing steps necessary for the handwriting recognition tasks. Visual comparisons between binarization and post-processing algorithms are described. Note that automatic word segmentation is not addressed in this work. Chapter 4 defines metrics used in the evaluation of lexicon reduction algorithms. Chapter 5 begins from the pre-processed image and shows an approach for the automated recognition of the handwritten words on the form. Once an ASCII version of the medical form is available, then it is open to text processing algorithms. Chapter 6 compares several handwriting recognition experiments and evaluates the performance of the lexicon reduction algorithm defined in Chapter 5. Chapter 7 compares the retrieval effectiveness on medical forms before and after the use of the lexicon reduction algorithm. Chapter 8 describes several practical applications that can utilize this research. Chapter 9 describes the software built to manage truthing and algorithm experiments. Chapter 10 concludes with specifics insights of this work. The appendices include ethics and security, a competing system, and a parallel processing architectural requirement.

The objective is to provide techniques for the handwriting recognition and retrieval of medical forms. This allows health surveillance and epidemiological software to have an entirely new resource of medical information. Prior work not only has its own dedicated chapter, but is included whenever the algorithm in question requires more detail to understand.

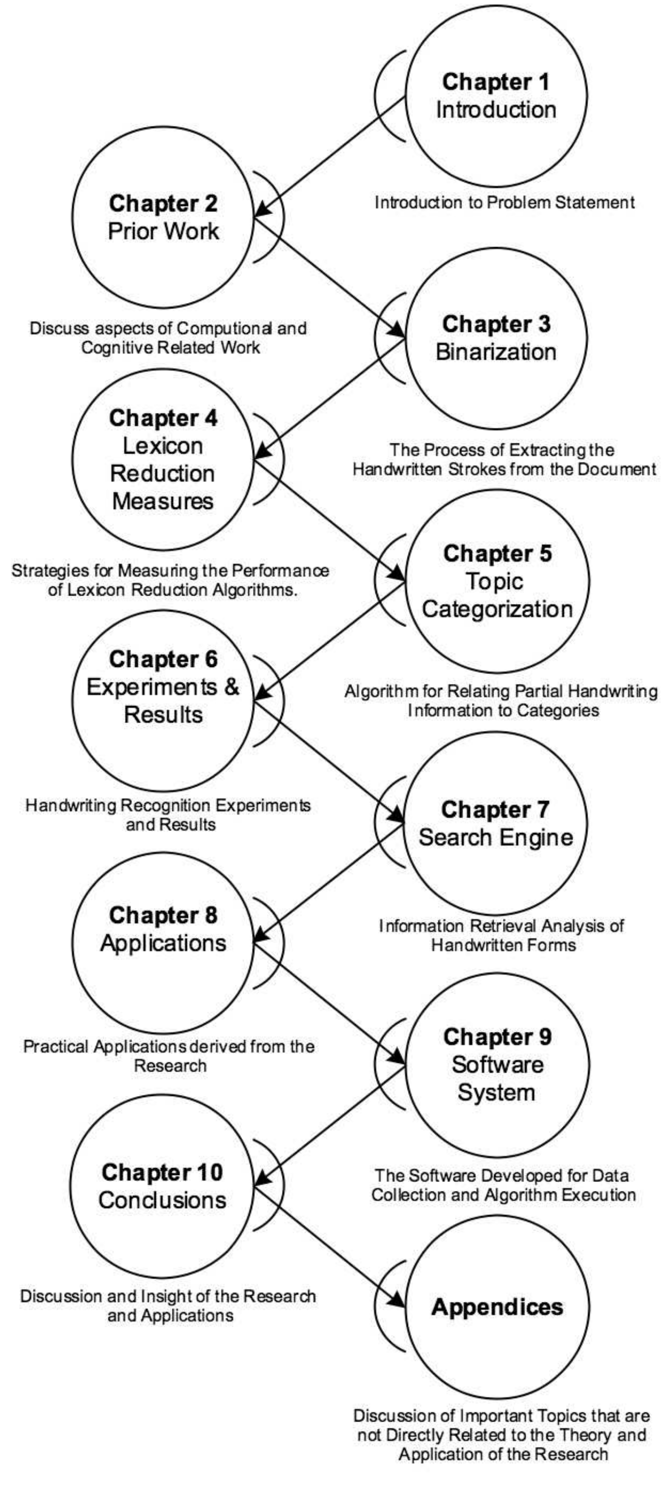


Figure 1.15: Dissertation Layout

Chapter 2

Prior Work

Prior work is found in three locations of this dissertation: (i) this chapter illustrates the prior work in handwriting recognition, lexicon reduction, forms processing, information retrieval systems, human cognition and ontology, (ii) chapter 3 contains prior work relevant to forms processing techniques, and (iii) chapter 5 discusses prior work relevant to lexicon reduction techniques.

2.1 Background

Although handwriting recognition and lexicon reduction [81] have been researched substantially over the years, many challenges still persist in the offline domain. Word recognition applications range from automated check recognition [65], postal recognition [38], historical documents recognition [39] [46], and now emergency medical documents [85] [86] [87]. Strategic recognition techniques for handwriting algorithms such as hidden markov models (HMM) [24] [57] [67] [82] [91], artificial neural networks (ANN) [15] [26] [27] [40] [96], and support vector machines (SVM) [7] [17] have been developed. Lexicon reduction, any process aiming to eliminate irrelevant entries from a lexicon, has been shown to be critical to improvement of performance primarily because of the minimization of possible choices [47]. Systems reporting high recognition even with a large vocabulary corpus have also been successful [67] [68].

Additionally, other lexicon reduction strategies have used the extraction of character information for lexicon reduction, such as that by Guillevic, et al. [50], which uses an HMM. However, that research reduces the lexicon from a single category, namely cities in Finland. In addition, usage of word length estimates for a smaller lexicon are available, and the binarization [1] appears significantly cleaner [50]. Caesar, et al. [18] also states that prior reduction techniques [97] [116] [120] are unsuitable since they can only operate on very small lexicons due to enormous computational burdens [18]. Caesar [18] further indicates that Suen's [123] approach of n-gram combinatorics is sensitive to segmentation issues, a common problem with medical form handwriting [18]. However, Caesar's method [18] and those that are dependent on using the character information, or the character information of only one word to directly reduce the lexicon, suffers if one of the characters is selected incorrectly [18]. This is observable in the cursive or mixed-cursive handwriting types.

Many existing schemes, such as that of Zimmermann [143], assume that acceptable characters can be extracted. However, in the medical handwriting domain, there are very high error rates. Therefore, operating a reduction scheme that can be robust to incorrectly chosen characters is necessary. As a result, this research has moved in the direction of an alternate organization, namely, sequences of characters are used to determine a category that has a lexicon of its own, thereby reducing the issues of using the character information directly. Similar to the study by Zimmermann et al. [143], here the length encodings of words are involved with the terms. However, a term in this case is a phrase rather than an individual word, and the use of wildcards [2] is found to increase run-time and degrade performance. In addition, the approach of Zimmermann et al. [143] provided an optimization,

[1] A process for extracting foreground handwritten stroke pixels from the document background.
[2] A regular expression pattern using tokens, such as *, to match alpha-numeric text.

whereas our research also shows recognition improvements.

Kaufmann, et al. [64] present another HMM strategy, which is primarily a distance-based method and uses model assumptions that are not applicable in the medical environment. For example, Kaufmann [64] assumes that "...people generally write more cooperatively at the beginning of the word, while the variability increases in the middle of the word." In the medical environment, variability is apparent when multiple health care professionals enter data on the same form. The medical environment also has exaggerated or extremely compressed word lengths due to erratic movement in a vehicle and limited paper space. Kaufmann [64] only provides a reduction of 25% of the lexicon size with little to no improvement in error rate, and the tests are run on a small sample of words.

Relatively little research has been done with the linguistic model for the purposes of lexicon reduction and information retrieval from degraded handwritten images. On about 15% of the medical forms, half of the documents were completely indecipherable by human beings. This illustrates the challenges of automated recognition.

2.2 Handwriting Recognition

Handwriting recognition (HR) is divided into two categories: online and offline. An HR survey paper by Tappert [125] discusses the processing and recognition of on-line handwriting across multiple languages and compares the differences between on-line and off-line recognition. Online recognition is performed on hardware devices, such as PDAs, and generally has the advantage of positional and temporal knowledge, and the disadvantage of having to process information rapidly to avoid user frustrations. The offline recognition process, such as the recognition of postal mail or historic documents, performs all processing on a document that has been completely handwritten. In many circumstances, there is more available run-time (e.g. a user scanning a document on his personal computer), however offline recognition lacks positional and temporal knowledge. This study addresses the offline recognition problem in which all text is available to the machine. However, if there is a substantial amount of handwriting text in an online recognition environment, the algorithms can still be used. Migration to an online environment needs previous context (e.g. a sentence) and the first one or two characters of a word being written (since lexicon reduction occurs during the writing process).

The recognition of postal mail [38] [66] in the United States, for example, is an off-line procedure in which time constraints have become increasingly critical as society has become dependent on automation. While postal recognition is a challenging problem, many algorithms have been used to solve the problem. Some algorithms incorporate postal databases to reduce the possibilities of words in a corpus. This is an example of *lexicon reduction*. This dissertation will introduce an alternative approach. However, not all scenarios have the luxury of an all-encompassing postal database engine. This research addresses another generation in recognition in which lexicons can not only be large, but they can also be on multiple heterogeneous topics. The task of handwriting recognition in forms involves enormously complex handwriting, lacks a consistent texture, exists at various locations within forms, involve multiple writers on the same document, suffers exposure to rescue situations, consists of free-form text, and has no database lookups.

Nevertheless, if people can read the handwriting on a medical form, then it should be possible to create automated systems as well. This statement is supported by evidence of high recognition performance by other automated systems with similar problems, such as the postal mail recognition problem [38] [66]. Initially it is expected that the machine will not perform as well as humans. This is acceptable for two reasons: (i) this is the first attempt at recognition of this kind, and (ii) the proposed system can operate in synchronization with humans (i.e. as an assistive process where the machine submits an interpretation to the human for verification and/or modification).

It is imperative to note McDermott's warning with the usage of words in artificial intelligence [83]. For example, while the words *semantic* or *reasoning* are used to describe the process in solving a problem, it can be considered a philosophical illusion. A stochastic process, for example, may be used to solve a problem in which humans solve by *reasoning*, however, the computational technique may or may not be performing an act of *reasoning*. McDermott specifically uses the example that a human may name a computational function *UNDERSTAND* which computes some output based on the input. However, if the function were named *G0034*, it is difficult to presume that the function is *understanding*. Therefore, the use of *syntactic* and *semantic* are more precisely defined in this research.

Handwriting recognition algorithms have generally started with syntactic [3] approaches and included semantics [4] along the way [45] [49]. The syntactic approach can be thought of as a parser which does not exhibit any knowledge other than geometric (e.g. concavity, curvature) information. Guillevic, et al. [49] discusses contextual, syntactic and semantic details in cursive script using psychological models. This classification of syntactic and semantic analysis by human readers as a *guessing game* was shown by Goodman [45]. An important issue that syntactic and semantic categories raise is related to two layers of ambiguity: (i) confusion with characters that are visually similar or identical, and (ii) the *true*[5] context of words in a sentence or paragraph. To address both situations, this research has two syntactic recognition steps (i.e. character extraction and handwriting recognition discussed later) separated by a *semantic* step (i.e. lexicon reduction involving the mapping of terms to categories; discussed later).

In this research, *semantic* is defined as the mapping of words or word encodings to

[3] Techniques for interpretting the structure of handwritten words.

[4] Techniques for utilizing the meaning of handwritten words.

[5] The notion of truth is defined as the human interpretation.

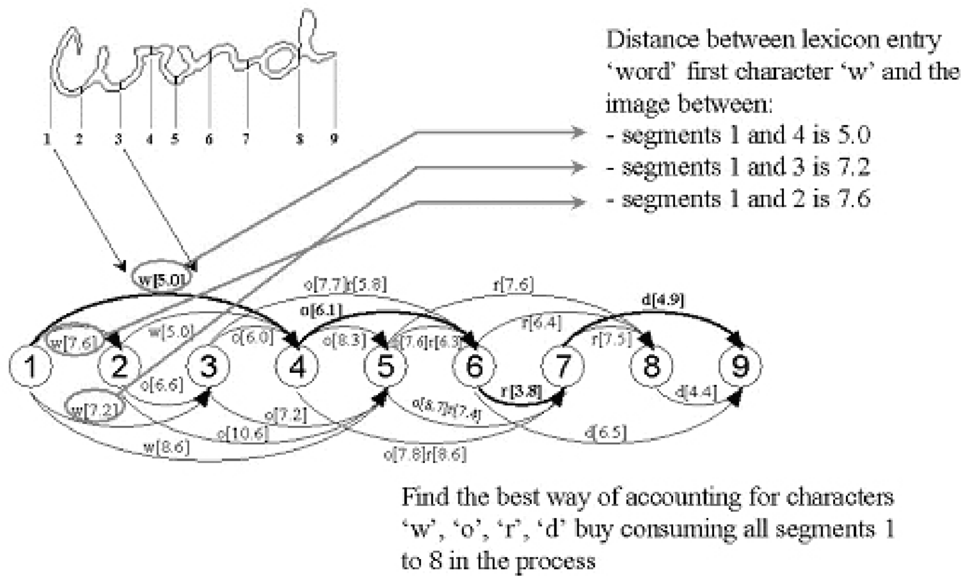


Figure 2.1: Lexicon Driven Word Recognizer Algorithm [66]

a finite set of topic *categories*. This combination of syntactic and semantic approaches is shown to improve the performance by allowing the lexicon to be reduced using these *categories*. Our research will show that is possible to use only those words found on a particular subject. This is then followed by the second syntactic recognition step on this reduced lexicon. In other words, by organizing heterogeneous information into homogeneous categories, the ambiguity issues are intuitively reduced; thereby, improving recognition performance. While our algorithm combines different classifiers, it differs from the multiple classifier combination (MCC) problem. In MCC, the task is to make a classification using the outputs of several independent classifiers. A simple solution for this problem is the majority voting principle [124]. Another probabilistic approach in the context of numeral handwriting recognition can be found by Xu, et al. [137]. Our research uses different classifiers which are serially dependent and therefore the MCC problem does not apply.

While many word recognition engines exist, the HR algorithms chosen for this research have had widespread trusted use with the United States Postal Service (USPS) (see Figure 2.1) [66]. In word model based recognition, all lexicon entries are treated as isolated words and matched against the input word image containing handwriting to recognize. The lexicon entry with the best match is the top choice. In character model based recognition, segments are matched against individual characters without using any contextual information implied by the lexicon. Word hypotheses are generated by the character recognition results. If the best hypothesis is found in the lexicon, then the recognition is done; otherwise, the second best hypothesis is generated and tested, and so on. Since the lexicon driven word recognizer (LDWR)[66] has the highest recognition rates, and it is lexicon dependent, this research focuses on how to get the LDWR the smallest lexicon possible.

While the LDWR has excellent performance in small lexicons (e.g. less than 100 words), it does not perform well on larger lexicons (e.g. $\geq$ 5,000 words) due to the confusion in selecting from many choices. The lexicon reduction approach presented in this work addresses the inadequacies of this recognizer's performance by supplying it a smaller lexicon. However, a challenge in reducing the lexicon is the retention of the actual word after the reduction. Another inadequacy of LDWR is its high error rate due to the segmentation procedure when applied to medical handwriting. Therefore, a new binarization and post-processing algorithm is introduced to reduce segmentation failures due to broken strokes. In order to address the LDWR's confusion, this research shows that a couple of characters from a word can provide sufficient semantic information. For example, if "BL" only matched the word "BLOOD", then our lexicon should be geared only to those topic categories involving "BLOOD." This has the potential of reducing the lexicon while maintaining correct word retention since the semantic analysis incorporates the characters, by the same recognizer, as a suggestion. The more suggestions provided, the better the guess of the semantic category.

2.3 Lexicon Reduction

Lexicon reduction is any process that takes an unknown input image of a word and a set of possible ASCII words (i.e. the lexicon), and produces a subset of words (i.e. the reduction). It has been shown that the reduction in lexicon size has strong potential for improving the recognition performance [47] [66] [81]. The objective of lexicon reduction is generally to: (i) produce the smallest subset of words possible, and (ii) retain the desired word, which is obviously unknown, within this subset. Clearly, if the desired word is not in the lexicon during the time of word recognition, the word image will fail recognition.

Several approaches for lexicon analysis and reduction have been performed previously. Holistic approaches, such as Madhvanath's [80] [81], is motivated by human reading studies and utilizes word shapes such as length, ascenders and descenders. Holistic algorithms provide a visually intuitive approach to reduction. This becomes compromised in the emergency setting where health-care professionals are writing in several environments: (i) movement in large emergency vehicles which are changing speeds, (ii) walking with the patient into the emergency room (ER), (iii) rescuing and writing with medical gloves, and (iv) the existence of multiple writers. This severely impacts all holistic aspects of a word. For example, a moving ambulance that changes speed affects word length, un-smooth driving surfaces affect word height, and different writers affect all directions and structure of the words. Consider the image in Figure 2.2, in which the central letter is out of alignment with the rest of the word and has strong potential for being classified as an ascender [80]. Since holistic approaches are not involved in character level segmentation, this approach is intuitively problematic in such situations.

Figure 2.2: Problematic example for holistic approaches (word displayed is *"sternal"*)

2.4 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a theory and procedure for computing the relationships between the context of words and terms to a semantic category.

> . . . LSA represents the meaning of a word as a kind of average of the meaning of all passages in which it appears, and the meaning of a passage as a kind of average of the meaning of all words it contains. [70]

Landauer, et al. states here that while the human approach at semantic comprehension is unknown, LSA mimics human sorting of words into categories and simulates "passage coherence" [70]. The term *latent semantic* is coined due to the semantic inferences it attempts to achieve [70]; latent is used in the connotation of a *concealed* meaning. In this research, we are concerned with modeling the semantic relationships between partially recognized handwritten characters and a category [6]. More specifically, this research shows that this character information, from words describing patient treatment, is sufficient for modeling and later querying for a medical category. This facilitates both improved handwriting recognition performance and improved search engine results.

LSA represents these term-category associations in multiple orthogonal dimensions simultaneously. This allows a reduction performed on LSA to minimize those parameters necessary to produce a deeper semantic meaning [34] [70]. LSA, which is a statistical approach for constructing these relationships, does not take any input other than the words

[6] This bares resemblance to contextual vocabulary acquisition by Rapaport [103]

and categories. It is still able to compute such relationships (i.e. no external knowledge bases are necessary). LSA performs these computations by applying the Singular Value Decomposition (SVD) [121] to a matrix. The SVD equation $X = U \bullet S \bullet V^T$ computes the eigenvalues and eigenvectors [5] from a rectangular matrix $\mathbf{X}$ [70] [34]. This matrix consists of weighted and normalized frequencies indicating the relationships of terms (rows) to categories (columns) computed beforehand. $\mathbf{U}$ contains the row eigenvectors of terms, $\mathbf{V^T}$ contains the column eigenvectors of categories, and $\mathbf{S}$ contains the now decomposed singular values (note: *singular values* are equal to the square-root of the eigenvalues), computed from the matrices, in descending order along the diagonal. The values along this diagonal represent the degree of variance such that the first element in the diagonal has the most variance and the last element has the least variance. The reduction can be performed by deleting the smallest values in this diagonal matrix [34] [70]. Once decomposed into orthogonal vectors (i.e. all vectors are perpendicular to each other in multiple dimensions), the classification of an unknown vector of terms against the appropriate vectors in the decomposed matrices can be performed by computing the cosine of the vectors. This allows the unknown vector of terms to be matched against a category. This can be thought of as a query into the multiple dimensional semantic category space [21] [22] [34]. The theorem and proof by induction that all matrices have an SVD can be found in [127].

SVD has also been used to model data in diverse areas such as gene expression [4], protein molecular dynamics [41] [106], weather forecasting [43], call-routing [21] [22], image compression [104], face recognition [52], cryptanalysis [89], and information retrieval [11] [12] [34] [140]. The most stable utility found to compute the SVD in the Java programming language [60] is provided by NIST's JAMA package [59]. This package utilizes QR-Decomposition (a.k.a. QR-Factorization) [5] instead of approaches such as Gram-Shcmidt [5], which suffers from rounding errors during computation [1]. The SVD numbers computed in this research were verified against the appendix by Deerwester et al. [34].

2.5 Forms Processing

In this section, some examples of form recognition types are illustrated from more to less constrained. The purpose is to illustrate the increasing complexity in form recognition.

• Figure 2.3: Bank checks are forms in which there is a small lexicon of numbers and their string counterparts (e.g. 1,000 and One Thousand) [2] [65]. This figure depicts the steps in performing bank check recognition: pre-processing, segmentation, word recognition, generation of numerical cost candidates, and finally a list of recognition results with confidence scores. The use of word recognition to restrict the lexicon of cost candidates is an example of lexicon reduction.

• Figure 2.4: Census forms are restricted to a small quantity of dictionary words relating to information such as occupation, employment, location, etc. (e.g. School) [79].

• Figure 2.5: Postal mail-piece recognition contains handwriting related to street addresses, cities, states, countries, barcode markings, various stamps, and codes. Address block information (noting both the return address and the destination address), which contain either the hand or machine print address, need to be recognized in such an application [38] [66].

Historical documents (although not strictly a form) contain unconstrained handwriting, may use an archaic vocabulary, and are written on a complex surface [39] [46] [126]. They represent the challenge of handwriting recognition (see Figure 2.6) when no form structure applies.

In order for work on Figures 2.3 - 2.6 to be performed, it is necessary to have form extraction algorithms available [8] [94]. Such algorithms already exist. We assume that

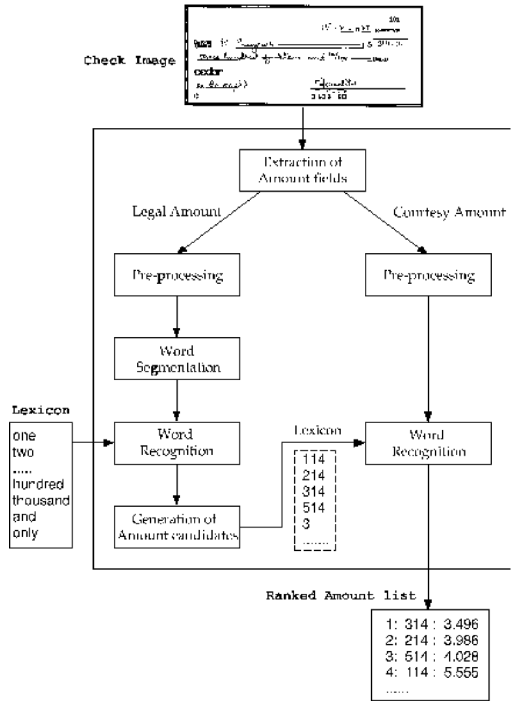
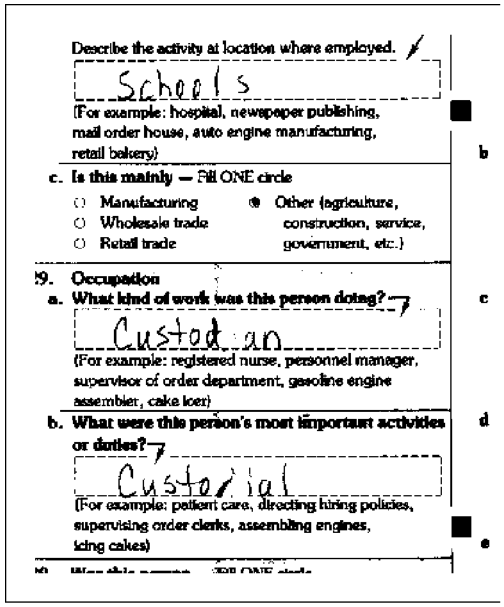


Figure 2.3: Bank Check Recognition Example [65]

existing algorithms are sufficient for the extraction of the handwriting regions of the PCR, and that the five major handwriting components of the PCR are already extracted. These regions are static and hence easily located on the form.

2.6 Information Retrieval

Information retrieval generally involves the indexing of information followed by a query to a search engine. Surveys of these technologies can be found in [25] [48] [53] [58] [71] [72] [102]. This research is concerned with the retrieval of information using semantic-based indexing in the medical domain using a vector space model [21] [22] [34] [71] [72][140].



Describe the activity at location where employed.

Schools

(For example: hospital, newspaper publishing, mail order house, auto engine manufacturing, retail bakery)

c. Is this mainly — Fill ONE circle

○ Manufacturing
○ Wholesale trade
○ Retail trade
● Other (agriculture, construction, service, government, etc.)

29. Occupation

a. What kind of work was this person doing?

Custodian

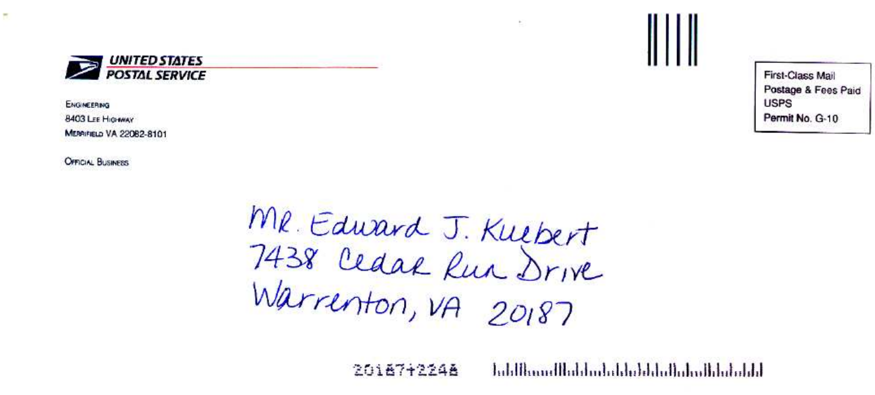
(For example: registered nurse, personnel manager, supervisor of order department, gasoline engine assembler, cake icer)

b. What were this person's most important activities or duties?

Custodial

(For example: patient care, directing hiring policies, supervising order clerks, assembling engines, icing cakes)

Figure 2.4: Census Forms [79]



UNITED STATES POSTAL SERVICE

ENGINEERING
8403 LEE HIGHWAY
MERRIFIELD VA 22082-8101

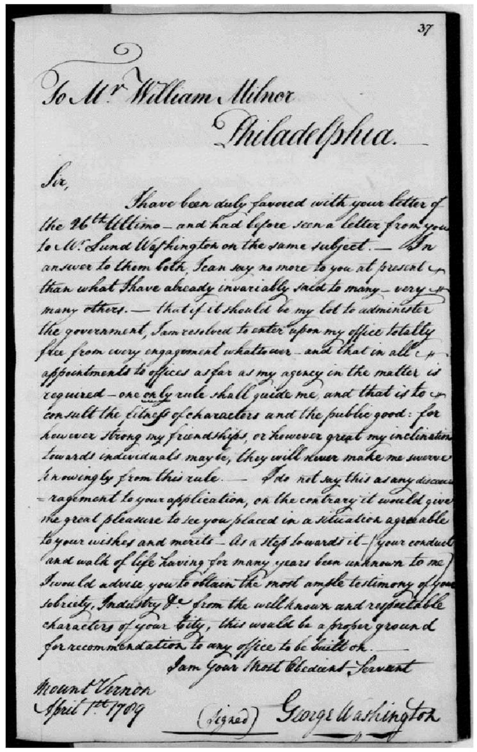
OFFICIAL BUSINESS

First-Class Mail
Postage & Fees Paid
USPS
Permit No. G-10

Mr. Edward J. Kuebert
7438 Cedar Run Drive
Warrenton, VA 20187

20187+2248

Figure 2.5: Postal Letter [66]



37

To Mr. William Milnor
Philadelphia.

Sir,

I have been duly favored with your letter of the 26th Ultimo — and had before seen a letter from you to Mr. Lund Washington on the same subject. — In answer to them both, I can say no more to you at present than what I have already invariably said to many — very many others. — that if it should be my lot to administer the government, I am resolved to enter upon my office totally free from every engagement whatsoever — and that in all appointments to offices as far as my agency in the matter is required — one only rule shall guide me, and that is to consult the fitness of characters and the public good: for however strong my friendships, or however great my inclination towards individuals may be, they will never make me swerve knowingly from this rule. — I do not say this as any discouragement to your application, on the contrary it would give me great pleasure to see you placed in a situation agreeable to your wishes and merits — As a step towards it — (your conduct and walk of life having for many years been unknown to me) I would advise you to obtain the most ample testimony of your sobriety, Industry &c from the well known and respectable characters of your City, this would be a proper ground for recommendation to any office to be built on. —

I am Your Most Obedient Servant

Mount Vernon
April 1st 1789

(Signed) George Washington

Figure 2.6: Historic Document [46]

Yang et al. [140] learned the association between words and indexing terms of documents using a Linear Least Squares Fit (LLSF) approach with the MEDLINE [84] database and Mayo patient records [140]. In order to manage the problems of unrestricted vocabularies, the work draws associations based on those assigned by human subjects. In addition, it introduces the phrase *"surface-based matching"* which denotes any search technology approach that determines a match only when the document contains the words from a query. It finds that other such *surface-based matching* approaches have poor performance when unrestricted vocabularies are used, and concludes that a better approach is to solve the problem using concept-based categorization and retrieval. This approach involves the categorization of document words in relation to a category using vector space [140]. Yang et al. [140] differs from Salton [110] and Deerwester [34] by mapping between two vector spaces; a source space of words (x-axis) to texts (y-axis) and a target space of document category sets (y-axis) to categories (x-axis). Yang et al. [140] uses it to map the text

"AIDS and Guillain Barre Syndrome" to the two categories (i.e. category set) *"Acquired Immunodeficiency Syndrome"* and *"Polyradiculoneuritis"*. It computes the Singular Value Decomposition (SVD) [74] [44] on a co-occurrence matrix which solves the mapping function needed for LLSF. At this point, the text is essentially biased towards certain target categories. Given a source text, the category can be queried by computing the cosine in vector space and using the highest score to represent the highest candidate [140]. Evaluation was conducted with respect to Categorization and Retrieval, multiple weighting schemes, and differing morphological processing. It was executed on three data sets: (i) *A Data Set of Clinical Categorization (SURCL)* containing reports in natural language by physicians, (ii) *A Data Set of MEDLINE Retrieval (MEDIR)* containing title and abstract from a MEDLINE citation, and (iii) *A Data Set of MEDLINE Document Indexing (MEDCL)* containing word and category information.

Chu-Caroll et al. [21] [22], (to be discussed in greater detail in the following sections) focuses on a similar approach using a single SVD. While Yang et al. [140] performed the analysis on text, Chu-Caroll [21] [22] took voice recognition data as an input and produced a caller destination to solve the call-routing problem. This research is closer to Chu-Caroll [21] [22] in using recognition information as an input. In addition, their work illustrates robustness to noisy data, reduction in error rates, and a high call-routing success rate. Our research concentrates on the handwriting domain.

The Yang et al. [140] and Chu-Carroll [21] [22] techniques have similarities with the lexicon reduction problem addressed here. Similarities are found in medical and recognition approaches. Yang's et al. [140] approach involves a mapping between vector spaces of information in the medical domain; however, the data is in a known text format. Chu-Caroll's [21] [22] approach involves mapping of voice recognized words to a caller destination. However, those terms are constructed using the voice recognition information as the input. In this approach, the question is whether low confidence recognition characters from a pair of unknown words have a cohesion which can be mapped to a category in the medical domain. Using this information, the recognition engine itself can be bootstrapped to improve the recognizer and then use all of the information to index medical forms containing only handwriting.

2.7 Human Cognition and Ontology

The question of human perception is still an open problem under study in many disciplines, such as Artificial Intelligence (AI), Cognitive Science, Philosophy, and Neuroscience. Ontology is a branch of cognitive science and philosophy using such things as topologies, axioms, and logic to describe reality. Since AI is not a solved problem, a gap exists between how a human mind computes a solution and how a computational system solves the same problem. Therefore, creating a standard ontological framework in bio-informatics has not yet been achieved. A biomedical ontology is conerned with the relationships among classes or categories [25]. This research exhibits traits of biomedical ontologies in which a human classifies human injury and ailments to a higher level anatomical category. These relationships are similar to several biomedical ontologies. The Foundational Model of Anatomy (FMA), developed by the University of Washington, models information and categories of the human body [25]. Rosse [107] breaks down a biomedical entity into several sub-categories, one of which denoted *material anatomical entity*. Several other references and examples of anatomical ontologies can also be found in [25].

In the framework of ontology, our research maps partially known and inaccurately recognized handwritten medical text to anatomical categories that can be used to loop back into the handwriting recognition and information retrieval algorithms. Description logic used with electronic patient health records for defining such free-form medical text relationships to escape the more primitive "is_a" type relationships, can be found in Smith [117]. The application of the ontology, namely the mapping of terms to categories, can be approached from a statistical perspective rather than employing description logic. This is a requirement since it is not known how human beings interpret handwritten information. This paper promotes the theory of terminology mappings in health records in accordance with Smith [117] in two ontologies: (i) the recognition of information, and (ii) the interpretation of information. While this approach creates an alternate ontology in phase (i), further medical inferences can be used by Smith's [117] during phase (ii). It should be further noted that Smith [117] also incorporates temporal knowledge like logic into the set theory. This is used in the medical domain to determine a cause-effect relationship (e.g. "mechanism of injury" due to a car accident).

The motivation for using anatomical categories in the model is twofold: (i) intuition that treatment is related to a patient's condition in the emergency medical environment, and (ii) philosophical theories in Smith's *Anatomical Information Science* [118]. This framework uses a FMA, a computational system consisting of a collection of 1.5 million statements involving 70,000 kinds of anatomical relationships [118]. Since the relationship mappings involve known text which is unavailable during the recognition process, this system is much larger than what would be required in this medical environment, which typically employs fewer than 30 concepts in a specific domain. While it is possible that global deployment of this research in multiple fields may one day take advantage of such systems, it is not currently feasible to use a general knowledge base for specific forms. The theory defined in Smith's work focuses on actual anatomical connections, locations and containment of anatomy within the body, the theory maps terms to anatomical categories. This dissertation contributes to the newly established *Anatomical Information Science* [118] by introducing an alternative type of term mapping involving character recognition information.

2.8 Summary

The section introduces prior work in the areas of handwriting recognition and information retrieval. This research assumes that the medical form can be segmented into its regions using the afforementioned form processing strategies. The greatest challenge is the improvement of the handwriting recognition algorithm to facilitate medical form searching. The Lexicon Driven Word Recognizer (LDWR) [66] is the state-of-the-art handwriting recognition engine that will be improved on by the use of a new lexicon reduction strategy. This reduction technique uses latent semantic analysis to relate partial handwriting character information to anatomical categories. The selection of the anatomical categories has its roots to *anatomical information science*, which argues that the human information can be categorically mapped as an anatomical ontology. In addition, the patient treatment is related to those anatomical positions that are treated. Therefore, the hypothesis of the lexicon reduction algorithm is that some recognized characters can be used to represent anatomical categories.

Chapter 3

Binarization of Carbon Copy Images

This research evaluates several algorithms that extract handwriting from medical form images (see Figure 3.1) to eventually provide the best handwriting recognition performance. This extraction of handwritten stroke pixels from the image is known as binarization. The research copy of the NYS PCR [135] is a carbon mesh document where both the foreground handwriting and the background carbon paper use approximately the same intensity values. While the handwriting on the top form has direct contact between ink and paper, the carbon does not transfer to the paper if there is insufficient pressure. This loss of complete character information in the carbon copy causes character strokes to break after binarization, which leads to recognition failures (the phrase *pressure sensitivity issues* will refer to this situation). Prior binarization algorithms have been reported to better manage noisy and complicated surfaces [42] [75] [136] [141]. However, the broken/unnatural handwriting due to ambulance movement and emergency environments, as well as carbon smearing from unintentional pressure to the form, add further complexity to the binarization task. A lexicon-driven word recognizer (LDWR) [66] is used for evaluation of the binarization methods. Analysis of the LDWR, as well as a full view of an actual NYS PCR image, can be found in [86].

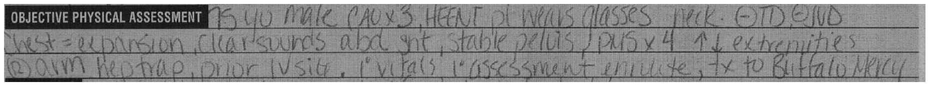


Figure 3.1: NYS PCR Objective Physical Assessment

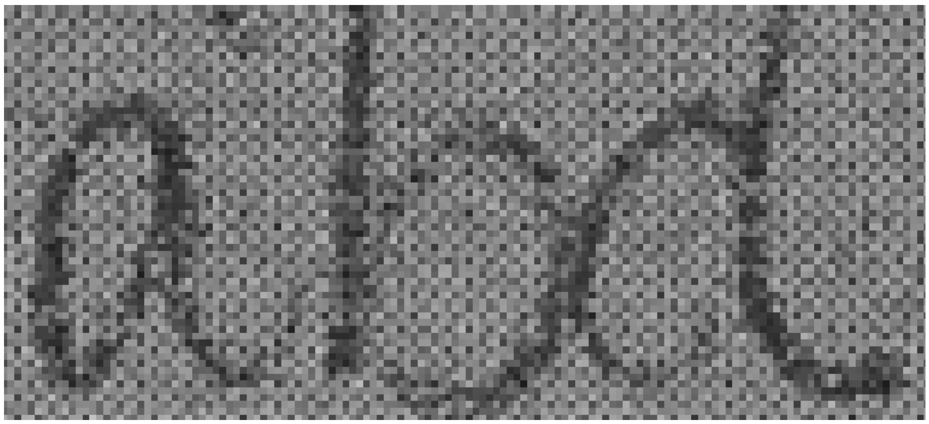


Figure 3.2: Grayscale 256 Carbon Mesh Handwriting Example (400% Zoom)

3.1 Carbon Paper

The inconsistent carbon paper, which shows varying grayscale intensities (see Figures 3.1 and 3.2), is referred to as carbon mesh. Figure 3.1 shows an example of the “Objective Assessment” region of the NYS PCR form. It provides an overview of the complex nature of the handwriting on the carbon paper. Figure 3.2 shows a 400% zoom of one word from Figure 1. It shows the carbon paper mesh integrated with the carbon handwriting stroke. The displayed word *abd*, in Figure 3.2, is a common abbreviation for abdomen. Since the carbon paper causes the paper, the stroke, and any artifacts to have the same intensities, the binarization problem becomes complex. Details of the application of existing algorithms will be discussed in the following sections. This paper describes an algorithm for binarizing the handwriting on carbon paper while preserving the handwriting stroke connectivity better than prior algorithms.

Pressure sensitivity issues, as a result of light strokes in penmanship, affect the extent of character connectivity after binarization. In order for the carbon copy to receive a reasonable representation of the top copy original, the health care professional needs to press down firmly with the writing instrument. Since the emergency environment is not conducive to good penmanship, the binarization and cleanup algorithms need to compensate.

The carbon paper forms also contain guidelines, which often interfere with the character strokes. These lines can be detected by those pixels with a grayscale value less then a pre-determined threshold; this is consistent across all forms in our data set. To reduce stroke fragmentation, it is sufficient to retain the pixels near the line, thus keeping most character ascenders and descenders reasonably connected. This form drop-out step is performed before binarization.

3.2 Prior Work

In this section, methods described in previous works are compared with our algorithm presented in this research. First we consider the processing of the image in Figure 3.3a, using various filters. The histogram of this image, shown in Figure 3.4, shows that the foreground (handwriting stroke) and background (carbon paper) use the same intensities in the supplied range. A split at any position in the histogram results in loss of both foreground and background information. The x-axis of the histogram represents the grayscale values 0-255 such that the left most position 0 represents black and the right most position 255 represents white. The y-axis of this histogram is the quantity of pixels for its corresponding grayscale intensity. The mean, median and standard deviation are computations on the grayscale intensities. The standard deviation shows the statistical dispersion of grayscale intensities with respect to the mean. The smaller standard deviation value indicates the grayscale values are clustered around the mean intensity value. The evaluation of pre-processing filters followed by the application of existing binarization algorithms on image 3.3a is discussed throughout this research.

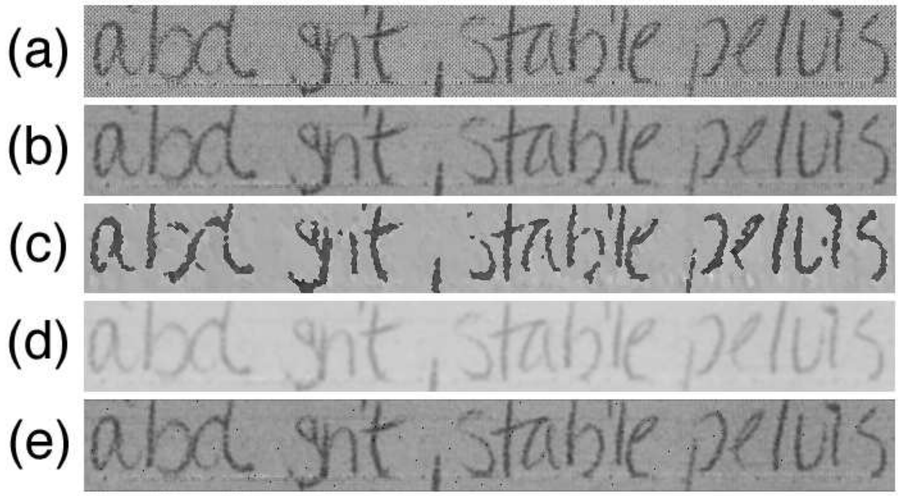


Figure 3.3: Smoothing Operations (a) Original image + Form drop out (b) Mean filter (c) Median filter (d) Gaussian filter (e) Weiner filter

Gaussian, median, mean and Weiner filtering/smoothing have been studied in previous works, as a base step, or an integrated step, for noise removal and image enhancement [42] [51] [119] [128] [142]. Mean filter (Figure 3.3b) shows the least damage to strokes in our experiments. Median filter (Figure 3.3c) illustrates severe character damage. Gaussian filter (Figure 3.3d) demonstrates characters being washed into the background. Weiner filter (Figure 3.3e) produces an image very similar to the mean filter, except the background surface is slightly lighter and stroke edges are sharper. Gatos, et al. [42] uses the Weiner filter as a pre-processing step to filter image noise.

Global thresholding algorithms determine a single threshold and apply it to the entire image. In the PCR application, the high pressure sensitive areas are binarized well, whereas medium to low pressure areas run the risk of being classified as background.

Other works use algorithms that address some weaknesses of the Otsu [76] [99] method, such as with degraded documents. Any algorithm that computes a global threshold

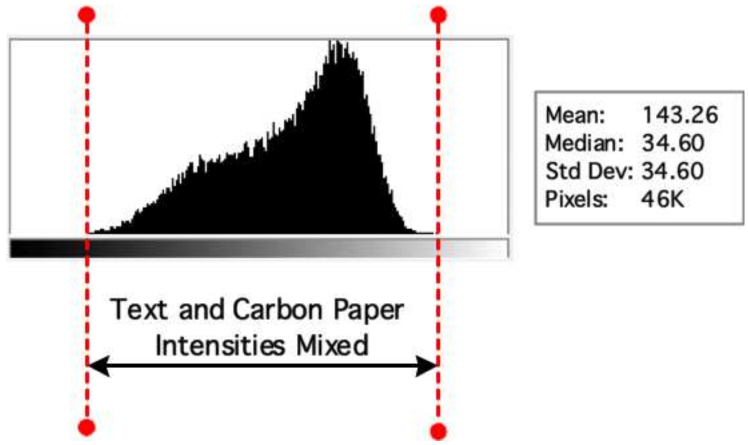


Figure 3.4: Histogram for Image in Figure 3.3a

splits the histogram into foreground and background pixels using that threshold. However, since both foreground and background pixels can have the same intensities at different positions on the image, splitting the histogram globally will incorrectly classify foreground pixels and backgrounds. The Wu/Manmatha [136] method expects at least two histogram intensity peaks and globally splits the histogram. This causes large portions of the handwriting to be lost to the background. To compensate, a histogram split is allowed to occur directly before the largest intensity peak in the image (note the highest histogram peak in Figure 3.4). This improves the performance of the algorithm, but, still suffers from stroke and background pixels trapped in the largest histogram peak (see Figure 3.4).

The Niblack binarization [93] algorithm is an adaptive technique that has been compared to other methods in applications such as image and video text detection and extraction [134], low quality camera images [114], low quality grayscale utility maps (such as cable and hydro maps with various intensity and noise issues [128]), and low quality historical documents [42]. This algorithm results in severe noise, jagged edges and broken character segments. While post-processing improves the algorithm performance, the broken character strokes result in lower performance. This is due to mean-variance computations occurring at lighter stroke regions.

Sauvola binarization [111] modifies the Niblack algorithm [93] and attempts to suppress noisy areas. In the cases of stronger handwriting pressure, Sauvola [111] has positive results. However, Sauvola [111] has fewer positive results than Niblack [93] in our experiments. Sauvola's [111] noise suppression affects the lighter strokes thereby causing incorrect recognizer segmentation.

Gatos, et al. [42] introduced an algorithm that incorporates Sauvola [111], but this implies that the performance of Gatos, et al. [42] will drop along with that of Sauvola [111]. While Gatos, et al. [42] does illustrate a performance improvement over Sauvola [111], this combination still under-performs Niblack [93] after post-processing. This is because Gatos often loses holistic features due to incorrect background estimation of the paper.

Logical binarization uses heuristics for evaluating whether a pixel belongs to the foreground or background. Other adaptive binarization strategies are integrated with such heuristics. The Kamel/Zhao algorithm [63] finds stroke locations and then later removes the noise in the non-stroke areas using an interpolation and thresholding step. Various stroke width combinations from 1-10 pixels were tried. However, the stroke is not adequately traced using this algorithm.

The Yang/Yan [141] algorithm is a variant of the method developed by Kamel/Zhao [63]. The modifications are meant to handle low quality images affected by varying intensity, illumination, and artifacts such as smearing. However, the run analysis step in this algorithm is computed using only black pixels. Neither the foreground (handwritten stroke) or background (carbon paper) of the carbon copy medical forms have black pixels; nor are the foreground pixels the same intensity throughout. Therefore, the stroke-width computation, which is dependent on the run-length computation, cannot be trivially determined in the carbon paper forms.

In addition to the binarization algorithms, various post-processing strategies are commonly used. The despeckle algorithm is a simple noise removal technique using a 3x3 mask to remove a foreground pixel that has no D8 neighbors [119]. The blob removal algorithm is a 9x9 mask that removes small pixel regions that have no neighbors [119]. The amorphous artifact filter removes any connected component whose pixel area is less than a threshold (60 pixels in this research) [119]. The Niblack [93] + Yanowitz and Bruckstein method [142] was found to be the best combination strategy by Trier and Taxt [128]. The Shi and Govindaraju method is an image enhancement strategy that has been used on postal mail-pieces [115].

3.3 Proposed Algorithm

Prior algorithms have relied on techniques such as histogram analysis, edge detection, and local measurements. However, these techniques are less effective on medical forms. Our algorithm uses a larger central NxN mask, which determines the intensity of one region, and compares it with the intensities of multiple dynamically-moving smaller PxP masks (see Figures 3.5, 3.6 and 3.7).

One hypothesis in managing the varying intensities of the carbon mesh and its similarities with the stroke is to use a wave trajectory (see Figures 3.6 and 3.7) for the D8 positioned masks (see Figure 3.5), as opposed to a linear trajectory (see Figure 3.8). A wave/trajectory is a path, in a Cartesian system, that undulates across an axis in 2D space with an amplitude and frequency that can be adjusted (see Figure 3.6). The experiments illustrate that the use of a wave trajectory is beneficial for the following reasons: (i) There is a better chance of the trajectory of the mask to evade a stroke. (ii) The possibility of finding a background region as close as possible to the central mask is enhanced. Note that as one goes further out from the center mask, the more likely it is to find that the carbon mesh of

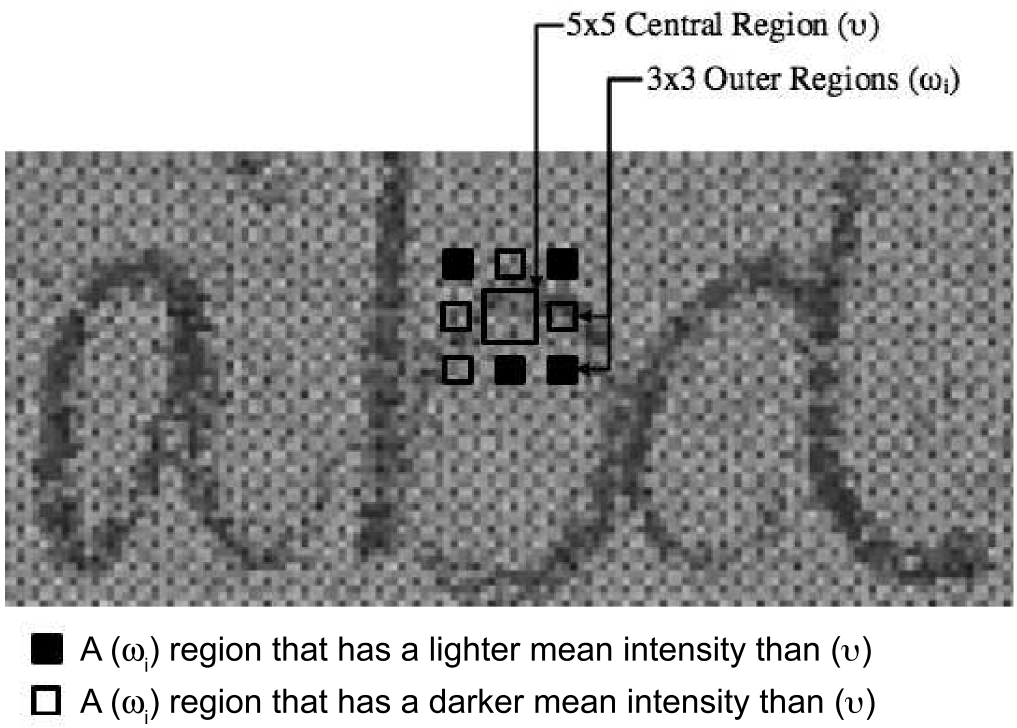


Figure 3.5: Initial Mask Placement Example (N=5 and P=3)

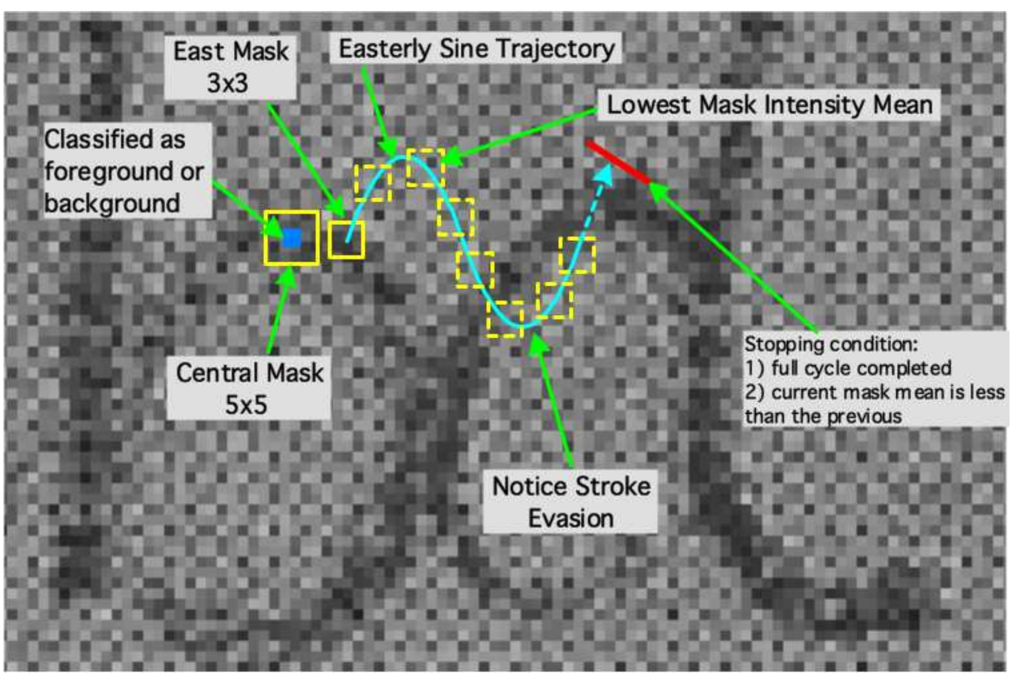


Figure 3.6: Sine Wave Trajectory

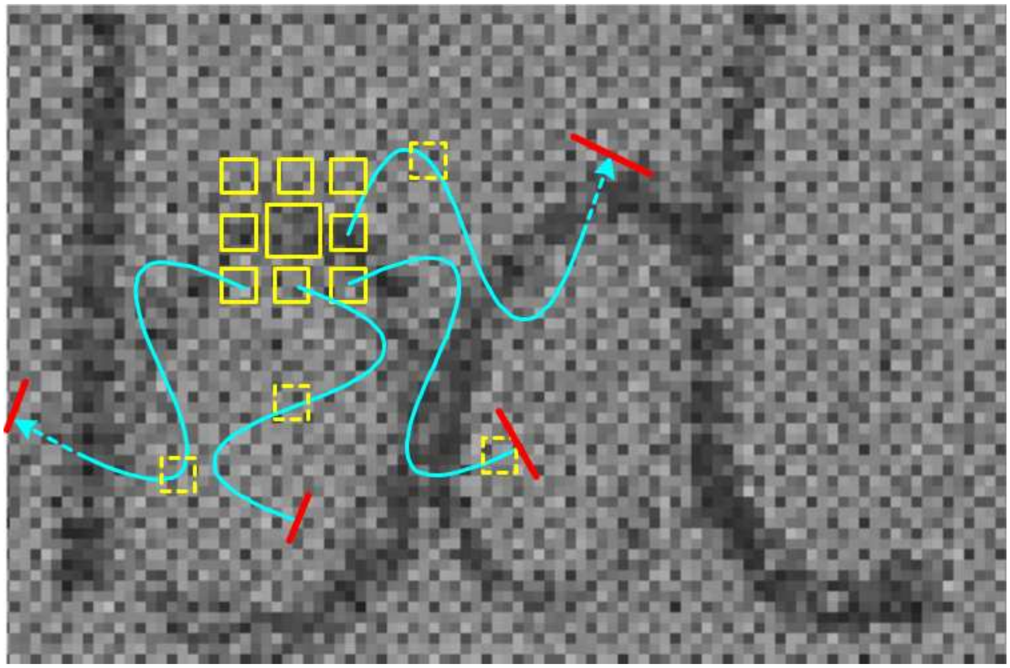


Figure 3.7: Sine Wave Coverage

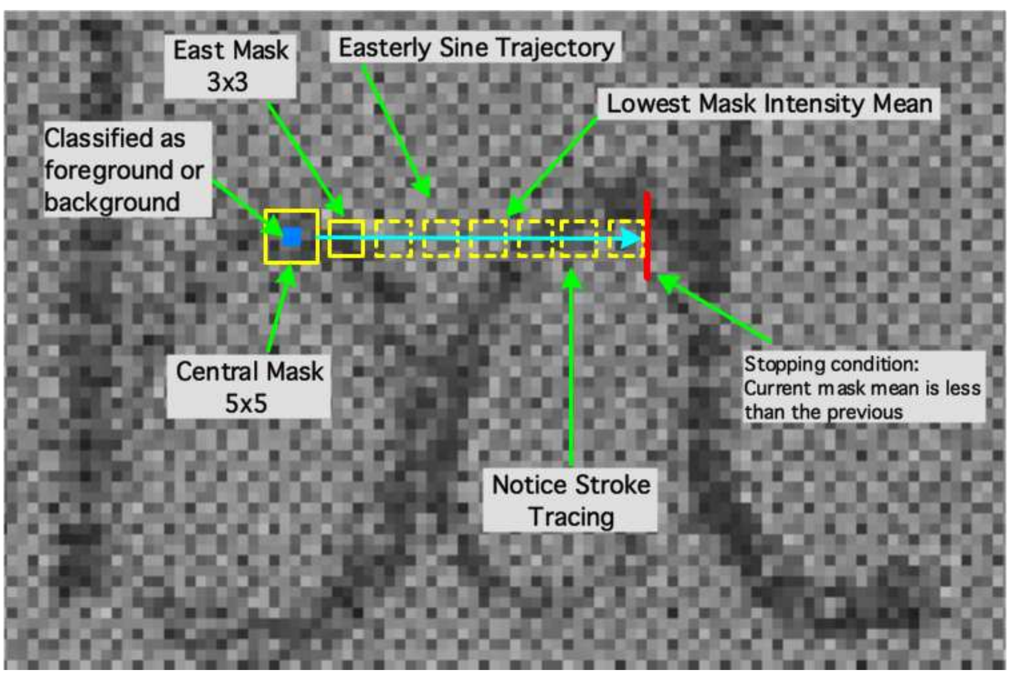


Figure 3.8: Linear Trajectory

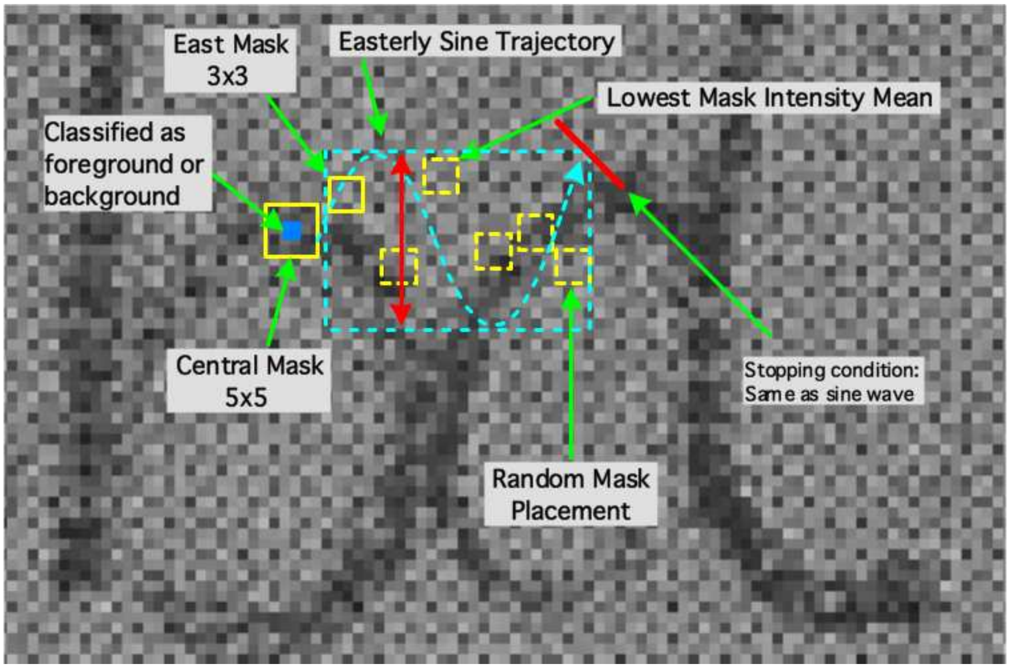


Figure 3.9: Random Trajectory

the background can change. (iii) The best background region to compare to a handwriting stroke may or may not be the edge of the stroke. (iv) Areas surrounding a stroke in the same trajectory can be observed. (v) Eight points of comparison, one for each trajectory, are performed (as opposed to one point used on other algorithms). With the inclusion of a stopping condition operating independently on each trajectory, this approach, as opposed to other global and adaptive approaches, does not get confined to square mask windows which are relative to a central position. In this context, the wave trajectory for scanning can be thought of as searching for lighter pockets in the intensity fluctuation of the carbon mesh (see Figures 3.6 and 3.7).

A sine wave trajectory offers the benefit of beginning at the origin and allowing a continuous trajectory regardless of distance (i.e. the wave will continue until the stopping condition is met as opposed to being confined to an arbitrary box). It allows the control of frequency and amplitude that is necessary to adjust for stroke width. Sinusoidal waves have been used in other contexts for the modeling of human motor function for on-line handwriting recognition, feature extraction and segmentation [9], shape normalization of Chinese characters [77], and signal canceling of pathological tremors while writing [56]. Based on these studies, and the knowledge of the English character set, it was possible to scan out from a character stroke at a certain frequency. This allows a handwriting stroke to be maneuvered, as opposed to traced, in the search for background regions. The sine trajectory can be thought of as a path which has the potential to cross handwritten strokes. This allows the background paper on both sides of the stroke, in all directions, and with a dynamic distance, to be evaluated. Intuitively, more space can be searched and both sides of the stroke can be evaluated in the same computational step at variable distances. It is also presumed that in a moving ambulance, carbon smearing is more likely since the writer will press harder on the document to maintain balance in the vehicle. While strokes in the English language contain both curves and straight lines, at the pixel level they can be considered piecewise linear movements such that a linear scan will trace the stroke and reduce the likelihood of finding the background. Furthermore, holistic features (such as the area in the letter "D") are typically small. Missing the carbon paper inside such character holes may result in missed background analysis. This motivated the use of a higher sine wave frequency so that the trajectory would pass through the center of holistic features as frequently as possible. Additionally, since the thickness of characters fluctuates, it is difficult to precisely calculate the true stroke width.

An input grayscale image 0 (black) to 255 (white) is the input, and a binarized image is the output. At a given position on the image, there are 9 masks. A single mask is denoted as Ξ. The mean intensity of all pixels within a single mask is denoted by M(Ξ). The central mask which slides across the image is denoted by (υ) and has a size NxN, such that N $\geq$ 3 and consists of numerically odd dimensions (e.g. 3x3, 5x5, and 7x7). The size of (υ) is based on the estimated stroke width constant denoted by ϕ. The value of ϕ has been estimated to be 5 pixels, therefore (υ) is of size 5x5. At each (υ) position over the image, 8 masks are initially stationed in each D8 position (Figure 3.5) and are denoted by (ω_i) where

$1 \leq$ i ≤ 8. The mask size of (ω_i) is PxP such that $3 \leq$ P $\leq \lceil$N/2$\rceil$. Note that P $\leq \lceil$N/2$\rceil$ allows a small mask the opportunity of preserving small holistic features when moving on the sine curve, and also making sure that the mask will not overlap (υ). Each (ω_i) is initially stationed as close to (υ) as possible so as to avoid the mask overlapping between (ω_i) and (υ). Each (ω_i) moves in its respective D8 direction, either linearly (see Figure 3.8), randomly (see Figure 3.9) or via a sinusoidal wave (see Figure 3.6). The M$(\omega_i)$is computed at each position along the trajectory and stops at a position after one cycle and when the current mask average intensity is lighter than the previous one on the sine trajectory. A list of mean values for each position on that trajectory is denoted by $M(\omega_i)_q$ where *q* is a coordinate on the sine curve. The lowest intensity mask on a single sine wave trajectory is represented by the equation $M(\omega_i)_{min} = min(M(\omega_i)_{\forall_q}))$. Next, a comparison of all the D8 $M(\omega_i)_{min}$ positions are made against M(υ). If there are 3-4 (empirically determined) of the 8 $M(\omega_i)_{min}$ values which satisfy the equation $M(\omega_i)_{min} - M(\upsilon) \geq \kappa$, such that κ is a small constant (using κ = 10), then the center pixel of (υ) is classified as a foreground pixel. The value κ defines a tolerance with respect to the localized intensity fluctuation of the carbon paper and denotes the carbon intensity similarity rule. Given that a new image has been initialized to white background pixels, it is only necessary to mark the foreground pixels when they are found. A dynamic programming step is used to store each $M(\omega_i)$, corresponding to the appropriate region on the image beforehand, to improve the run-time performance.

The sinusoidal trajectory is defined by Equation 3.3.1.

$$y = 2\phi sin(\frac{1}{2}x) \tag{3.3.1}$$

The coordinate (x, y), on a sinusoidal trajectory is relative to its starting location (origin). A nearest neighbor approach is sufficient for conversion of real coordinates to pixel coordinates. Each ω_i is computed on the sine curve trajectory (see Figures 3.6 and 3.7). Note that using ϕ as the amplitude in Equation 3.3.1, without the coefficient, will result

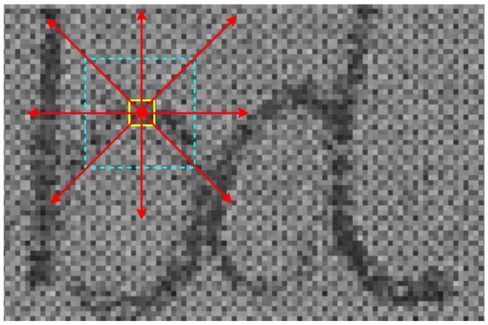


Figure 3.10: Sliding Square Window

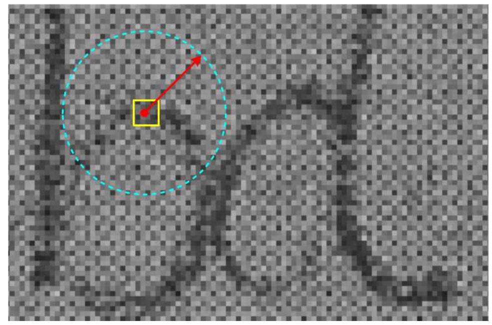


Figure 3.11: Sliding Circular Window

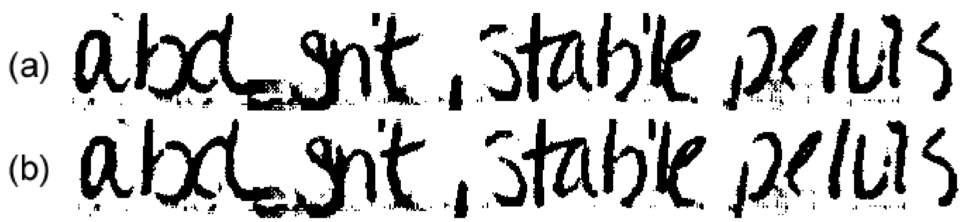


Figure 3.12: Window Binarization Examples: (a) Sliding Square Window (b) Sliding Circular Window

in a distance of 2ϕ between the highest and lowest y-axis points (using ϕ as the stroke width). In addition, using 2ϕ as the amplitude yields a distance of 4 times the stroke width, to account for the possibility of 2 touching strokes (i.e. two touching letters). In this way, the curve efficiently exits a stroke while searching for the background. The constant $\frac{1}{2}$ is used in Equation 3.3.1 so that the sine frequency does not trace the handwritten stroke. The constant chosen is handwriting style specific. The objective is to evade the stroke, but since the stroke pixels are not known, nor easily approximated, a constant is chosen.

One alternative approach to the sinusoidal approach would be to search for the lightest mean intensity mask within a sliding window that is the same distance as one complete cycle of the sinusoidal wave. Figures 3.10 and 3.11 illustrate the window structures and Figure 3.12 shows the output. The procedure for computing these window structures is to calculate the intensity average of all 3x3 masks within a window. It is then possible to compare the lightest of these averaged intensity masks with the average intensity of the central 5x5 mask. However, only a single mask value within the window is compared to the central mask which requires κ (as defined above) to be larger (κ = 30 in Figure 3.12) resulting in a completely black image. This is due to the larger area of coverage which increases the likelihood that a lighter area will always be found relative to the center mask, thereby classifying the area as foreground. In addition, there is a degradation of holistic features, close-proximity characters combining and more broken characters than in the sine wave approach (see section 5 results). This requires κ to be based on the image. The sine wave approach rarely suffers from this situation since at least 3 of the trajectories authenticate a foreground value instead of one value. Therefore, the carbon intensity similarity rule breaks down, and causes the effectiveness of these sliding window approaches to be as problematic as a global thresholding technique. In this way, κ becomes the global threshold.

Another possible approach computes the Otsu [99] algorithm in small windows rather than over the entire image. However, this results in an output image nearly identical to that obtained by computing Otsu [99] globally. The thresholds chosen are negligibly different between windows and, therefore, the image is still noisy and many strokes are still broken.

An alternate strategy to the sine wave is a randomized mask movement (Figure 3.9). Instead of the outer masks moving on the sine wave trajectory, they move on the y-axis randomly within the same rectangular area of the sine wave movement. It may be expected that the randomized version (see Figure 3.9) will perform as well as the sine wave. However, since the window involved in the sine wave trajectories is reasonably small, if a stroke is present within that window, and the randomized approach is used, there is no guarantee that the stroke will be evaded. Therefore, if a random position is chosen, and that rests on a stroke as opposed to the background, then the desired background position is missed. The sine wave approach is more likely to cross the stroke rather than tracing it. Furthermore, due to the nature of randomized approaches, the recognition results may not be consistent.

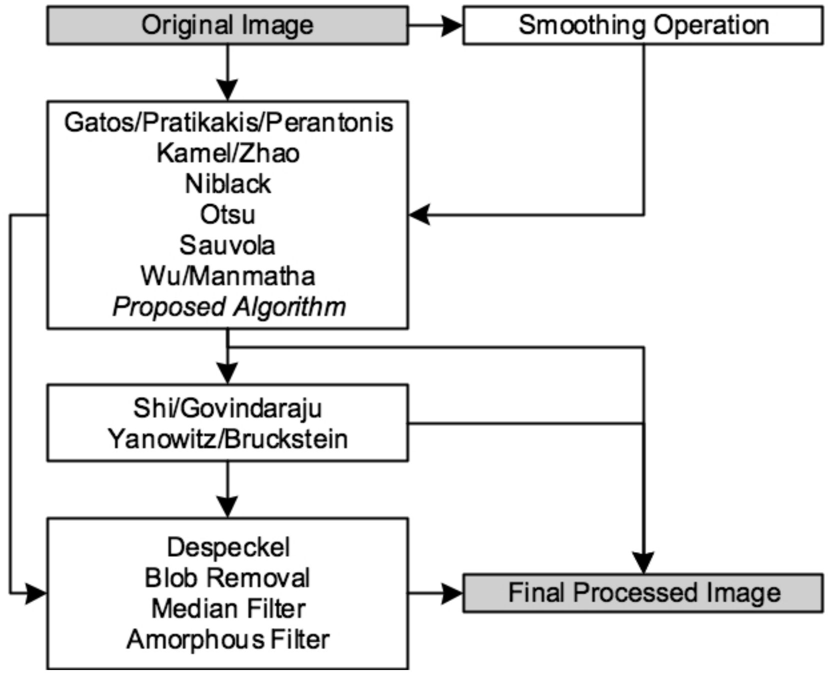


Figure 3.13: Image Processing Combinations

3.4 Results

All experiments (see Figure 3.13) were performed on a set of 62 PCRs consisting of $\sim 3,000$ word images and various size lexicons (see Figures 3.16 and 3.17). The linear, random, square and circular strategies (see Figures 3.8-3.12) were outperformed by Otsu [99]. The sine wave strategy presented here outperformed all prior algorithms, with a 11-31% improvement. After post-processing there was a 4.5-7.25% improvement (see Figures 3.16 and 3.17).

The handwriting phrase depicted in Figure 3.14 and 3.15, “abd snt, stable pelvis” means “abdominal soft-not-tender, stable pelvis.” Figures 3.14 and 3.16 show the performance of the aforementioned binarization strategies with no post-processing support.

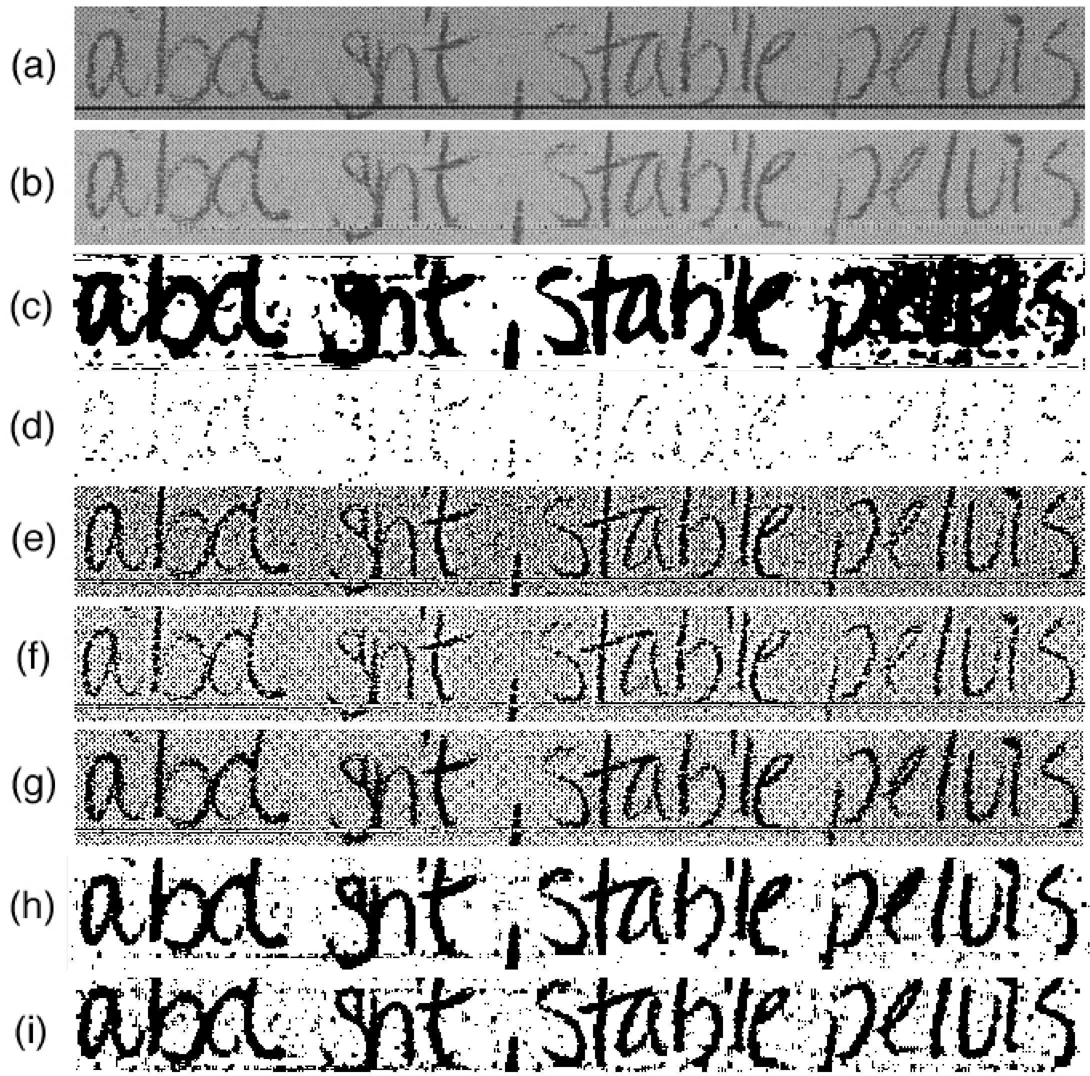


Figure 3.14: Comparison of Binarization Algorithms Only: (a) Original image (b) Original image with form drop out (c) Wu/Manmatha Binarization (d) Kamel/Zhao Binarization (e) Niblack Binarization (f) Sauvola Binarization (g) Otsu Binarization (h) Gatos/Pratikakis/Perantonis Binarization (i) Sine Wave Binarization

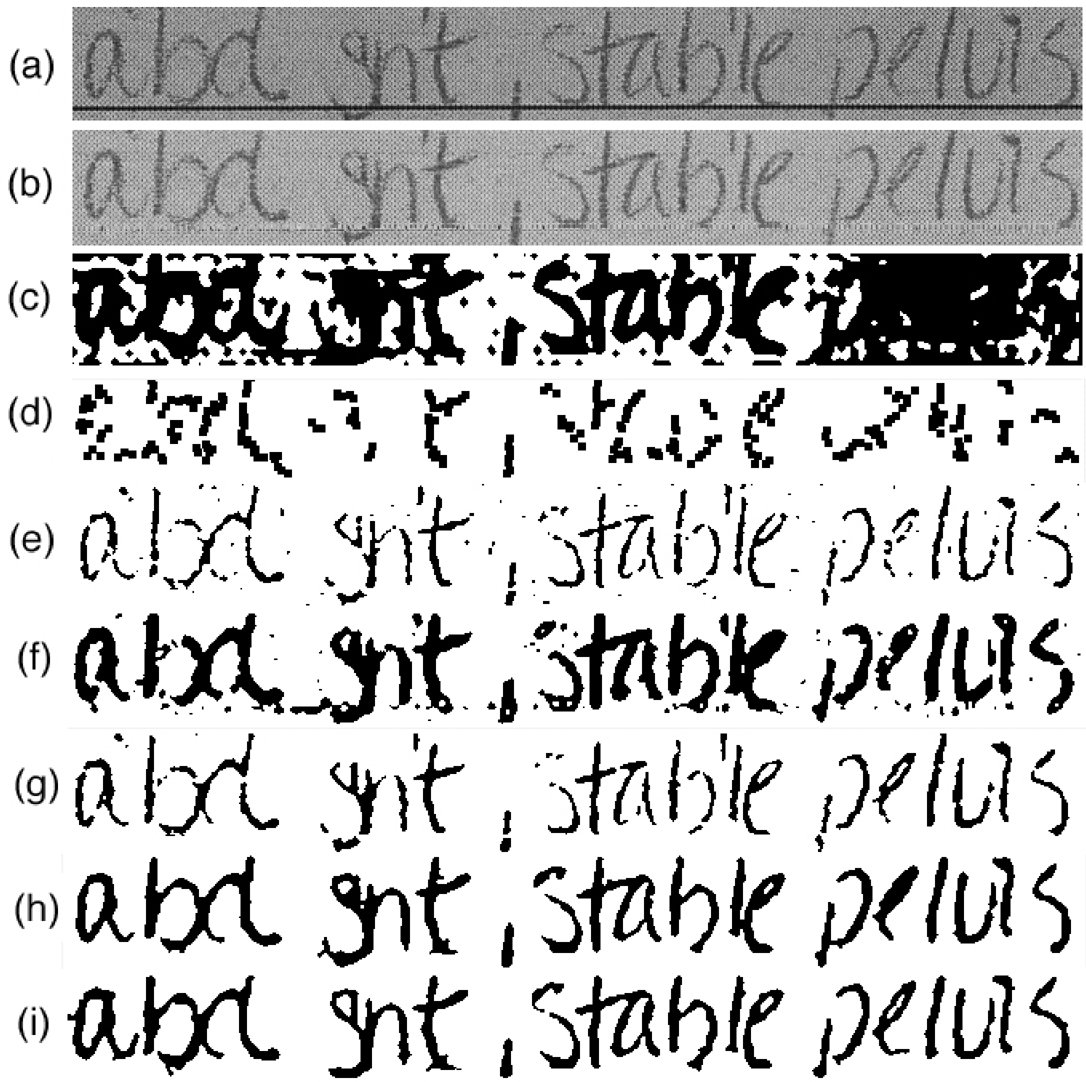
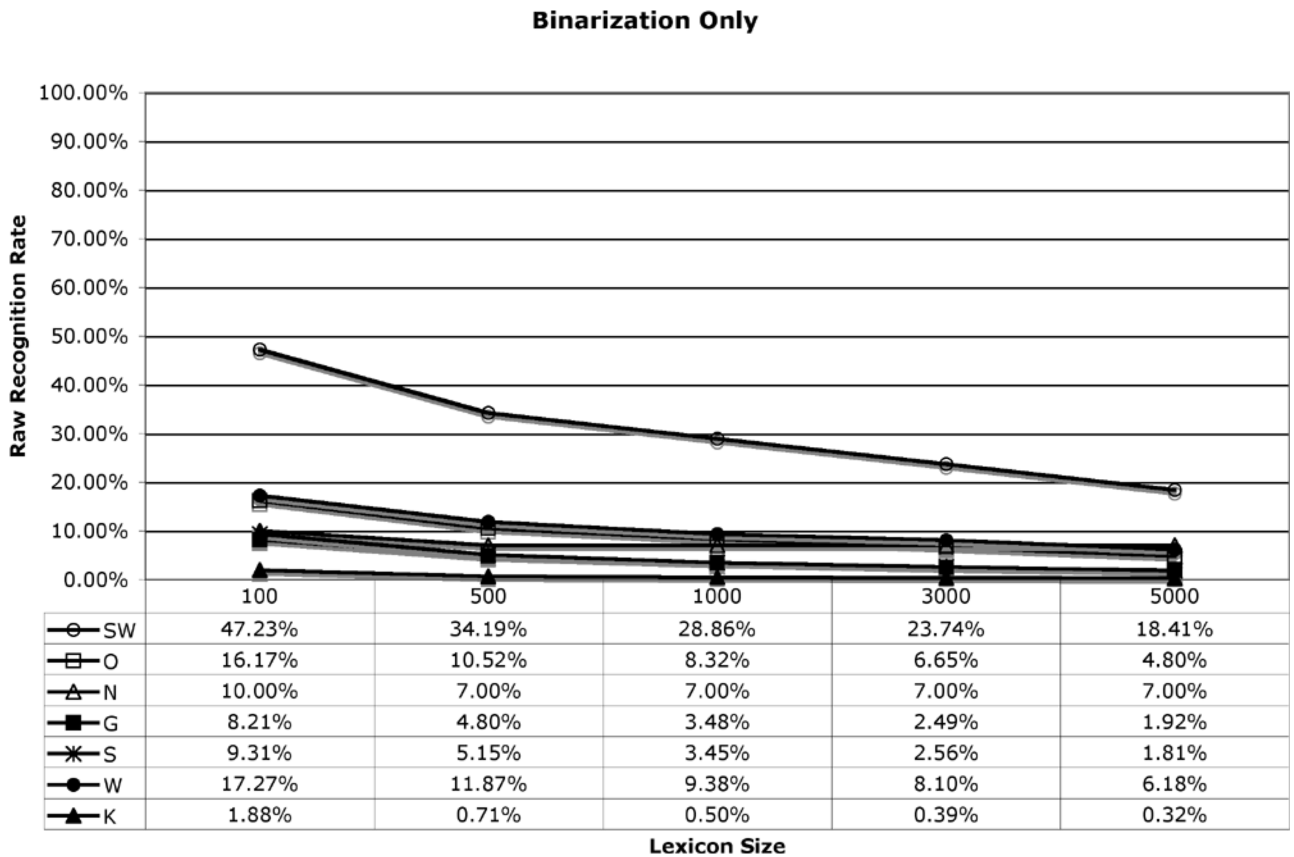
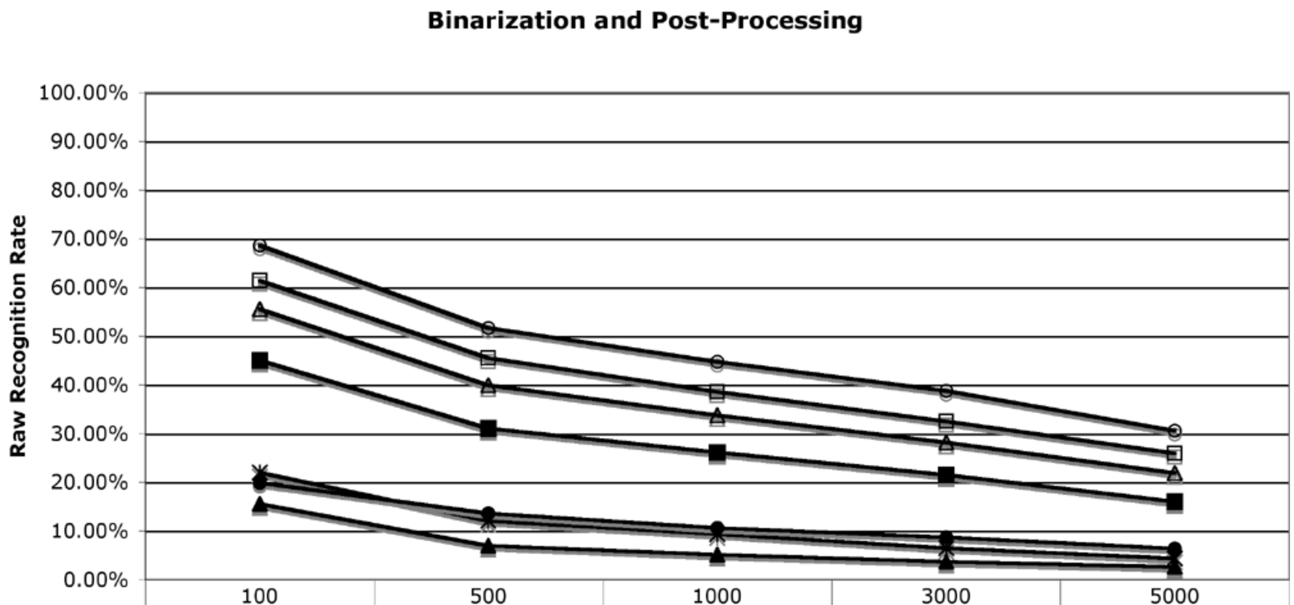


Figure 3.15: Comparison of Binarization Algorithms with their Best Post-Processing Strategy: (a) Original image (b) Original image with form drop out (c) Wu/Manmatha Binarization + Shi/Govindaraju + Despeckel (d) Kamel/Zhao Binarization + Shi/Govindaraju + Despeckel + Amorphous Filter (e) Niblack Binarization + Despeckel + Amorphous Filter (f) Sauvola Binarization + Yanowitz/Bruckstein + Despeckel (g) Otsu Binarization + Despeckel + Blob Removal (h) Gatos/Pratikakis/Perantonis Binarization + Despeckel + Amorphous Filter (i) Sine Wave Binarization + Amorphous Filter



	100	500	1000	3000	5000
SW	47.23%	34.19%	28.86%	23.74%	18.41%
O	16.17%	10.52%	8.32%	6.65%	4.80%
N	10.00%	7.00%	7.00%	7.00%	7.00%
G	8.21%	4.80%	3.48%	2.49%	1.92%
S	9.31%	5.15%	3.45%	2.56%	1.81%
W	17.27%	11.87%	9.38%	8.10%	6.18%
K	1.88%	0.71%	0.50%	0.39%	0.32%

Figure 3.16: Binarization Only Performance



	100	500	1000	3000	5000
SW	68.55%	51.60%	44.71%	38.70%	30.56%
O	61.37%	45.49%	38.56%	32.48%	25.94%
N	55.50%	39.84%	33.69%	28.11%	21.89%
G	45.02%	31.02%	26.08%	21.46%	16.03%
S	22.00%	12.04%	9.31%	6.43%	4.30%
W	19.86%	13.54%	10.59%	8.70%	6.32%
K	15.53%	6.93%	5.08%	3.62%	2.63%

Lexicon Size

Figure 3.17: Binarization + Post-Processing Performance

Figures 3.15 and 3.17 reflect the performance of the best respective post-processing combinations from Figure 3.13.

With respect to Figures 3.16 and 3.17, the y-axis represents the percentage of correctly recognized words by the LDWR [66] versus the lexicon size on the x-axis. It is expected that the performance decreases with an increase in lexicon size since the LDWR [66] is lexicon driven and, therefore, has more choices from which to select. The definition of correctly recognized words is in the context of raw recognition performance. The error rate is consistently high in this application. Furthermore, words in the form region were manually segmented. The LDWR [66] algorithm uses pre-processing strategies for its own noise removal and smoothing before executing its recognition algorithm [66] [113] [16]. Therefore, a noisy image submitted to the LDWR algorithm will be internally pre-processed by the handwriting recognizer. The first letter of an author's name is used to refer to the algorithms: (G)atos, (K)amel, (N)iblack, (O)tsu, (S)auvola, (W)u. (SW) designates Sine Wave binarization.

3.5 Conclusions

Several methods address the problem of extracting degraded text; however, they generally cause broken gaps and lost holistic features. It appears that algorithms relying on histogram separation, interpolation and mean-variance perform poorly. In addition, these algorithms determine a foreground pixel based on a single value from a larger sliding window, sometimes computed after an intensity interpolation. However, our algorithm classifies a pixel based on 8 masks that observe other pixel regions in a non-linear fashion. The results seem to improve the readability for humans as well as improve automatic recognition performance substantially. This provides insight into the human ability to effectively extract the stroke. The following chapters use word images, binarized and processed by the sine wave and amorphous filter, and proceed with the handwriting recognition.

Chapter 4

Lexicon Reduction Measures

Recall the definition of lexicon reduction from Chapter 2.3 in which a process produces a smaller lexicon using an unknown word image and a complete lexicon for inputs. The purpose of the reduction is to improve both performance and run-time of the recognition algorithm [47] [80] [81]. The following sections discuss the hypothesis, reasoning, and performance measures involved in the reduction approach.

4.1 Lexicon Category Hypothesis

This research proposes the following hypothesis, which is verified experimentally: A sequence of confidently recognized characters, extracted from an image of handwritten medical text, can be used to represent a topic category. These categories are a finite list determined by a human trainer and stored in a knowledge base. A lexicon can be reduced by keeping only those words belonging to those categories. The topic categories used in this research pertain to human anatomy and are found in Table 4.1.

The associations between category and form word relationships are performed by a person skilled in health care. Each category contains a lexicon of words extracted from only those medical forms assigned with a respective category; overlapping may occur. It is the aim of the lexicon reduction algorithm to determine the categories for an unknown form. Only words assigned to the determined categories are used during the recognition process. It was originally conceivable that the complete lexicon could be determined from medical data sets. However, in reviewing the OHSUMED [132], NLM MeSH [92], and NLM UMLS [129] data sets, the text information appears inconsistent with the text information found in the PCR reports. These medical databases are clinical, laboratory, and/or research-based information in which the semantic scope is too detailed for PCR handwriting recognition. Since there is no obvious correlation between the phrases or categories from these medical data sources, they cannot be used. It has been determined by prior research that the use of such databases may decrease retrieval performance [53].

Another approach to the extraction uses the bubble sheet values (see Figure 1.2 Location 10) on the medical PCR forms as categories; typically, these types of form blocks have high recognition. Most of the PCR forms suffer from the following: (i) Bubbles that should have been filled in are not, (ii) bubbles are not always correctly filled in, (iii) some bubbles still require the entry of a handwritten phrase next to them, (iv) a high frequency of transcription of other documents over the carbon copy forms result in incorrect carbon markings over the entire region, and (v) the requirement of bubble information in the modeling research requires a form with a similar organization (recall the fields in Figure 3). Therefore, while form information may appear to be useful, it was found to be incomplete and inaccurate. As a result, an alternative approach for determining categories using the form handwriting was used.

4.2 Anatomical Categories

All PCRs are manually tagged with up to five categories from Table 4.1. During the testing, the system detected categories automatically. This work is dependent on the semantic of words and categories in the emergency medical domain. This anatomical topology, used as the PCR categories, corresponds to the patient ailment location(s) (see Table 4.1). A PCR can be tagged with multiple categories. Anatomical categories were chosen to test the hypothesis that the patient's treatment (information on the medical form) is related to the patient's ailment locations, which are anatomical by nature. In our data set, no form had more than five category tags.

The subjectivity involved in determining the categories makes the construction of a hierarchical chart representing all patient scenarios with respective prioritized anatomical regions a difficult task and exceeds the scope of this research. The following are some examples for classifying medical form text into categories (see Table 4.1):

Example 1: A patient treated for an emergency related to her pregnancy would be classified under the *Reproductive System* category (see Table 4.1).

Example 2: A conscious and breathing patient treated for gun shot wounds to the abdominal region would fall into the *Circulatory/Cardiovascular System* due to potential loss of blood, as well as being categorized for *Abdominal, Back, and Pelvic* categories (see Table 4.1).

4.3 Lexicon Category Methodology

The recognition of a word out of context can be a difficult task even for human beings. Consider the task of reliably identifying the words in Figures 4.1 or 4.2 out of context. The words may be in the domain of English, medicine and/or pharmacology. The interpretation of the handwriting in Figures 4.1 and 4.2 require context even for humans.

In reference to Figure 4.3, consider the same words in their actual context on a medical form. While some doubt still remains with respect to the first two words, it is expected

10 Body Systems	6 Body Range Locations	4 Extremity Locations	4 General
Circulatory/Cardiovascular	Abdomen	Arms/Shoulders/Elbows	Fluid/Chemical Imbalance
Digestive	Back/Thoracic/Lumbar	Feet/Ankles/Toes	Full Body
Endocrine	Chest	Hands/Wrists/Fingers	Hospital Transfer/Transport
Excretory	Head	Legs/Knees	Senses
Immune	Neck/Cervical		
Integumentary	Pelvic/Sacrum/Coccyx		
Musculoskeletal			
Nervous			
Reproductive			
Respiratory			

Table 4.1: Categories are denoted by these Anatomical Positions

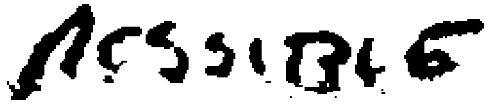


Figure 4.1: Unknown Word 1 with no Context [86]

Figure 4.2: Unknown Word 2 with no Context [86]

that one would have a higher confidence in the identity of these words because of contextual clues. Furthermore, if told that this phrase is found on the objective assessment of a medical form, the certainty increases further. The simple phrase is "Several Possible Wounds" and on the medical form from which this example was extracted, an individual had fallen from a dangerous height [86].

Although syntactic recognition algorithms are still necessary, research has shown that recognition performance is lexicon driven [20] [47] [50] [64] [66] [81] [139] [143]. There is a need to compensate for the degradation in recognition accuracy caused by large lexicons. The modeling of associations between medical form text and categories restricts the recognizer classification to specific topics [86].

Games such as The Wheel of Fortune[1] and crossword puzzles involve the relationships of characters to words and words to phrases. Similarly, some characters in the example are easily recognizable while others are not. If it is assumed that some characters

[1]Game show website can be found here: http://www.wheeloffortune.com

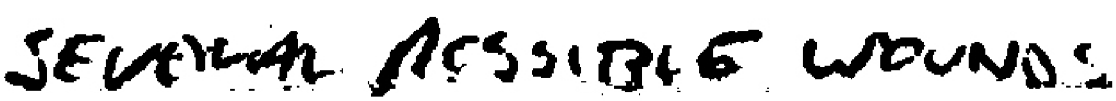


Figure 4.3: Unknown Words in Context ("several possible wounds") [86]

are recognized with high confidence, it is possible: (i) to infer the meaning from the context and use the information to improve word recognition, and (ii) use partially recognized character information to decipher the context.

4.4 Measuring Lexicon Reduction Performance

The performance measures for lexicon reduction as described by Madhvanath [80] and Govindaraju, et al. [47] are discussed in this section. A lexicon reduction algorithm takes an input word image x_i and a lexicon L_i and computes a reduced lexicon Q_i such that $Q_i \subseteq L_i$ and *i* which indicates the position within a list of *n* images such that $1 \leq i \leq n$. The truth of the image x_i is represented as t_i. The function *E* computes the expectation (the mean is used to estimate the expectation) [14] of a random variable. A random variable is a function or mapping representing the outcome of an experiment. These performance measures will be used when gaining insight into the effectiveness of the lexicon reduction technique. In this research, all words on a PCR receive the same reduced lexicon. In this situation *i* is not used to distinguish independent lexicons for each word and can be omitted.

• *Accuracy of Reduction:* $\alpha = E(\mathbf{A})$ such that $\alpha \in [0, 1]$ [80].

The value α simply represents the mean existance of a word in the reduced lexicon. $\mathbf{A}$ is a random variable [14] which indicates whether or not the truth of the image exists in the lexicon. In the context of $\mathbf{A}$, the variable *t* represents the truth for an image and *Q* represents a reduced lexicon.

$$\mathbf{A} = \begin{cases} 1, & if\ t \in Q \\ 0, & otherwise \end{cases}$$

• *Degree of Reduction:* $\rho = E(\mathbf{R})$ such that $\rho \in [0, 1]$ [80].

The value ρ simply represents the mean size of the reduced lexicon. The value $\mathbf{R}$ is a random variable [14] representing the extent of the reduction. In the context of $\mathbf{R}$, the variable L represents a complete lexicon and Q represents a reduced lexicon.

$$\mathbf{R} = \frac{(|L| - |Q|)}{|L|}$$

• *Reduction Efficacy:* $\eta = \Delta_{LDWR} \times \alpha^{1-\rho}$ such that $\Delta_{LDWR}, \eta, \alpha, \rho \in [0,1]$.

Reduction efficacy was originally defined as $\eta = \alpha^k \cdot \rho$ such that $\alpha, \rho, \eta \in [0,1]$ by Madhvanath [80]. The value η is a single value which incorporates both the accuracy of reduction and degree of reduction for determining the effectiveness of the reduction. The constant *k* allows a weighting of the accuracy relative to the degree of reduction and is empirically determined [80].

However, the equation $\eta = \alpha^k \cdot \rho$ appears to be counter-intuitive for the following reasons:

◦ *Case 1:* Suppose that the recognition rate improves by 50% or drops by 50%. Since the efficacy measure does not take into account the recognition rate, the true effectiveness of the reduction is not apparent. One assumption that appears to exist in this equation is that a reduction with the word still remaining in the lexicon means a good reduction. However, suppose that the word images themselves are so complicated that the recognizer still fails in the interpretation, regardless of the existence of the word in the lexicon. Since the intent of the reduction is to improve the recognizer peformance, this dependence must be incorporated into the measure. This case is especially of concern in the difficult interpretation of medical handwriting.

◦ *Case 2:* Suppose that α = .9 and ρ = .1; when multiplied together, in the case that *k*=1, the result will be same as α = .1 and ρ = .9. In other words, the value of a high degree of reduction with low accuracy is equivalent to a high degree of accuracy with low reduction. These cases are actually not equivalent and therefore should not receive the same value. Given the same values for α and ρ, as *k* increases, the measure in accuracy is penalized.

This is not desired since the accuracy of the reduction is most important, considering a lexicon driven word recognizer requires the existence of the word in the lexicon. Conversely, as *k* decreases, the measure of the output of the equation changes exponentially. Therefore, selecting a constant in the exponent, especially if *k* is fractional, is not intuitive because of the non-linear nature of the computation. While the existence of *k* appears to be for weighting the degree of accuracy over the degree of reduction, the output is misleading.

○ *Case 3:* The purpose of the *reduction efficacy* is to provide a means of comparison between different recognizers. However, the application specific constant of *k* would therefore have to be the same for all applications to make an appropriate comparison. In other words, one cannot penalize α in the first reduction algorithm, promote α in the second reduction algorithm, and then use η as a basis of comparison between the two algorithms.

An alternative measure for reduction efficacy can be as follows:

$$\eta = \Delta_{LDWR} \times \alpha^{1-\rho}$$

The Δ_{LDWR} represents the difference in the recognizers performance run before and after the reduction. In other words, $\Delta_{LDWR} = LDWR_{after} - LDWR_{before}$ such that $LDWR_{after}$ and $LDWR_{before}$ represent the recognition rate after and before the reduction, respectively. The introduction of Δ_{LDWR} addresses *Case 1*. The accuracy (α) is the base relative to the reduction (ρ) to weight the importance of the accuracy higher than the reduction. Therefore, an increase in accuracy and decrease in reduction is better than an increase in reduction with a decrease in accuracy. This addresses *Case 2*. This new metric also allows the comparison of the same reduced lexicon on two different recognizers. The original equation did not allow this. This has been made possible by removing the need for the empirically chosen constant *k* (this satisfies *Case 3*) as well as previously satisfying *Cases 1 and 2*. Note that a small η number does not imply a poor reduction effectiveness.

It is merely a value used to compare the reduction between different recognizers. A negative η value indicates a drop in recognizer performance due to the lexicon reduction. The equation balances the effectiveness of both the change in recognition performance and the accuracy relative to the reduction. The larger the efficacy value is, the better the effectiveness of the reduction for one recognizer versus another. Note that this does not measure the effectiveness of the recognizers since only the recognizer improvement (Δ_{LDWR}) relative to the reduction is considered.

• *Lexicon Density:* $\varrho_{LDWR}(\mathcal{L}) = (\upsilon_{LDWR}(\mathcal{L}))(f_{LDWR}(n) + \delta_{LDWR})$ [47].
The value ϱ represents the density of the lexicon, with respect to a given recognizer, as described by Govindaraju et al. [47]. A larger density value indicates that the words compared are *similar* or *closer* [47]. The LDWR [66] is the recognizer which operates on a lexicon $\mathcal{L}$ of words $\omega_1 ... \omega_n$. The value δ_{LDWR} is a recognizer independent constant in which *ln 2* (i.e. the natural log defined as $log_e 2$) is used [47].

$$\upsilon_{LDWR}(\mathcal{L}) = \frac{n(n-1)}{\sum_{i \neq j} d_{LDWR}(\omega_i, \omega_j)}$$

The function $\upsilon_{LDWR}(\mathcal{L})$ is defined as the reciprocal average distance between all word pair combinations. The standard function $d_{LDWR}(\omega_i, \omega_j)$ is a recognizer dependent computation used to denote a distance metric between two supplied words essentially measuring the confusion between the words by LDWR [66] [47]. However, even with the dynamic programming step, the process is computationally intensive. Although the LDWR algorithm is a segmentation based algorithm like WR-1 [47], its performance is impacted by the size of the lexicon and therefore it is more complex then WR-1 [47]. The computation of the *slice distance*, which is a comparison between the possible segmentation of the recognizer to lexicon entries is therefore cumbersome. Therefore, we use the string edit distance, also suggested as a natural alternative by Govindaraju et al. [47], and we also introduce the notion of *reduction density* (discussed shortly). The notion of the distance metric is similar to the notion of perplexity in the speech recognition community [6] [47].

The function $f_{LDWR}(n) = ln(n)$ was shown by Govindaraju, et al. [47] to be the most effective.

• *N-Gram Lexicon Distance Metric:* $d_{LDWR}(\omega_i, \omega_j) = \gamma(\omega_i, \omega_j)/\Gamma(\omega_i, \omega_j)$.

The *n-gram lexicon distance metric*, is an alternate distance metric introduced in this research, that is substituted into the Govindaraju et al. [47] lexicon density equation ϱ. This formula computes the lexicon confusion due to the lexicon reduction algorithm. In contrast to making the lexicon density dependent on the recognizer, as done with Govindaraju et al. [47], the dependence is to the lexicon reduction algorithm. Γ denotes the total number of combinations (see definition in Section 5.1.1.3) of uni/bi-gram terms generated between ω_i and ω_j. The value γ represents the number of uni/bi-gram terms that are *not* common between ω_i and ω_j. This keeps the equations compatible with the reciprocal. It allows the density function, using this alternative distance metric, to be computed on both a complete lexicon and a reduced lexicon showing the similarity of uni/bi-gram term occurrences within the lexicon. Note that since the recognizer takes only a lexicon of words as an input and computes its own distance information, only the NSI (NSI denotes "No Spatial Information" used as the encoding procedure for the uni/bi-gram terms; see details in Section 5.1.1.3) term encodings are used. Since the lexicon reduction involves the ESI (ESI denotes "Exact Spatial Information" used as the encoding procedure for the uni/bi-gram terms; see details in Section 5.1.1.3) term encodings, the application of this *n-gram lexicon distance metric* metric only provides an approximation. The computation of γ involves two steps: (i) generating the NSI encodings for ω_i and determining their occurence in ω_j and, (ii) the same step applied by reversing the ω values. These two values are averaged and then returned as the value of γ.

In order to distinguish between the *lexicon density distance metric* and the *n-gram lexicon distance metric* equations, the values ϱ' and ϱ'' will be respectively used. Using the two equations, the confusion among lexicon words and uni-bi-gram terms can be shown.

This is important since it is likely that the LDWR [66] will produce the same high character confidence scores used for anchor points, for a word as originally computed during the uni/bi-gram term recognition extraction. Although likely, this situation is still not guaranteed since the choice of the words in the reduced lexicon is different from the complete lexicon.

4.5 Discussion

The performance measures in this section provide insight into the effectiveness of the lexicon reduction algortihm. The reduction of the lexicon improves recognition performance only if the unknown word remains in the lexicon afterwards. In addition, if the lexicon is too dense (i.e. the words are too similar), then it is also possible that the recognizer will select a word which is geometrically similar (e.g. *maximize* and *minimize* may be considered similar). Therefore, an increase lexicon density indicates the recognition can drop as well. The formulas in this section will be used when discussing the results of any lexicon reduction.

Chapter 5

Topic Categorization

5.1 Proposed Algorithm

Topic categorization in this research is the process of assigning categories representing the topic content of the form. This challenge is similar to the call routing problem. In the call routing problem, researchers at Lucent Technologies took voice recognition information as an input and produced the call destination as an output [21] [22]. Here, we take characters with the highest recognition as an input and produce higher level anatomical categories. Both problems receive recognition data as an output and produce a topic as output.

The road map in Figure 5.1 illustrates the layout of the proposed algorithm. This is broken up into three areas: (i) training, (ii) recognition, and (iii) retrieval. A knowledge base is constructed during the *training phase* from a set of PCR forms. This contains the relationships between terms and categories that are used by the other two areas. The *recognition phase* takes an unknown form, and reduces the lexicon using the knowledge base. This phase is evaluated using a separate testing deck. Finally, after all content of the PCR form has been recognized, a search can take place by entering in a query. This phase is tested by querying the system with a deck of phrase inputs. The forms are then ranked accordingly and returned to the user.

In the training phase, a mechanism for relating uni-grams and bi-grams (henceforth: uni/bi-grams discussed in the following section) as well as categories from a PCR training deck are constructed. The testing phase then evaluates the algorithm's ability to determine the category from a test form by using a Lexicon Driven Word Recognizer (LDWR) [66] to extract the top-choice uni/bi-gram characters from all words. A maximum of two characters per word is considered, since LDWR [66] successfully extracts a bi-gram with spatial encoding information 40% of the time. If ≥ 3 characters are selected, then the LDWR [66] will only successfully extract a character $\leq$ 1% of the time. Hence the limit of two was selected (see examples in Figure 5.4).

5.1.1 Training

The training stage involves a series of steps to construct a matrix that represents relationships between terms and categories. Recall that each form can have up to five categories. In the first phase, lexicons are constructed using all the words from all forms under a category. In the second phase, phrases are extracted from the form using a cohesion equation. These phrases are then converted to ESI encoding terms (ESI denotes "Exact Spatial Information" used as the encoding procedure for the uni/bi-gram terms; see details in Section 5.1.1.3). A matrix is then constructed utilizing the ESI terms for the rows and the categories in the columns. The matrix is then normalized, weighted, and prepared in Singular Value Decomposition format.

5.1.1.1 Filtering

Stopwords are those words that are not used for category determination in this application. The list of $\sim$400 stopwords provided by PubMed are omitted from the lexicon [95]

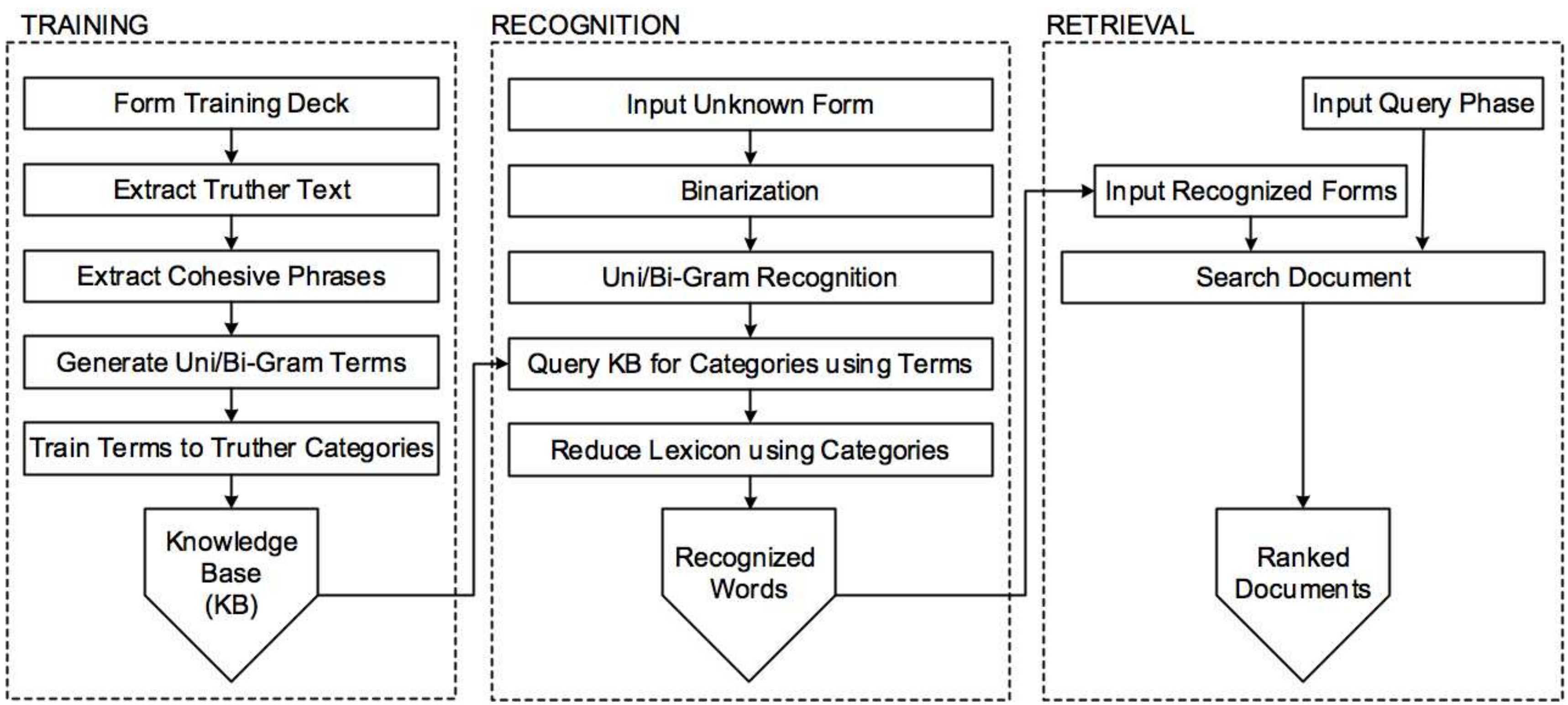


Figure 5.1: Proposed Algorithm Road Map

[54]. An additional list of $\sim$50 words (e.g. patient, staff, nurse, etc.) found in most PCR's, which have little bearing on the category, are omitted from the cohesion analysis (the frequency of two words co-occurring versus occurring independently; see Equation 5.1.1) but retained in the final lexicon. It is common to apply other filters to reduce the likelihood of morphological mismatches [54]. However, strategies such as 'stemming' [54] cannot be applied before recognition because the text is not yet ASCII and is therefore unknown. Consider a handwritten word image representing "rhythms" that needs to be recognized. The alteration of "rhythms" to "rhythm" in the lexicon will affect recognition performance. However, at the end of classification, these words are considered equivalent. Therefore, word stemming is applied after the LDWR [66] has determined the ASCII word translation.

5.1.1.2 Phrase Construction

A phrase is defined as a sequence of adjacent non-stopwords found in [37]. Although an empirical study in Fagan [37] indicates that important phrases may wrap around stopwords [37], the inclusion of stopwords degrades performance in the training experiments here. Furthermore, since longer sequences of words as well as longer sentences have been shown to be more successful than shorter contingent words [37], phrases are computed within the text area of a single PCR region utilizing a natural language cohesion technique used by Fagan [37] [54].

A passage P is the set of all words w for a PCR form under a category C treated as a single string. For each C, every pair of passages, denoted P_1 and P_2, is compared. Here we denote w_x as a word located at position *x* within a passage P. If $w_a \in P_1, w_a^{'} \in P_2, w_b \in P_1, w_b^{'} \in P_2$ such that $b^{'} > a^{'}$ and $b > a$, then a potential phrase consisting of exactly two words is constructed. The cohesion of phrases under each C is then computed. If the cohesion is above a threshold, then that phrase represents that category C. Thus a category C is represented by a sequence of high cohesion phrases using only those PCR passages

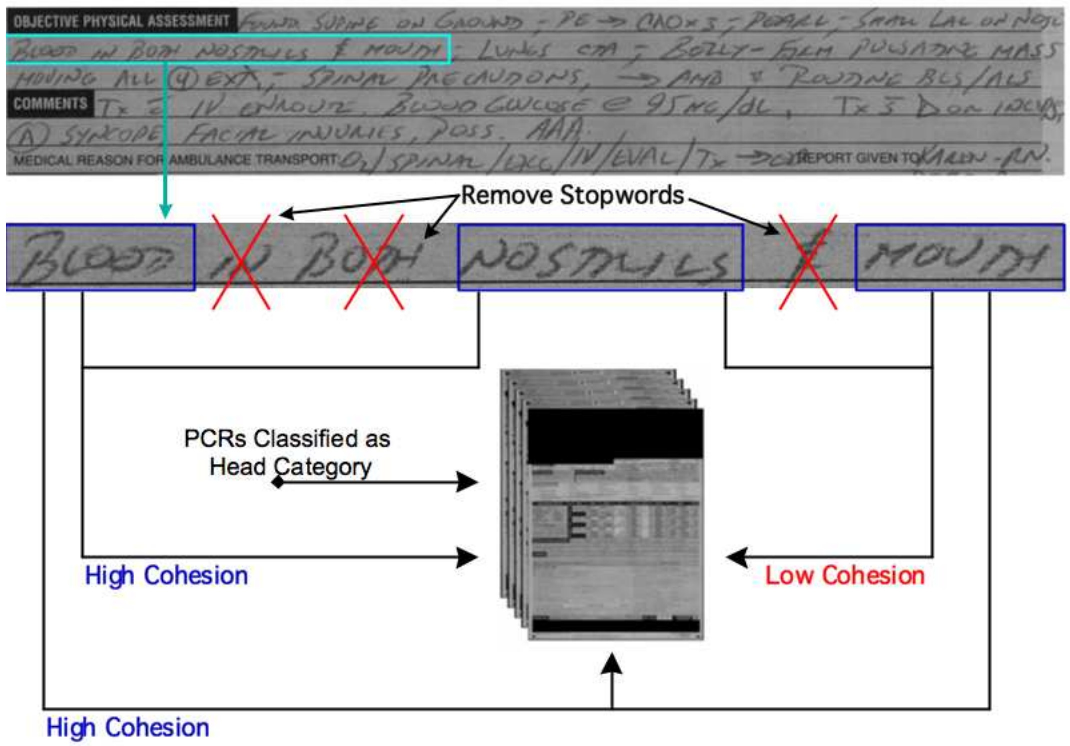


Figure 5.2: Phrase Candidate (High vs. Low Cohesion)

manually categorized under C.

$$cohesion(w_a, w_b) = z \bullet \frac{f(w_a, w_b)}{\sqrt{f(w_a)f(w_b)}} \tag{5.1.1}$$

The cohesion between any two words w_a and w_b is computed by the frequency that w_a and w_b occur together versus existing independently. The top 40 cohesive phrases are retained for each category (see Equation 5.1.1).

Consider the following two unfiltered text phrases S_1 and S_2 under the category *legs*:

S_1: “right femur fracture”

S_2: “broken right tibia and femur”

The candidate phrases CP_1 and CP_2 after the filtering step are:

CP_1: “right femur” ... “right fracture” ... “femur fracture”

CP_2: “broken right” ... “right femur” ...

The phrase “right femur” is computed from CP_1 and CP_2, since w_a and $w_a^{'}$ = “right”, w_b and $w_b^{'}$ = “femur”, and the conditions $b > a$ and $b' > a'$ have been met. If the cohesion for “right femur” is above the threshold across all PCR forms under the *legs* category, then this phrase is retained as a representative of the category *legs*.

Tables 5.1 and 5.2 illustrate some top choice cohesive phrases generated. Notice that digestive system and pelvic region are anatomically *close*. However, different information is reported in these two cases resulting in mostly different cohesive phrases. The phrase *CHEST PAIN* occurs in both categories, however, have different cohesion values. This implies that the term frequencies will also likely be different and therefore commonly occurring terms need to be weighted appropriately to their category (this will be discussed in more detail in Section 5.1.1.5). Phrases sometimes may not make sense by themselves; however, this is the result of using a cohesive phrase formula in which words may not be adjacent.

FREQUENCY	**COHESION**	**PHRASE**
6	0.67	DCAP BTLS
166	0.35	CHEST PAIN
91	0.38	PAIN 0
1860	2.49	PAIN HIP
144	0.34	HIP JVD
112	0.39	PAIN CHANGE
275	0.81	HIP FX
110	0.37	HIP CHANGE
82	0.38	PAIN 10
163	0.40	JVD PAIN
106	0.40	CAOX3 PAIN
202	0.50	PAIN JVD
213	0.55	PAIN LEG
205	0.42	CHEST HIP
3	0.33	PERPENDICULAR DECREASE
121	0.33	FELL HIP
118	0.36	PAIN FX
2251	3.01	HIP PAIN
390	0.83	PAIN CHEST
288	0.59	HIP CHEST

Table 5.1: Top Cohesive Phrases for the Category: Pelvis

FREQUENCY	COHESION	PHRASE
30	0.72	PAIN INCIDENT
5	0.31	PAIN TRANSPORTED
42	0.54	PAIN CHEST
52	0.81	STOMACH PAIN
9	0.25	HOME PAIN
6	0.43	VOMITING ILLNESS
39	0.51	CHEST PAIN
4	0.24	CHEST SOFT
25	0.54	PAIN SBM
31	0.37	PAIN X4
31	0.47	PAIN JVD
11	0.34	PAIN EDEMA
25	0.44	PAIN PMSX4
6	0.21	PAIN SOFT
3	0.21	SBM INCIDENT
11	0.25	PAIN LEFT

Table 5.2: Top Cohesive Phrases for the Category: *Digestive System*

5.1.1.3 Term Extraction

There are three term encoding formats: NSI, ESI and ASI. Terms of a particular encoding will later be associated with an anatomical category and used as the essential criterion for lexicon reduction.

No Spatial Information (NSI):

An asterisk (*) indicates that zero or more characters are found between C_1 and C_2. NSI encodings are the most simple form of encoding (see Figure 5.4 examples).

UNI-GRAM ENCODING: $*C*$

BI-GRAM ENCODING: $*C_1 * C_2*$

BI-GRAM ENCODING EXAMPLE: BLOOD → *L*D*

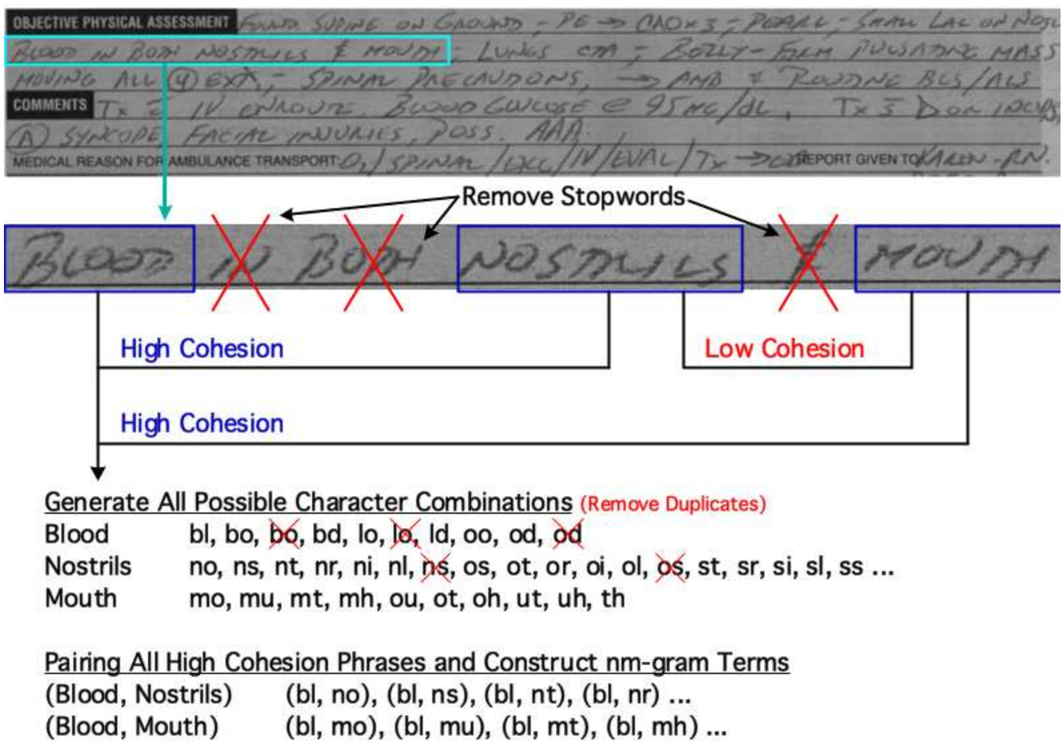


Figure 5.3: Term Extraction from High Cohesive Phrases

Exact Spatial Information (ESI):

The integers (x, y, z) represent the precise number of characters between C_1 and C_2. ESI encodings are an extension of the NSI encodings with the inclusion of precise spatial information. In other words, the number of characters before, after and between the highest confidence C_1 and C_2 characters are part of the encoding. These encodings are the most successful since there are fewer term collisions involved. Hence the ESI encodings are preferred.

UNI-GRAM ENCODING: xCy

BI-GRAM ENCODING: xC_1yC_2z

BI-GRAM ENCODING EXAMPLE: BLOOD → 1L2D0

Approximate Spatial Information (ASI):

The integers (x_a, y_a, z_a), denoted as length codes, represent an estimated range of characters between C_1 and C_2. A '0' indicates no characters, a '1' indicates between one and two characters, and a '2' represents greater than 2 characters. The ASI encodings are an approximation of ESI encodings designed to handle the case that the precise number of characters may not be known with high confidence.

UNI-GRAM ENCODING: x_aCy_a

BI-GRAM ENCODING: $x_aC_1y_aC_2z_a$

BI-GRAM ENCODING EXAMPLE: BLOOD → 1L1D0

Combinatorial Analysis

The quantity of all possible NSI, ESI and ASI uni-gram and bi-gram combinations, for a given word of character length n, such that $n \geq 1$, is represented by the mathematical series of Equation 5.1.2. Regardless of the encoding, the same quantity of combinations exists since the distance between characters is known.

$$\mathcal{C}(n) = \left(\left(\sum_{i=1}^{n-1} (n-i) \right) + n \right) = \left(\left(\left(\frac{n}{2} \right) (n-1) \right) + n \right) \tag{5.1.2}$$

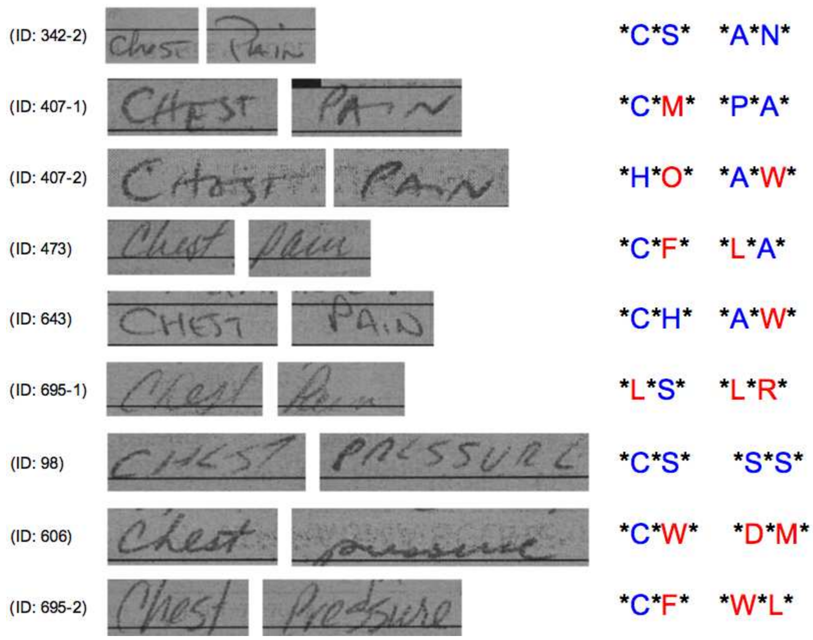


Figure 5.4: NSI Encodings Example (Blue Letters: LDWR[66] successfully extracted)

However, the function $\mathcal{C}$ only considers the combinations of an individual entry. The combination inflation of a uni/bi-gram phrase is shown by Equation 5.1.3. The equation parameters *a* and *b* represent the string lengths of the words considered in a phrase.

$$\mathcal{P}(a, b) = \mathcal{C}(a) \cdot \mathcal{C}(b) \tag{5.1.3}$$

For example:

Let the phrase to evaluate uni/bi-gram combinations be *PULMONARY DISEASE*.

Let n = length(“PULMONARY”) = 9

Let m = length(“DISEASE”) = 7

$\mathcal{C}$(n) = 45 uni-gram + bi-gram combinations for “PULMONARY”

$\mathcal{C}$(m) = 28 uni-gram + bi-gram combinations for “DISEASE”

$\mathcal{P}$(n,m) = 1,260 uni-gram + bi-gram phrase combinations for *PULMONARY DISEASE*

Each of these encodings has its advantages and disadvantages. The choice is ultimately based on the quality of a handwriting recognizer’s ability to extract characters. If a handwriting recognizer cannot successfully extract positional information, then NSI is the best approach. If extraction of positional information is reliable, then the ESI is the best approach. However, NSI and ASI create more possibilities for confusion since distances are either approximated or omitted. ESI is more restrictive on the possibilities as the precise spacing is used, leading to lesser confusion among terms.

Using the ESI protocol, all possible uni/bi-gram terms are synthetically extracted from each cohesive phrase under each category. For example, BLOOD can be encoded to the uni-gram 0B4 (zero characters before 'B' and four characters after 'B') and the bi-gram 0B3D0 (zero characters before 'B', three characters between 'B' and 'D' and zero characters following 'D'). All possible synthetic positional encodings are generated for each phrase and chained together (a '$' is used to denote a chained phrase). For example, CHEST PAIN encodes to: 0C4$0P0A2 ... 0C4$1A2 ... 0C0H3$0P1I1 ... 0C0H3$0P2N0, etc. Therefore, each category now has a list of encoded phrases consisting of positional encoded uni/bi-grams. These terms are the most primitive representative links to the category used throughout the training process. In the training phase, the synthetic information can be extracted since the text is known. However, in the testing phase, a recognizer will be used to automatically produce an ESI encoding since the test text is not known. To improve readability, the notation (W_1, W_2) is used to represent an ESI encoding of a two-word phrase (e.g. Myocardial Infarction: (my, in), (my, if), (my, ia), etc ...).

5.1.1.4 Term-Category Matrix Construction

A matrix A, of size $|T|$ by $|C|$, is constructed such that the rows of the matrix represent the set of terms T, and the columns of the matrix represent the set of categories C. The value at matrix coordinate (t,c) is the frequency that each term is associated with the category. The term frequency corresponds to the phrasal frequency from which it was derived. It is the same value as the numerator in the cohesion formula (refer to Equation 5.1.1):

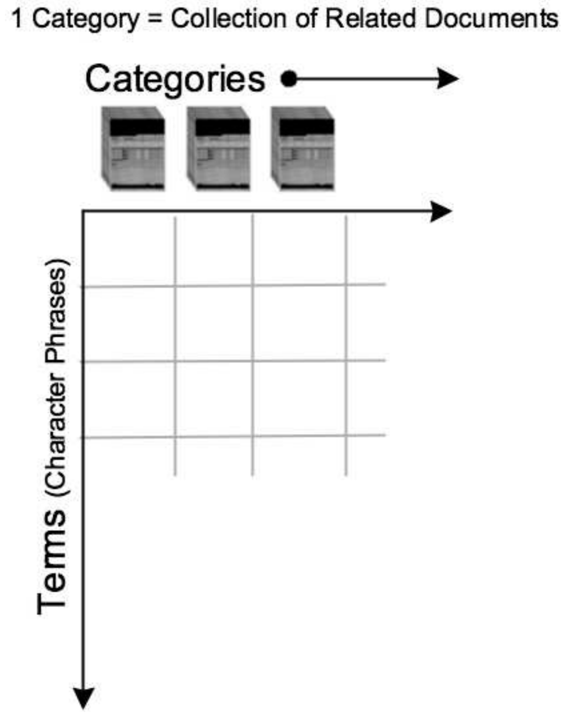


Figure 5.5: Term Category Matrix (TCM) Overview

$f(w_a, w_b)$. For example, if the frequency of CHEST PAIN is 50, then all term encodings generated from CHEST PAIN, such as (ch, pa), will also receive a frequency of 50 in the matrix.

Step 1: Compute the normalized matrix B from A using Equation 5.1.4 [21] [22]:

$$B_{t,c} = \frac{A_{t,c}}{\sqrt{\sum_{e=1}^{n} A_{t,e}^2}} \tag{5.1.4}$$

Matrix A is the input matrix containing raw frequencies, Matrix B is the output matrix with normalized frequencies, and (t,c) is a (term, category) coordinate within a matrix.

Step 2: Term Discrimination Ability

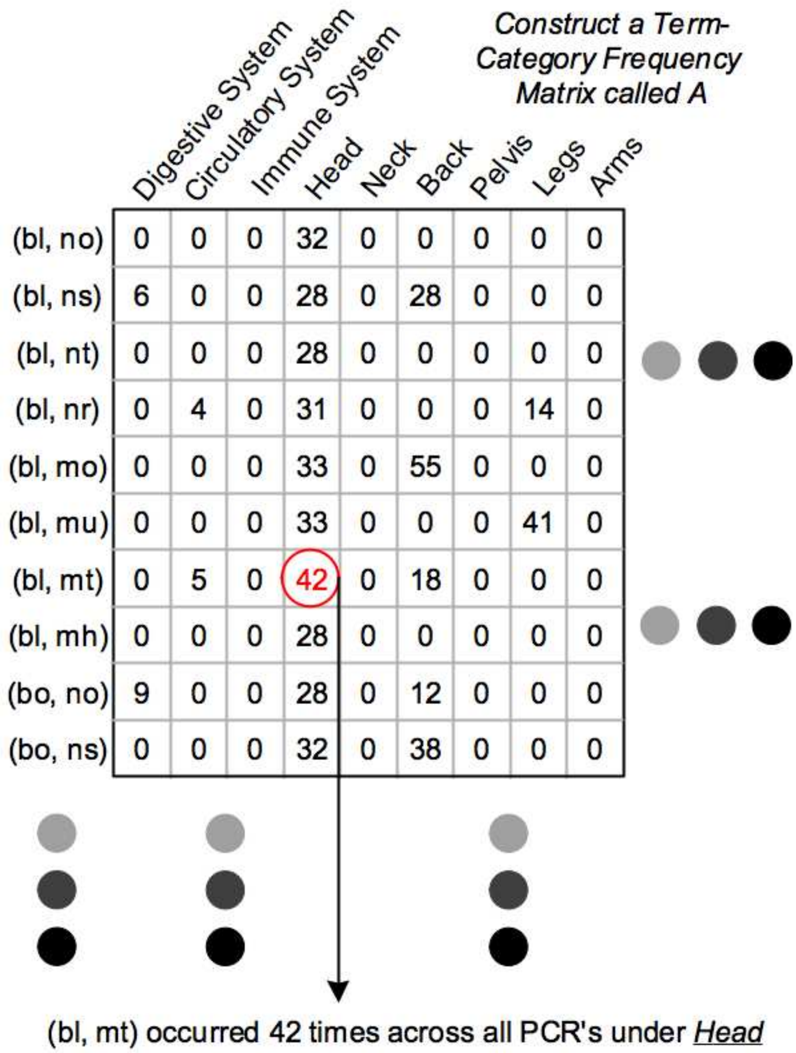


Figure 5.6: TCM Frequency Construction Example

The Term Frequency times Inverse Document Frequency (TF x IDF) weighting approach is used to favor those terms which occur frequently with a small number of categories as opposed to their existence in all categories [78] [109]. While Luhn [78] asserts that medium frequency terms would best resolve a document, it precludes classification of more rare medical words. Salton's [109] theory, stating that terms with the most discriminatory power are associated with fewer documents, allows a rare-medium frequent word to resolve the document.

STEP 2A Compute the weighted matrix X from B using Equation 5.1.5 [21] [22] [54]:

$$IDF(t) = log_2 \frac{n}{c(t)} \tag{5.1.5}$$

IDF computes the inverse-document-frequency on term t, and c(t) is the number of categories containing term t.

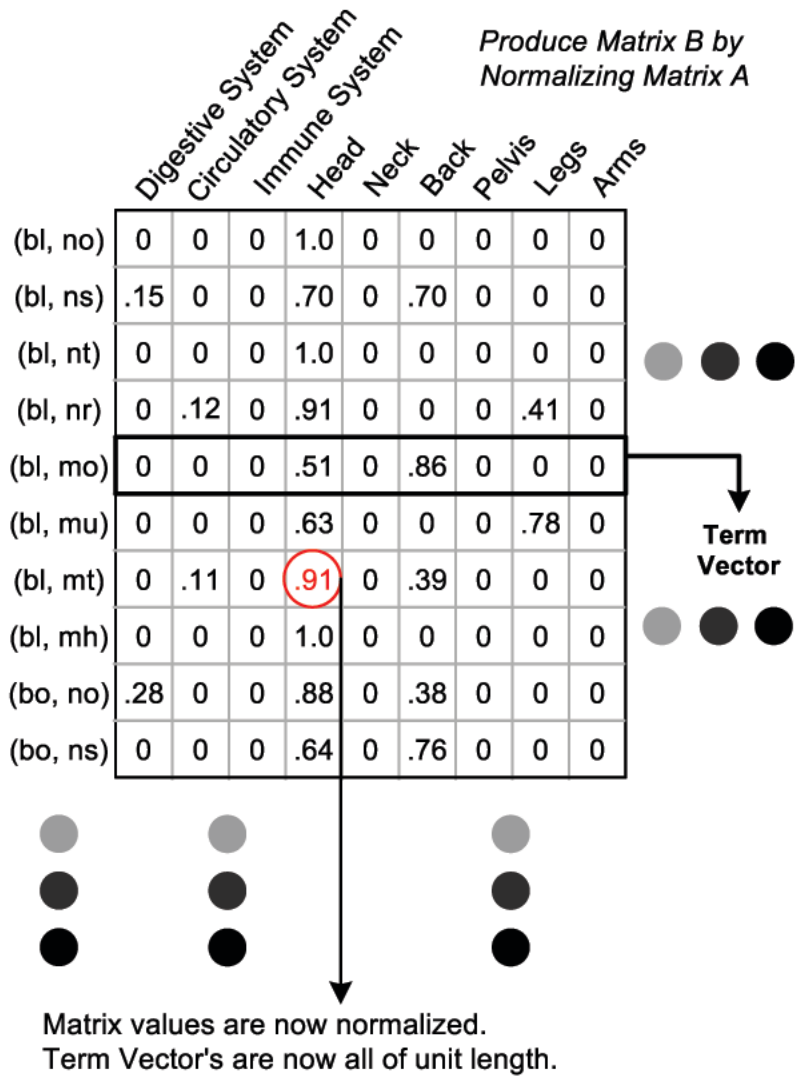
Step 2B Weight the normalized matrix by IDF values using Equation 5.1.6 [21] [22] [61] [54]:

$$X_{t,c} = IDF(t) \cdot B_{t,c} \tag{5.1.6}$$

Matrix B is the normalized matrix from Step 1, IDF is the computational step defined in Step 2, and Matrix X is a normalized and weighted matrix.

5.1.1.5 Reduced Singular Value Decomposition (R-SVD) [34]

The normalized and weighted term-category matrix can now be used as the knowledge base for subsequent classification. A singular value decomposition variant, which incorporates a dimensionality reduction step allows a large term-category matrix to represent the PCR training set (see Equation 5.1.7). This facilitates a category query from an unknown



	Digestive System	Circulatory System	Immune System	Head	Neck	Back	Pelvis	Legs	Arms
(bl, no)	0	0	0	1.0	0	0	0	0	0
(bl, ns)	.15	0	0	.70	0	.70	0	0	0
(bl, nt)	0	0	0	1.0	0	0	0	0	0
(bl, nr)	0	.12	0	.91	0	0	0	.41	0
(bl, mo)	0	0	0	.51	0	.86	0	0	0
(bl, mu)	0	0	0	.63	0	0	0	.78	0
(bl, mt)	0	.11	0	.91	0	.39	0	0	0
(bl, mh)	0	0	0	1.0	0	0	0	0	0
(bo, no)	.28	0	0	.88	0	.38	0	0	0
(bo, ns)	0	0	0	.64	0	.76	0	0	0

Figure 5.7: TCP Normalization

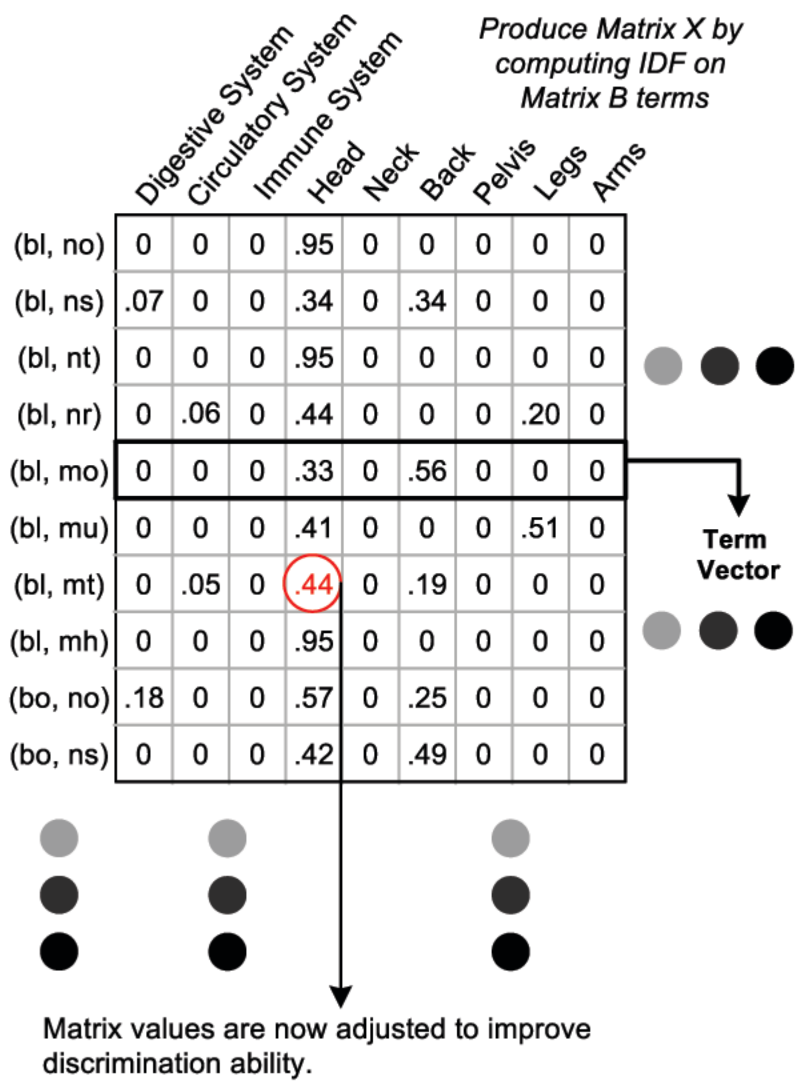


Figure 5.8: TCM Inverse Document Frequency (IDF)

PCR using the LDWR [66] determined terms [21] [22] [34].

$$X = U \bullet S \bullet V^T \tag{5.1.7}$$

Matrix X is decomposed into 3 matrices: U is a (T x k) matrix representing term vectors, S is a (k x k) matrix, and V is a (k x C) matrix representing the category vectors.

The value k represents the number of dimensions to be finally retained. If k equals the targeted number of categories to model, then SVD is performed without the reduction step. Therefore, in order to reduce the dimensionality, the condition $k < |C|$ is necessary to reduce noise [34].

5.1.2 Testing

Given an unknown PCR form, the task is to determine the category of the form, and use the reduced lexicon associated with the determined category to drive the LDWR [66]. In addition, the category determined can be used to tag the form which can be subsequently used for information retrieval. The query task is divided into the following steps: (i) Term Extraction, (ii) Pseudo-Category Generation, and (iii) Candidate Category Selection [21] [22].

5.1.2.1 Term Extraction

Given a new PCR image, all image words are extracted from the form, and the LDWR [66] is used to produce a list of confidently recognized characters for each word. These are used to encode the positional uni/bi-grams consistent with the format during training. All combinations of uni/bi-phrases in the PCR form are constructed. Each word has exactly one uni-gram and exactly one bi-gram. A phrase consists of exactly two unknown words. Therefore it is represented by precisely four uni/bi-phrases (BI-BI, BI-UNI, UNI-BI and

UNI-UNI).

5.1.2.2 Pseudo-Category Generation

An *(m x 1)* query vector Q is derived, which is then populated with the term frequencies for the generated sequences from the Term-Extraction step. If a term was not encountered in the training set, then it is not considered. Positional bi-grams are generated to yield the trained terms 37% of the time, and similarly positional uni-grams 57% of the time. The experiments here illustrate this to be a sufficient number of terms. A scaled vector representation of Q is then produced by multiplying Q^T and U.

5.1.2.3 Reduced Singular Value Decomposition (R-SVD)

Once the pseudo-category is derived, R-SVD is applied for the following reasons: (i) It converts the query into a vector space compatible input, and (ii) the dimensional reduction can help reduce noise [34]. Since the relationship between terms and categories is scaled by variance, the reduction allows parametric removal of less significant term-category relationships.

5.1.2.4 Candidate Category Selection

The task is now to compare the pseudo-category vector Q with each vector in $V_r \bullet S_r$ (from the training phase) using a scoring mechanism. The cosine score is used for matching the query [21] [22]. Both *x* and *y* are dimensional vectors used to compute the cosine in Equation 5.1.8. Vectors *x* and *y* in the equations represent the comparison of the vectors: pseudo-category Q to every column vector in $V_r \bullet S_r$.

$$z = cos(x, y) = \frac{x \cdot y^T}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}} \tag{5.1.8}$$

	Digestive System	Circulatory System	Immune System	Head	Neck	Back	Pelvis	Legs	Arms	Pseudo-Category Vector
(bl, no)	0	0	0	.95	0	0	0	0	0	0
(bl, ns)	.07	0	0	.34	0	.34	0	0	0	2
(bl, nt)	0	0	0	.95	0	0	0	0	0	1
(bl, nr)	0	.06	0	.44	0	0	0	.20	0	1
(bl, mo)	0	0	0	.33	0	.56	0	0	0	1
(bl, mu)	0	0	0	.41	0	0	0	.51	0	0
(bl, mt)	0	.05	0	.44	0	.19	0	0	0	2
(bl, mh)	0	0	0	.95	0	0	0	0	0	0
(bo, no)	.18	0	0	.57	0	.25	0	0	0	3
(bo, ns)	0	0	0	.42	0	.49	0	0	0	1

Vector Q is now added to the original normalized/weighted matrix X

Figure 5.9: Pseudo-Category Vector

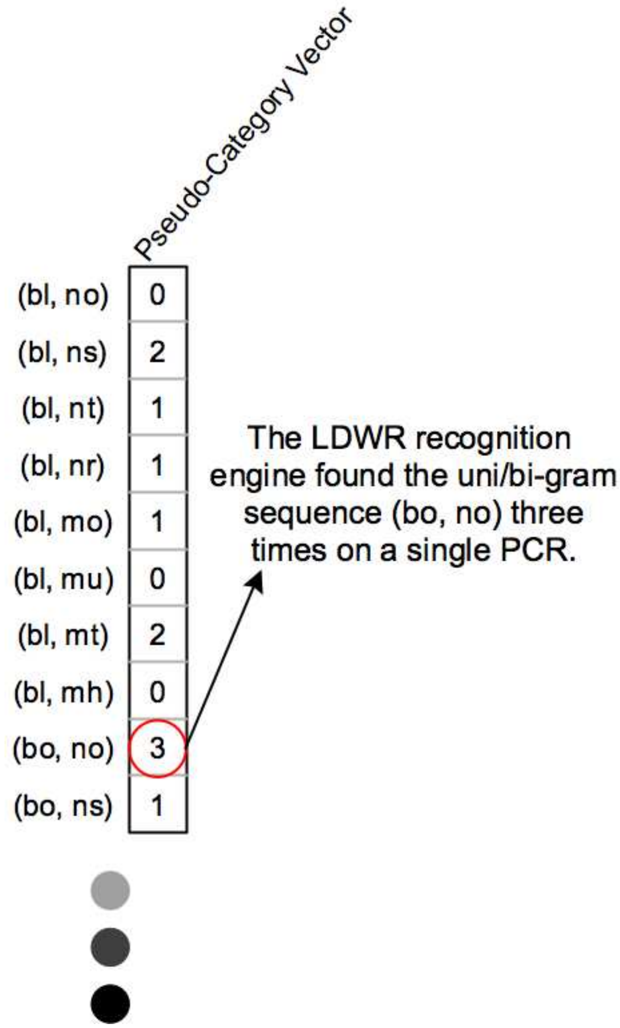


Figure 5.10: Pseudo-Category Integration

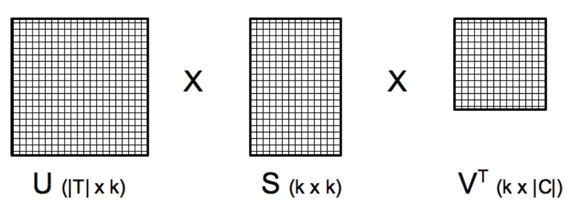


Figure 5.11: Matrix Decomposition Visual

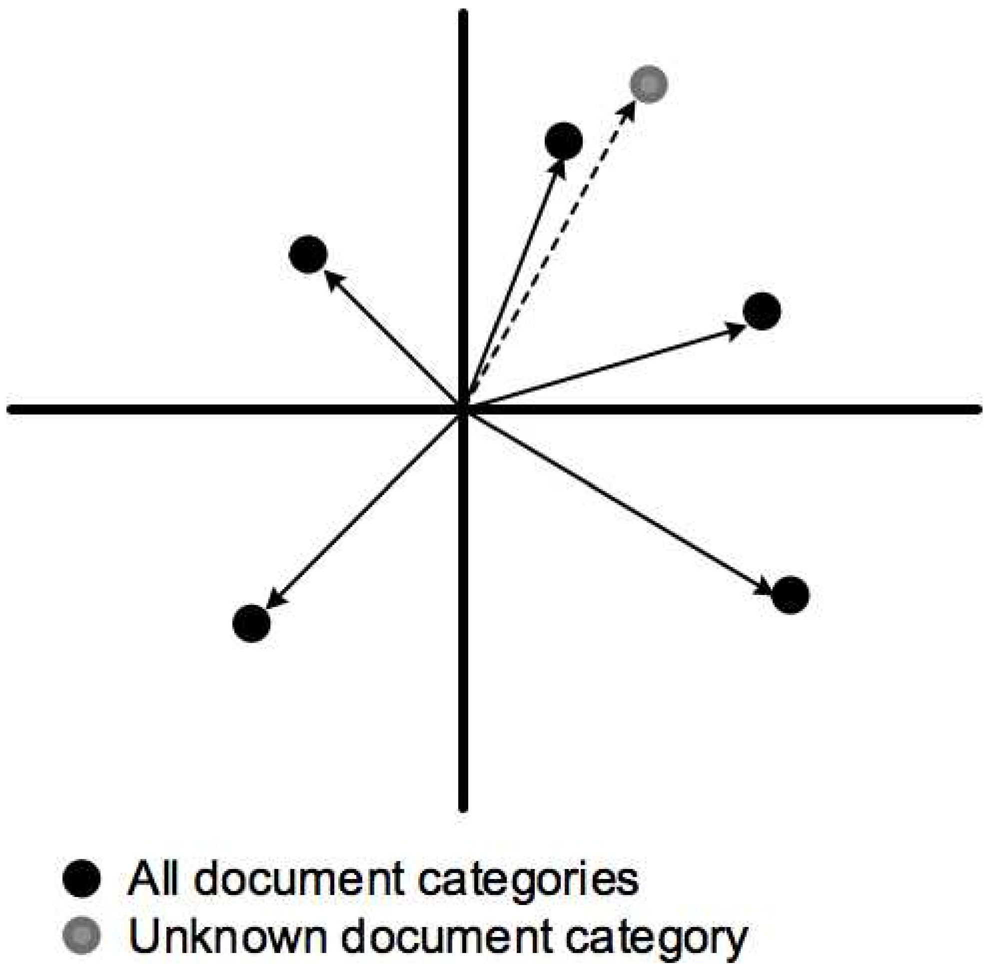


Figure 5.12: Simplified Multi-Dimensional Category Vector Space

Each cosine score is mapped onto a sigmoid function using the least square fitting method, thereby producing a more accurate confidence score [21] [22]. The least squares regression line equations used to satisfy the equation f(x) = ax + b are shown in Equations 5.1.9 and 5.1.10 [73]:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \tag{5.1.9}$$

$$b = \frac{1}{n}(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i) \tag{5.1.10}$$

The fitted sigmoid confidence is produced using the cosine score and the regression line, using equation (9):

$$confidence(a, b, z) = \frac{1}{1 + e^{-(az+b)}} \tag{5.1.11}$$

The confidence scores, one for each category, are then used to rank the categories. If a category is above an empirically chosen threshold, then that category is retained for the PCR. Multiple categories may be thus retained. All words corresponding to the selected categories are then used to construct a new lexicon which is finally submitted to the LDWR recognizer [66]. Given a test PCR form, and the reduced lexicon, the LDWR [66] converts the handwritten medical words in the form to ASCII.

5.1.2.5 Result

Each word which is recognized is compared with the truth. However, a simple string comparison is insufficient due to spelling mistakes and root variations of word forms which are semantically identical. This occurs 20% of the time within the test deck words. Therefore, a Porter stemming [62] [100] [105] and a Levenshtein String Edit Distance [13] of 1 allowable penalty are performed on both the truth and the recognizer result before they are compared. Levenshtein is only applied to a word that is believed to be ≥ 4 characters in length. For example, PAIN and PAINS are identical. However, this also results in an improper comparison in $\sim$11% of the corrections (see Table 5.3).

5.1.3 Time Complexity

Given a series of steps involving the lexicon reduction model, this section breaks down the individual run-times in worst case O-notation. The evaluation of training an entire knowledge base for the recognition of a single word is tabulated here. These complexities assume that the handwritten word image has been extracted, binarized and pre-processed.

FIGHT vs EIGHT vs LIGHT	FINE vs FIRE
MEDICAL vs MEDICATION	FOOD vs FOOT
1400 vs 2400	LEFT vs LIFT
BAIL vs RAIL	MOANING vs MORNING
BALL vs CALL	MARK vs MARY
MOLE vs MOVE	PUNCH vs LUNCH
CALF vs CALL	REACH vs REACT
CARD vs CARE vs CART	SCARE vs CARE
COLD vs TOLD	SEVER vs FEVER
NECK vs DECK	STABLE vs TABLE
FALL vs CALL	FEET vs FEED
FOUND vs BOUND vs SOUND vs POUND	

Table 5.3: Word Collisions

5.1.3.1 Training Complexity

The training performance is based on the summation of the following complexities:

• *Filtering:* $O(|l|)$ such that *l* represents the complete lexicon.

• *Cohesive Phrase Construction:* $O(\Sigma_{\forall i}|P_i|^2)$ such that P_i represents a paragraph of text.

• *Term Extraction:* $O(\Sigma_{\forall(i,j)p_i \in C_j} a_i^2 b_i^2)$ such that a_i = $|\omega_1^{'}|$ and b_i = $|\omega_2^{''}|$ where $\omega_i^{'}$ and $\omega_i^{''}$ represent the first and second words from the cohesive phrase p_i under category C_j.

• *Matrix Normalization and Weighting:* $(2 \times O(mn))$ such that *m* is the quantity of terms and *n* is quantity of categories.

Therefore total training run-time is:

$$O(|l|) + O(\Sigma_{\forall i}|P_i|^2) + O(\Sigma_{\forall(i,j)p_i \in C_j} a_i^2 b_i^2) + (2 \times O(mn))$$

5.1.3.2 Recognition Complexity

The recognition performance is based on the summation of the following complexities:

• *Term Extraction:* $O(\Sigma_{\forall i} g(h_i))$ such that $g(h_i)$ represents the run-time of the LDWR [66] on the handwritten image h_i where *i* denotes the index of each word image on a single PCR form.

• *Pseudo-Category Generation:* $O(m)$ such that *m* represents the quantity of possible terms in which the frequencies are tallied.

• *Singular Value Decomposition:* $O(mn^2 + m^2n)$ such that *m* is the quantity of terms and *n* is quantity of categories [44].

• *Cosine Score Computation:* $O(m) \times O(mn)$ such that a single vector of values $O(m)$ is compared against the matrix $O(mn)$ values.

• *Regression Computation:* $O(m) \times O(mn)$ such that a single vector of values $O(m)$ is compared against the matrix $O(mn)$ values.

• *Sigmoid Mapping:* $O(m) \times O(mn)$ such that a single vector of values $O(m)$ is compared against the matrix $O(mn)$ values.

Therefore total recognition run-time is:

$$O(\Sigma_{\forall i} g(h_i)) + O(m) + O(mn^2 + m^2n) + (3 \times (O(m) \times O(mn)))$$

5.2 Artificial Neural Network Implementation

Initially, the model used for relating terms to categories was a backpropagation artificial neural network (ANN) [88]. The training procedure involved the NSI encodings constructed from adjacent handwritten word images using the LDWR [66]. This encoding was then formatted to a bit string compatible input layer and trained against the neural network by setting the output layer to the target categories (see Figure 5.13). During the recognition phase the ANN would be queried using the terms for the top choice categories (see Figure 5.14). The lexicon would then be constructed from those top choice categories, in the same way that it had been done with the SVD approach, and then submitted to the LDWR [66] for handwriting recognition.

Unfortunately, the ANN approach ran into scalability issues hindering performance. The ANN could only be computed on a training deck of 30 PCR's and a test deck of 10 PCR's. A lexicon of approximately 800 words was reduced to 92% resulting in $\leq$ 3% improvement to recognition with a momentum of 0.1 and learning rate of 0.3. This approach was chosen for two reasons: (i) the ANN's ability to compute a non-linear decision surface which was expected due to the overlapping nature of terms under a category, and (ii) the ANN's resistance to noise which can be controlled to some extent by using the momentum constant to escape local minima in the hypothesis space. While the ANN implementation may perform well on a larger data set, the actual performance is unknown due to computational limitations.

5.3 Summary

The detailed mathematical and algorithmic structure for the lexicon reduction model has been presented. The latent semantic indexing model is more scaleable than the artificial neural network model. The training phase involves the extraction of cohesive phrases under each anatomical cateogory from a training deck. The terms are extracted from the cohesive phrases and then mapped against the categories. During the testing phase, the LDWR [66] is used to extract the highest confident character information and query the latent semantic model for the highest confident categories. The reduced lexicon is then constructed from these categories and provided to the LDWR [66] for a second interpretation of the input word image. The next section will compare the effectiveness of the recognition before and after the reduction.

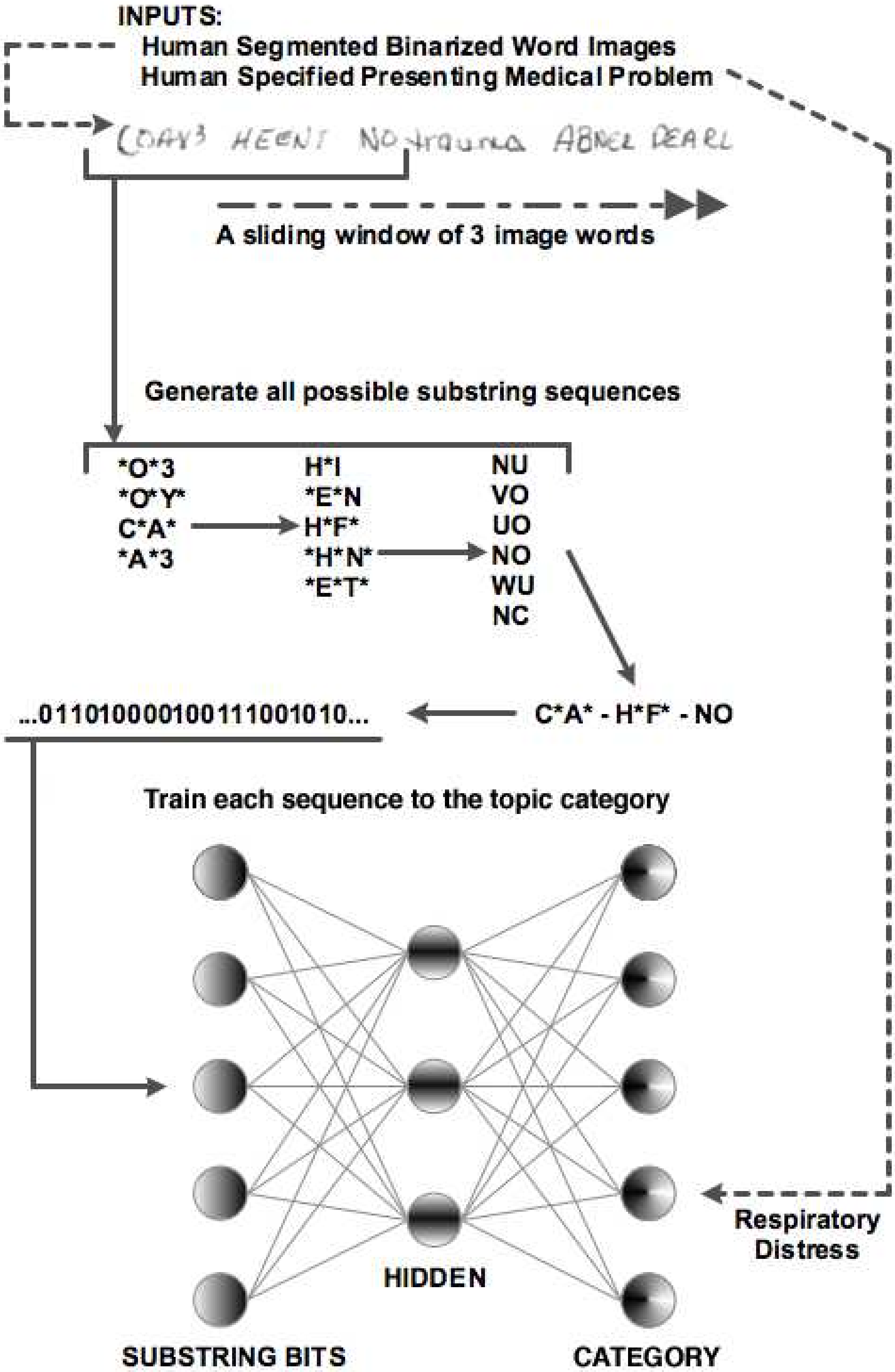


Figure 5.13: ANN Training

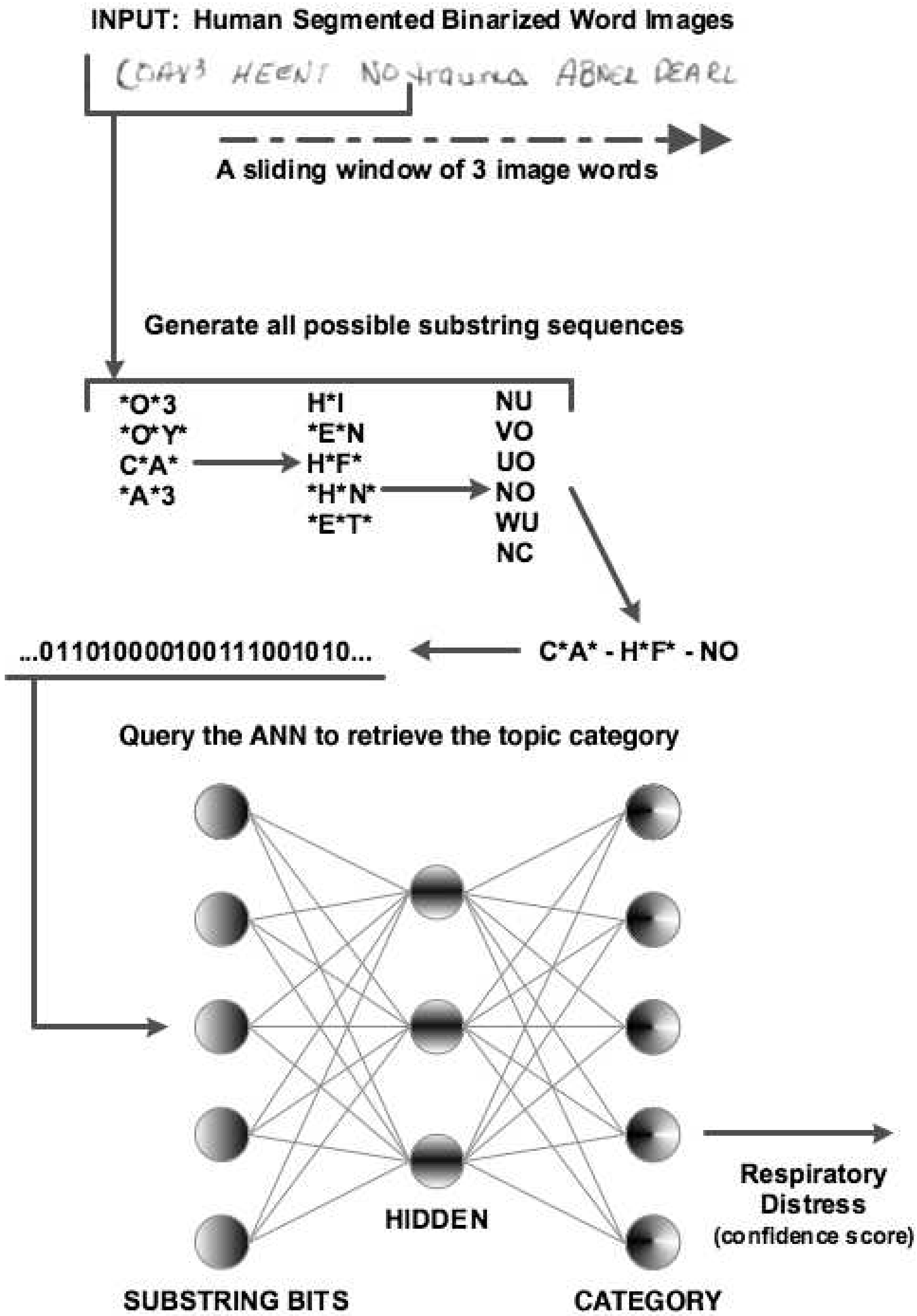


Figure 5.14: ANN Recognition

Chapter 6

Recognition Experiments and Results

In this chapter, the results of several experiments illustrate the effectiveness of our algorithm. Accept rate, error rate, and raw rates are reported for several experiments found in Table 6.1. Improvements to recognition rates and error rates are reported in Table 6.2. The effectiveness of the reduced lexicon is found in Table 6.3. A description of the training and test decks can be found in Table 6.4. The correct ranking of categories by the lexicon reduction algorithm is shown in Figure 6.1. The breakdown of lexicon sizes by category can be found in Figure 6.2.

	CL	**CLT**	**AL**	**ALT**	**SL**	**SLT**	**RL**	**RLT**
ACC	76.34%	76.92%	63.52%	66.59%	70.51%	71.51%	70.70%	71.06%
ERR	71.93%	69.65%	57.24%	47.12%	62.26%	59.44%	62.04%	59.45%
RAW	23.31%	25.32%	32.31%	41.73%	30.30%	32.73%	30.62%	32.63%
LS	5,628	8,156	1,193	1,246	2,514	2,620	2,401	2,463
!L	-	-	23.89%	8.02%	16.06%	10.46%	16.61%	12.23%
!HL	-	-	33.33%	97.98%	48.19%	73.99%	46.59%	62.96%

Table 6.1: Handwriting Recognition Performance

	CLT to RLT	**CL to RL**	**CLT to ALT**	**CLT to SLT**
RAW Match Rate	↑ 7.48%	↑ 7.42%	↑ 17.58%	↑ 7.42%
Error Rate	↓ 10.78%	↓ 10.88%	↓ 24.53%	↓ 10.21%

Table 6.2: Comparison between Handwriting Recognition Experiments

LEXICON ANALYSIS METRIC	**VALUE**
Accuracy of Reduction (α)	0.33
Degree of Reduction (ρ)	0.83
Reduction Efficacy (η)	0.06
Lexicon Density (ϱ')	1.07 → 0.87
Lexicon Density (ϱ”)	0.50 → 0.78

Table 6.3: Lexicon Reduction Performance between the Complete Lexicon (CL) and the Reduced Lexicon (RL)

ENVIRONMENT ITEM	**VALUE**
Training Deck PCR Size	750
Testing Deck PCR Size	62
Training Deck Lexicon Size	5,628
Testing Deck Lexicon Size	2,528
Training + Testing Deck Lexicon Size	8,156
Training Deck Words for Modeling	42,226
Testing Deck Words to Recognize	3,089
Modeled Categories / RSVD Dimensions	24
Category Selection Threshold	0.55
Maximum Categories per Form	5
Average Categories per form	1.40
Max Phrases Per Category	50
Apple OS X Memory Usage	520 MB
Apple OS X G4 1GHZ Train Time	15-20 mins/exp
Apple OS X G4 1GHZ Test Time	3 hrs/exp

Table 6.4: Handwriting Recognition System Environment

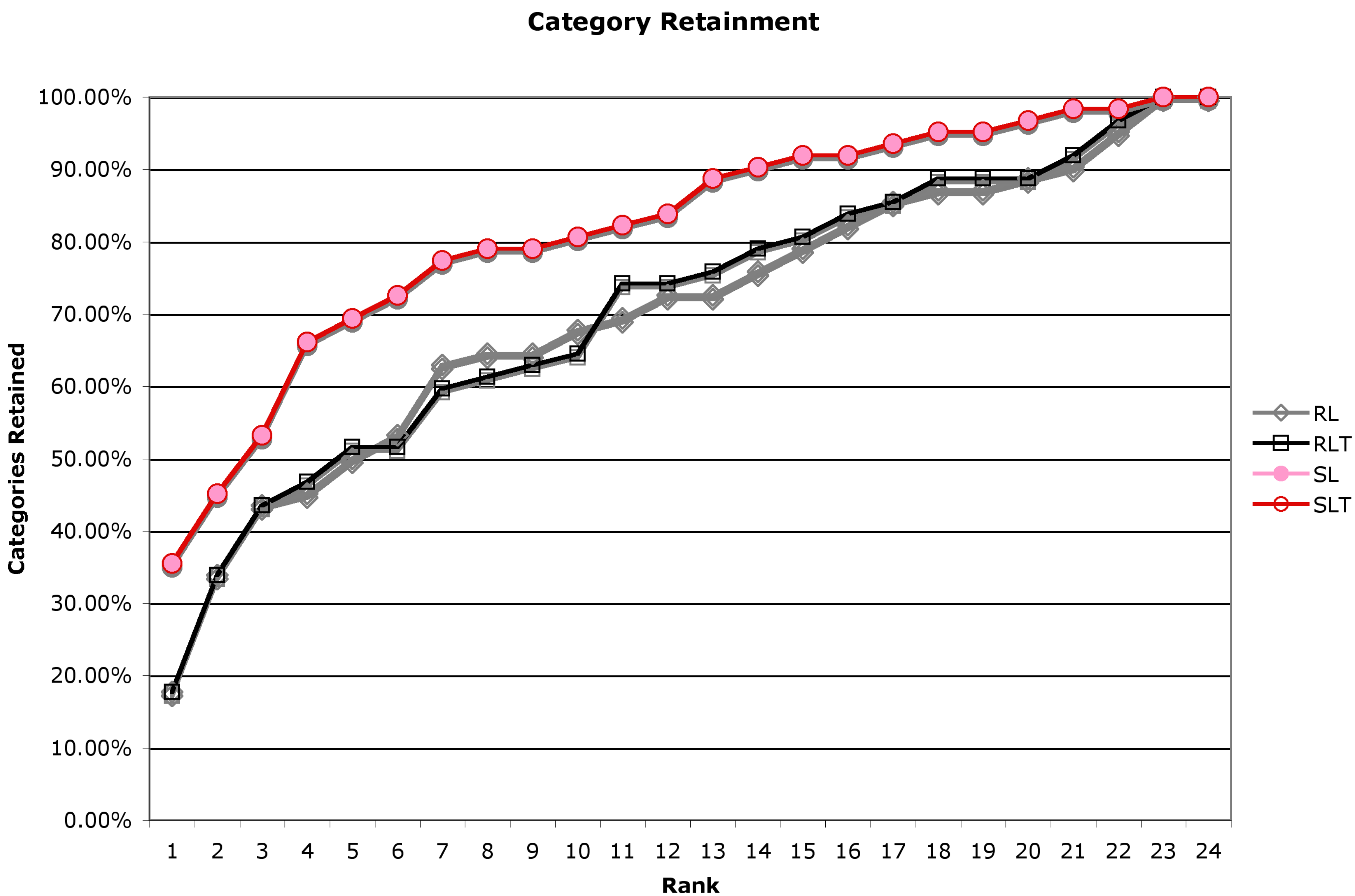


Figure 6.1: Category Retainment by Rank

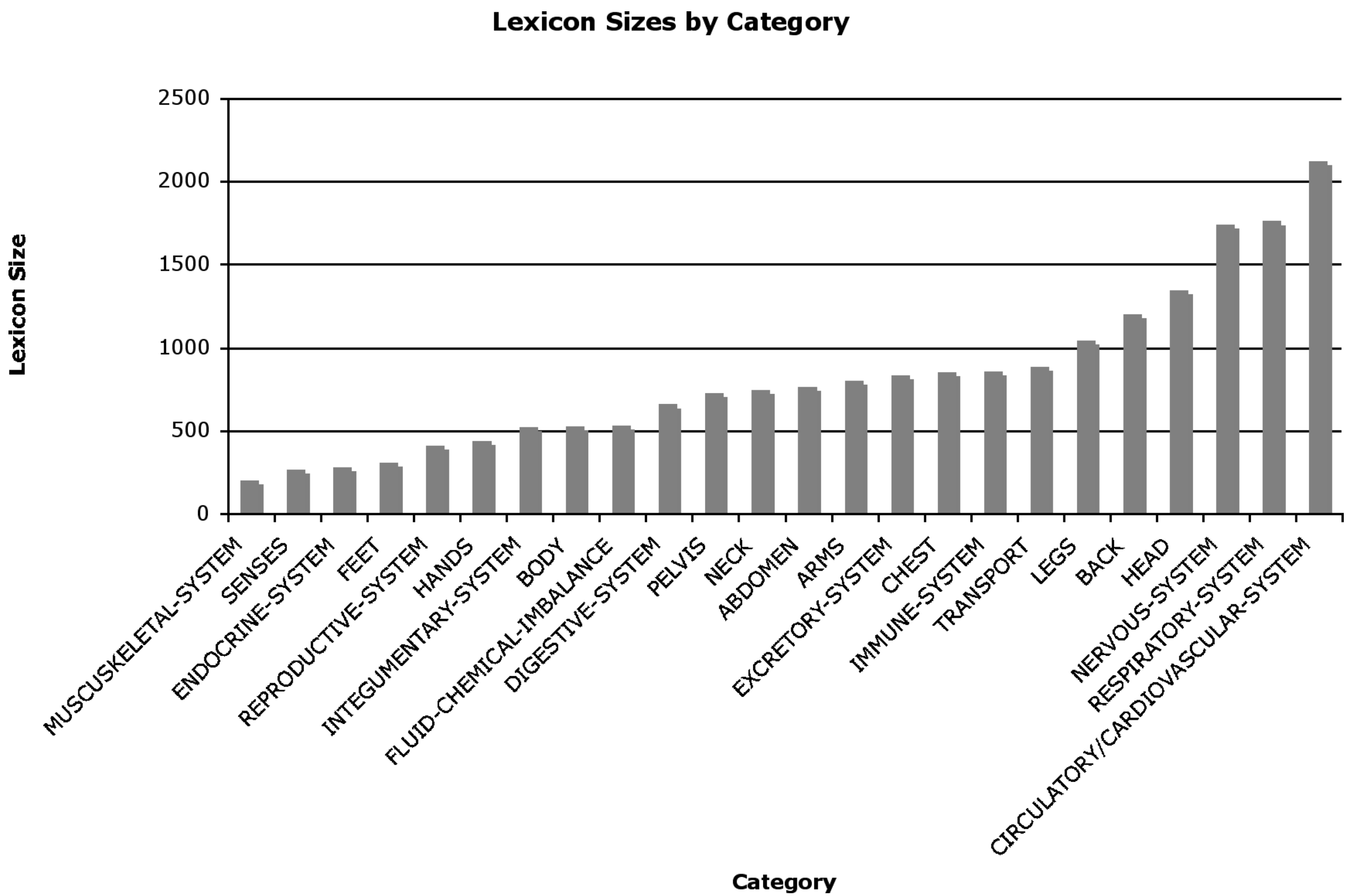


Figure 6.2: Sorted Lexicon Sizes by Category

6.1 Performance Measures

ACC (accept recognition rate): number of words the word recognizer accepts above an empirically decided threshold.

ERR (error recognition rate): number of words incorrectly recognized among the accepted words.

RAW (raw recognition rate): top choice word recognition rate without use of thresholds.

LS (lexicon size): the lexicon size for the experiment after any reductions.

!L (truther word not present in the lexicon): percentage of words (for a specific experiment) not in the lexicon as a result of incorrectly chosen categories or due to the absence of that word in the training deck.

!HL (human being could not completely decipher word): percentage of the !L set in which even human beings could not reliably decipher all or some of the characters in the word (given the context).

6.2 Experiments

CL (complete training lexicon): The union of all words in the training set.

CLT (complete training lexicon + test deck lexicon): The union of all the words in the training and test sets.

AL (assumed training lexicon): This is a reduced lexicon constructed from the training deck where the categories are determined by an Oracle.

ALT (assumed training lexicon + test deck lexicon): Same as AL except that all words from the test set are also inserted into the training deck category lexicon. This gives the upper bound for the effectiveness of the reduced lexicon strategy.

RL (reduced lexicon): The reduced lexicon from the training deck, which is the union of words from the top ranked categories returned by the word recognizer. This is a practical measure of the current performance of the system.

RLT (reduced lexicon + test deck lexicon): Same as RL except that all words from the test set are inserted into the training deck category lexicon. This shows the effectiveness of word recognition under the assumption that the category lexicons are complete.

SL (synthetic term generation): This is the reduced lexicon in which the categories are determined by a synthetic generation of the truth word. This is the theoretical upper bound of RL in which the handwriting recognition is a 100% accept rate with a 0% error rate.

SLT (synthetic term generation + test deck lexicon): Same as SL except that all words from the test set are inserted into the training deck category lexicon. This is the theoretical upper bound of RLT.

6.3 Discussion

In all experiments it is assumed that the word segmentation and extraction has been performed by a person. Also, forms in which 50% of the content is indecipherable by a human being are omitted. This occurs 15% of the time.

In reference to Table 6.2 which is computed from the most relevant changes in Table 6.1 : The theoretical RLT (i.e. comparing RLT to CLT) improves the RAW match rate by 7.48% and drops the error rate 10.78% with a *degree of reduction* ρ = 61.59%. The practical RL (i.e. comparing RL to CL) improves the RAW match rate by 7.42% and drops the error rate by 10.88%, with a *degree of reduction* ρ = 51.30%. The RLT and RL numbers are close due to the difference in the initial lexicon sizes: CLT/RLT starts with 6,561 words (i.e. training deck and testing deck lexicons) whereas the CL/RL starts with 5,029 words (i.e. training deck lexicon only). The RLT lexicon is more complete, but the lexicon is larger. The RL lexicon is less complete, but the lexicon is smaller. Thus, RLT gives the advantage that the recognizer has a greater chance of the word being a possible selection and RL gives the advantage of the lexicon being smaller. The ALT shows the theoretical upper bound for the paradigm: (i) the categories are correctly determined 100%, and (ii) the lexicon is complete. The ALT (i.e. comparing ALT to CLT) improves the RAW match rate by 17.58% and drops the error rate 24.53% with a *degree of reduction* ρ = 83.01%. The synthetic experiments (SL and SLT) also do not offer much improvement which shows perfect character extraction does not guarantee recognition improvement. This is due to two reasons: (i) a form is a representation of many characters and so some incorrectly recognized characters are tolerated, and (ii) the remaining words on the form to be recognized are difficult to determine even when the lexicon is constructed with only words of known uni/bi-gram terms. Table 6.3 provides insight into the effectiveness of the lexicon reduction from the complete lexicon (CL) to the reduced lexicon (RL) experiments. The *lexicon density distance metric* ϱ' shows less confusion among lexicon words considering all the characters are equally important. This implies that the reduced lexicon will be less confusing to the recognizer. The *n-gram lexicon distance metric* shows an increase in the quantity of words with common NSI encodings. This implies the recognizer has a greater chance of selecting a word using the confidently selected characters.

6.4 Summary

Both theoretical and practical recognition experiments have been shown both before and after the lexicon reduction algorithm was applied. The algorithm has shown approximately 8% improvement to raw recognition rate, a reduction in error rate by about 11% and a lexicon reduction of over 50% in practical experiments. In the theoretical situation that the categories are always correctly determined, a recognition improvement of about 18%, a reduction in error rate by about 25% and a lexicon reduction of about 83% are shown.

Chapter 7

Medical Form Search Engine

7.1 Search Method

In this section, various search engine approaches are compared. The inputs to the search engine are a set of PCR medical forms and a query. The output are those forms which match the input query.

All known available search engines are based on the assumption that the text is already in a digital text format. The technologies have focused on parsing and organizing the content in a variety of formats (e.g. PDF, PS, HTML, XML, and other proprietary document formats). There is no widely used search engine technology which can directly search and analyze the content of digital handwritten documents. This query ability is necessary for the Health Surveillance (see Appendix) application to access medical forms presented with a specific type of medically related condition.

In order to have a query deck of sufficient size, we use the leave-1-out strategy which is explained as follows. Suppose a total of 10 PCRs are available. Take the first PCR as the test deck and the remaining 9 PCRs as the training deck, and perform the recognition and tagging on that single PCR. Next, repeat the process, except that now the test deck consists only of the 2nd PCR while the training consists of the first PCR and the remaining 8 PCRs. The recognition and tagging on the 2nd PCR is now performed. This exhaustive processing of recognition and tagging repeats 10 times, thereby providing a training deck and an unbiased test deck of the same size. Applying this process to 800 PCR forms, the notion is the same, except the split is Leave-100; i.e. 8 experiments are performed using groups of 100. Finally, a set of 1,250 phrases, constructed from adjacent non-stopwords, are extracted from an isolated deck of 200 PCR forms (i.e. these 200 forms are not a subset of the 800 deck) such that each phrase is found in at least one form in the 800 deck.

A query is performed by scanning the forms in the 800 test deck for recognized words that match an input query phrase. Two query experiments are performed and displayed in Figure 7.1: CL and RL. In the CL (complete lexicon) experiment, the raw LDWR recognized words computed from the full lexicon are compared against the query. In the RL (reduced lexicon) experiment, the raw LDWR words computed from the reduced lexicon are compared against the query. A set of ranking rules are applied, relevance determined, and the recall-precision table generated (see Table 6.1 and Figure 7.1). A relevant PCR is a document in which a human truther classifies at least one occurrence of each word from the input phrase.

Ranking rules given an input phrase of exactly two words:

- Both words must match the recognized words or that PCR is not returned.
- A double precision rank is computed by summing the values in these two steps:
  - Summing the frequencies of the occurring phrase words in the document.
  - Summing the distance between all recognized word occurrences in the document using Equation 7.1.1. Let $d(a_i, b_j)$ be a function which computes the distance between the input phrase of two words, a_i and b_j such that *i* and *j* respectively represent the word position in the document.

$$d(a_i, b_j) = \frac{1}{|a_i - b_j|} \tag{7.1.1}$$

Unlike typical text retrieval systems, the words on a PCR may be incorrectly recognized by the handwritten recognition engine. In addition, general search engines need to be concerned about external influences such as spamming, which is not a concern in this application. Therefore, a more trivial ranking measure such as of nearness/proximity in Equation 7.1.1 is sufficient.

7.2 Results

The comparison of the complete and reduced lexicon queries can be found in Figure 7.1. The plot illustrates only those queries which returned at least one record. This is because the precision value ($\frac{RelevantRetrieved}{Retrieved}$) is undefined when no documents are returned [54]. Queries in the CL series returned 0 forms 73% of the time, and returned only 1 record on average. The RL returned 0 forms 23% of the time and returned 7.5 documents on average. Thus, with RL, about 3 times more queries had at least one response.

NOTE: The two curves for CL and RL in Figure 7.1 are not directly comparable as the CL curve reflects data from 50% fewer queries corresponding to cases when no forms were returned.

7.3 Discussion

One question that arises is the validity of the search engine approach. An alternative search engine approach involving the expansion of the query terms into their respective ESI combinations can be applied directly to the initial LDWR character recognition results. This would effectively bypass the more elaborate search engine except that this alternative approach significantly under-performs. While results are returned 99.8% of the time, with

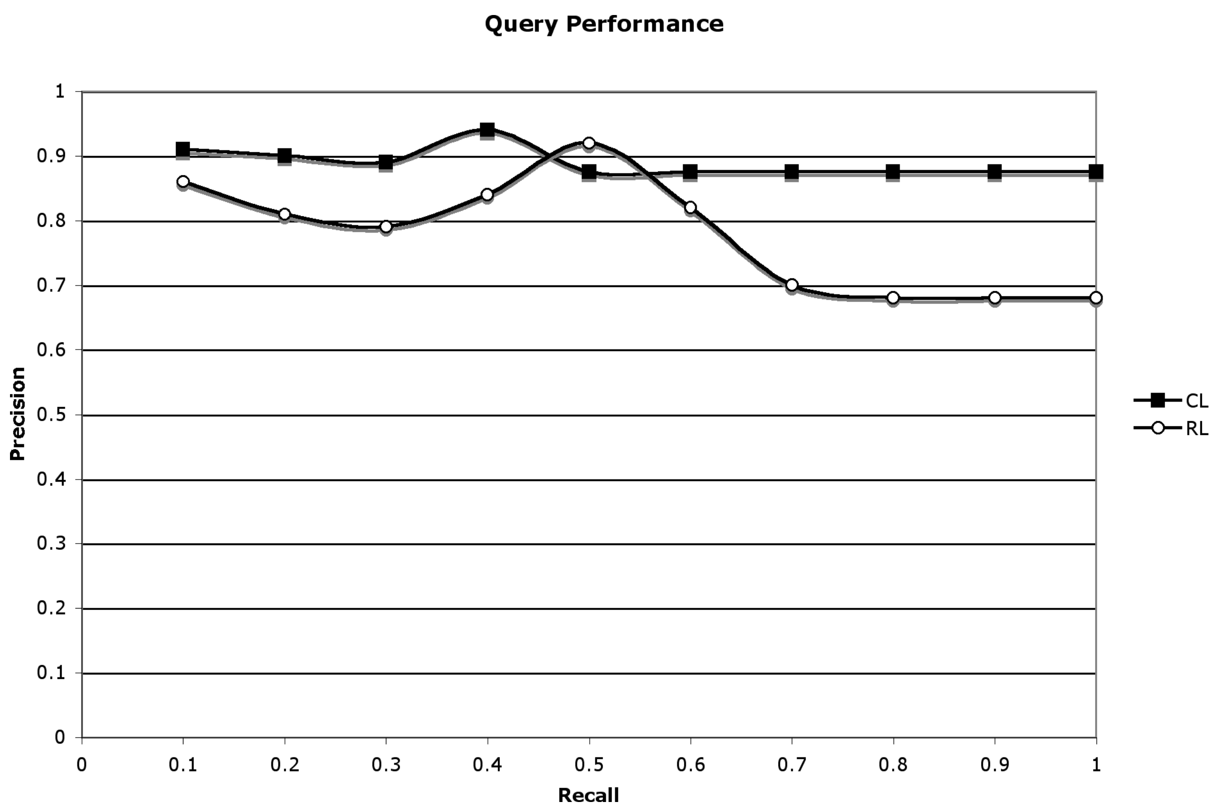


Figure 7.1: Recall/Precision Chart for Medical Form Search Engine

125 records returned on average, the precision of the results is very low. As intuitively expected, the uni/bi-grams match more terms due to the loss of word information. The recall/precision chart in Figure 7.2 illustrates a drop in retrieval effectiveness. This demonstrates the dependence of the searches to operate at the word level, rather then the character level. The lexicon reduction strategy which improves the handwriting recognition performance also improves the search effectiveness as expected.

For example, consider input query phrase *CHEST PAIN*:
CHEST is decomposed into: CH, CE, CS, CT, HE, HS, HT, ES, ET, C, H, E, S, and T.
PAIN is decomposed into: PA, PI, PN, AI, AN, IN, P, A, I, and N.

In addition, the spatial information is known since the input query is provided by a user.

The ESI encodings for *CHEST* is decomposed into: 0C0H3, 0C1E2, 0C2S1, 0C3T0, 1HE2, 1H1S1, 1H2T0, 2E0S1, 2E1T0, 0C4, 1H3, 2E2, 3S1, and 4T0.

The ESI encodings for *PAIN* is decomposed into: 0P0A2, 0P1I1, 0P2N0, 1A0I1, 1A1N0, 2I0N0, 0P3, 1A2, 2I1, and 3N0.

Finally, all possible ESI sequences are generated: 0C0H3$0P0A2, 0C0H3$0P1I1, 0C0H3$0P2N0, 0C0H3$1A0I1, etc...

If any of these ESI sequences match any of the character spatial encodings from the LDWR recognition, then that form is returned. Relevancy is determined if the input query words *CHEST* and *PAIN* are actually found on that form according to the truth.

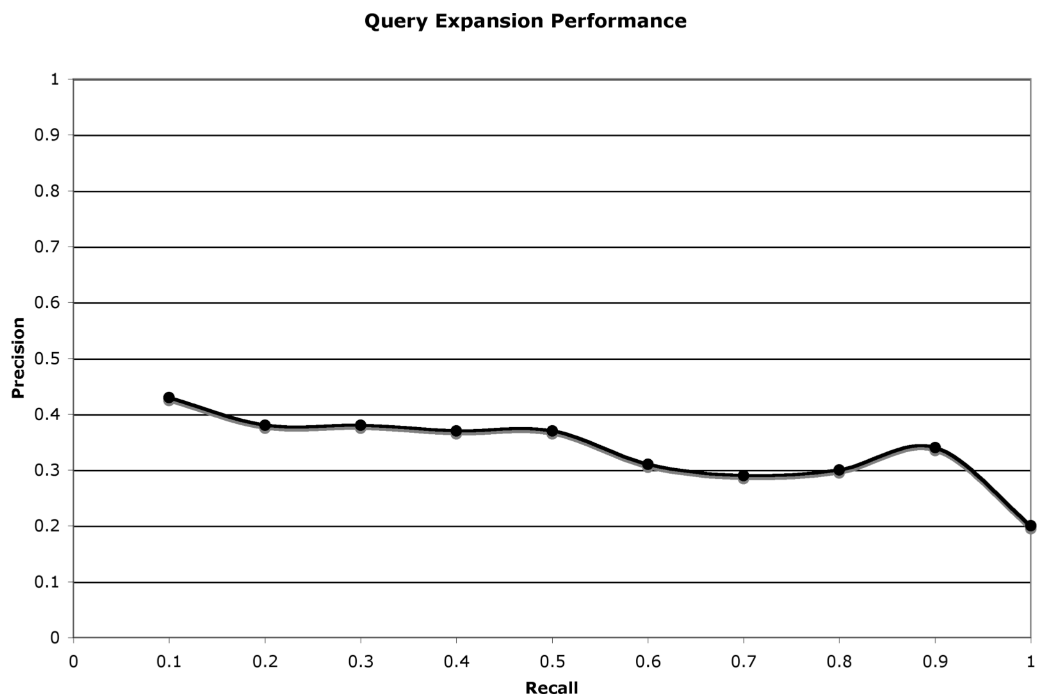


Figure 7.2: Recall/Precision Chart using Query Expansion

7.4 Summary

This section presents the effectiveness of a medical form search engine before and after the reduced lexicon. Two methods have been presented: one in which the recognized words are matched and the second in which only the highest confident characters are matched. The character matching scheme has very low performance because characters, with no mapping, have no meaning. The word level matching after the reduction provided an increase of successful queries by 50% with more possibilities returned to the user than had the reduction not been applied.

Chapter 8

Applications

Several applications are described in this section which can benefit from the real-time storage of PCR or other medical form data. The first example described is the procedure for a complete *health surveillance system* which takes a PCR form at the hospital, recognizes the content, stores and indexes the information in a centralized and secure data repository. Once stored, other applications such as *syndromic surveillance software* can be used to extrapolite trends and red flag regional medical concerns such as a pandemic. To improve the communications of such a health surveillance system, it is necessary to comply with existing standards involved in the exchange of medical information. This research is shown to be compliant with the Center for Disease Control and Health Level 7 information exchange protocols. Since the automated recognition of medical forms is not perfect, the integration of the technology with keying sites, to improve the efficiency of human data entry of such information, is also proposed. The data can also be used to evaluate the quality assurance of healthcare personnel by government authorities. Finally, the application of this research towards a prescription verification system at pharmacies is also described.

8.1 Health Surveillance

...We must prepare to minimize the damage and recover from any future terrorist attacks that may occur despite our best efforts at prevention. Past experience has shown that preparedness efforts are key to providing an effective response to major terrorist incidents and natural disasters. Therefore, we need a comprehensive national system to bring together and command all necessary response assets quickly and effectively... [131].

-United States Office of Homeland Security, 2002

The definition of a public health surveillance system is "the ongoing systematic collection, analysis, and interpretation of outcome-specific data for use in the planning, implementation and evaluation of public health practice" [28]. The most important theme in this definition is the reference to the collection of data; without it, there is nothing to be analyzed. As implied from the above quotation, access to medical data is one possible asset in the construction of a national emergency system.

It is highly probable that no completely autonomous system can collect data with a 0% error rate. To date, there is no automated and centralized system in the United States for retrieving significant medical information from paper forms. The amount of information to manually digitize by human beings is simply too great. However, with the introduction of this medical form-driven handwriting technology, it is now possible to add a new piece to the emergency system.

Figure 8.1 shows the flow of data extraction and dissemination for both state government (e.g. New York State) and federal systems (e.g. Center for Disease Control (CDC) [23]). The development of this system requires data in a standardized manner. The CDC constructed the NEDSS format to manage diverse data in a standard organized format [23]. Once the data are available in a secure and centralized repository, it will allow for continuous access to medical form data for the purposes of epidemiological analysis, outbreak

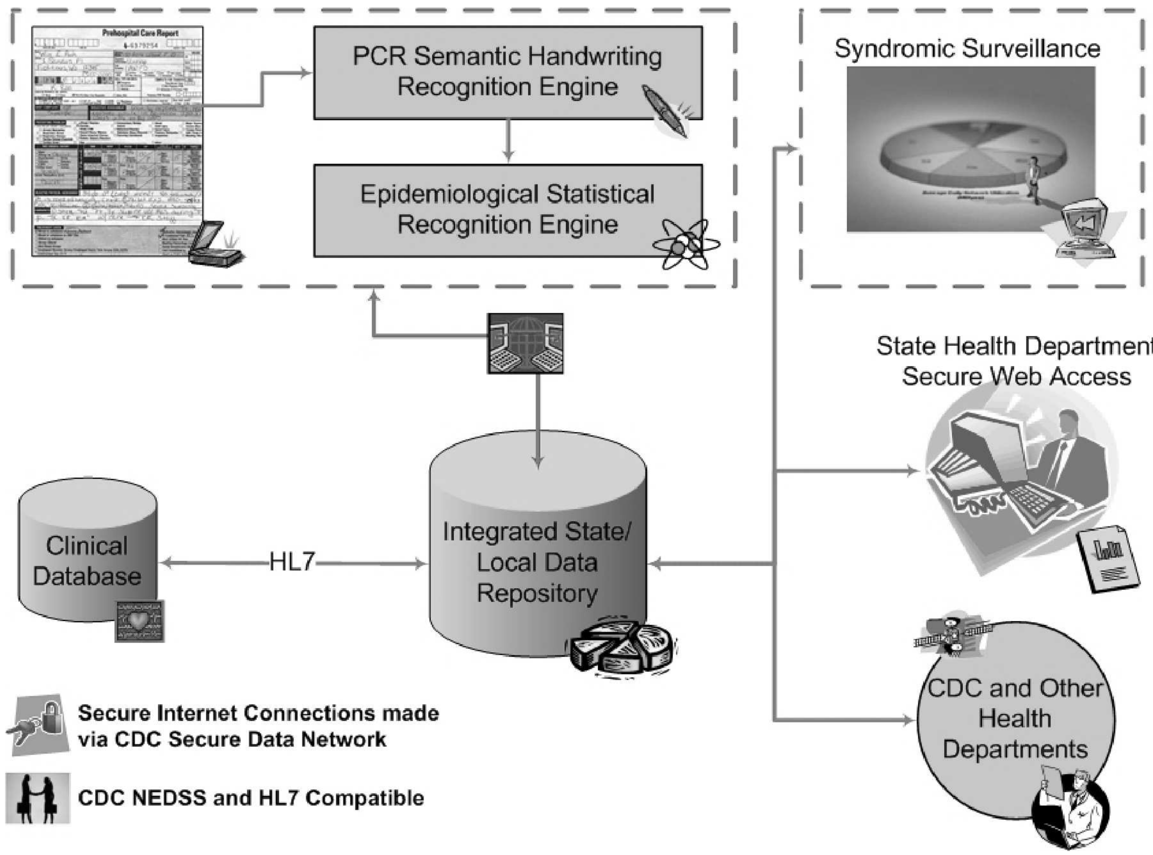


Figure 8.1: Possible Health Surveillance Protocol

detection, counter-bio-terrorist information technology, health-care quality assurance, and further enhancement to the medical student's studies and awareness.

The health surveillance application described in Figure 8.1 allows the creation of a data processing facility where high speed scanners can be used to scan stacks of PCR forms, feed the forms into the image processing and recognition algorithms, and produce ASCII text output as well as a PCR category, all with related confidence levels. These forms can then be sorted within each category on content severity. Once sorted, the forms can be provided to the various branches of the Department of Health responsible for health-care provider quality assurance. This is currently a non-automated process involving inefficient human labor. In addition, the forms appear to be entering the system at a faster rate than they can be analyzed, creating a continuous backlog of such information. Several problems are currently created due to this information access slow-down:

- A health care professional could continue several years without receiving comments on the quality of care. This means that all patients managed by that health-care professional might have been affected by an incorrect or non-optimal medical decision that could have been addressed sooner.

- Any outbreak (e.g. whooping cough, SARS, ebola, bird flu), may go undetected until a substantial number of people are affected. The proposed Health Surveillance System aims to address these needs.

- The ability to improve and standardize flaws in rescue tactics and techniques may take years to discover.

Further details on infectious disease informatics, outbreak detection, health surveillance and biological terrorism can be found in [25] [28] [35].

8.2 Center for Disease Control (CDC) Integration

The Public Health Information Network (PHIN) and the National Electronic Disease Surveillance System (NEDSS), standardized by the CDC, were created for such tasks as establishing medical trends [23]. However, there are vast amounts of data that are not being collected due to resource limitations. Their systems can be improved with the introduction of medical form recognition and information retrieval systems. The CDC also has a Secure Data Network Standards and Procedures protocol for allowing the secure communications of such information [23].

The Logical Data Model Data Dictionary (NLDM) [23] protocol modules of interest are:
• Observation class: allows subjective, objective and assessment-based input.
• Observation_Interp: allows a “very rough interpretation” of an observation.

Using the data provided by the algorithms in this research, the NLDM Observation classes can be populated so that the following CDC PHIN systems can use them: Early Event Detection (EED), Outbreak Management (OM), and Countermeasure and Response Administration (CRA) [23].

8.3 Health Level 7 (HL7) Integration

Health Level 7 (HL7) [55] is an OSI layer data exchange protocol designed as a standard for message communication between medical applications. While there is a grammar and vocabulary to this system, there is also the ability to have a custom message, called a Z-Segment. The Z-Segment proposed should be in XML format, and would contain the following information:

$\langle PCR - FORM \rangle$
$\langle IMAGE - ID \rangle$ [pcr-image-id] $\langle / IMAGE - ID \rangle$
$\langle CHECKSUM \rangle$ [checksum] $\langle / CHECKSUM \rangle$
$\langle ACCESS - PERMISSIONS \rangle$ [permissions] $\langle / ACCESS - PERMISSIONS \rangle$

$\langle CRYPTOGRAPHIC - HEADER \rangle$ [crypto-header] $\langle / CRYPTOGRAPHIC - HEADER \rangle$
$\langle CRYPTOGRAPHIC - EXPIRATION \rangle$ [crypto-expiration] $\langle / CRYPTOGRAPHIC - EXPIRATION \rangle$
$\langle CREATION - DATE \rangle$ [creation-date] $\langle / CREATION - DATE \rangle$
$\langle EXPIRATION - DATE \rangle$ [expiration-date] $\langle / EXPIRATION - DATE \rangle$
$\langle AUTOMATED - COLLECTION - SITE - ID \rangle$ [site-id] $\langle / AUTOMATED - COLLECTION - SITE - ID \rangle$
$\langle LOCALE - INFORMATION \rangle$ [locale-information] $\langle / LOCALE - INFORMATION \rangle$
$\langle RECOGNITION - CONFIDENCE \rangle$ [pcr-recognition-confidence] $\langle / RECOGNITION - CONFIDENCE \rangle$
$\langle HUMAN - VERIFIED \rangle$ [human-verified] $\langle / HUMAN - VERIFIED \rangle$
$\langle HUMAN - NOTES \rangle$ [human-notes] $\langle / HUMAN - NOTES \rangle$
$\langle HUMAN - LANGUAGE \rangle$ [human-language] $\langle / HUMAN - LANGUAGE \rangle$
$\langle PCR - TAGGED - CATEGORIES \rangle$
$\langle CATEGORY_1 \rangle$ [category-name] $\langle / CATEGORY_1 \rangle$
$\langle CATEGORY_2 \rangle$ [category-name] $\langle / CATEGORY_2 \rangle$
$\vdots$
$\langle CATEGORY_n \rangle$ [category-name] $\langle / CATEGORY_n \rangle$
$\langle / PCR - TAGGED - CATEGORIES \rangle$
$\langle RECOGNITION - DATA \rangle$
$\langle WORD_1 \rangle$
$\langle TEXT \rangle$ [text] $\langle TEXT \rangle$
$\langle RECOGNITION - SCORE \rangle$ [recognition-score] $\langle RECOGNITION - SCORE \rangle$
$\langle COORDINATES \rangle$ [coordinates] $\langle COORDINATES \rangle$
$\langle / WORD_1 \rangle$
$\langle WORD_2 \rangle$
$\langle TEXT \rangle$ [text] $\langle TEXT \rangle$
$\langle RECOGNITION - SCORE \rangle$ [recognition-score] $\langle RECOGNITION - SCORE \rangle$
$\langle COORDINATES \rangle$ [coordinates] $\langle COORDINATES \rangle$
$\langle / WORD_2 \rangle$
$\vdots$
$\langle WORD_n \rangle$
$\langle TEXT \rangle$ [text] $\langle TEXT \rangle$
$\langle RECOGNITION - SCORE \rangle$ [recognition-score] $\langle RECOGNITION - SCORE \rangle$
$\langle COORDINATES \rangle$ [coordinates] $\langle COORDINATES \rangle$
$\langle / WORD_n \rangle$
$\langle / RECOGNITION - DATA \rangle$
$\langle / PCR - FORM \rangle$

All segments must be protected by the maximum government approved security and cryptographic standard when transmitted over a network. Furthermore, the XML message should be encrypted with a private key-based algorithm. Hospitals would need to register to have access to a specific private key. This restricts hospital access to those designed by the appropriate Department of Health agencies.

8.4 Keying Sites

Currently, the only way to enter medical information is to have keyers enter this information at data entry warehouses. When great amounts of medical information are to be collected, it is inefficient to rely on human data entry. An improvement to this is to have an automated medical handwriting recognition system propose its interpretation of a form to a human keyer. The human keyer can visually confirm and, thereby, change only the information which is inaccurate. The improvement in data entry time can be estimated using the RAW match handwritten recognition rate.

8.5 Quality Assurance Improvement

When a trained professional assists a patient, how is the performance of that health-care worker evaluated? A backlog of medical forms currently exists due to the enormous daily influx of PCR forms. Therefore, a health-care professional who aids a patient may not be evaluated for years. Perhaps they are not even in the same jurisdiction or in the same career path. The spot checking of a small portion of medical forms in a short period of time is more effective than analyzing all forms with a long lag. It, therefore, follows that the forms to retrieve are those which most likely involve the more complicated and, hence, more error-prone rescue scenarios (e.g. respiratory arrest). The brute force approach is less effective.

8.6 Prescription Medications Protection System

In spite of the advancements of computer technology, paper is still used to file prescriptions. In some instances, misinterpretations of medical prescriptions have caused unnecessary suffering [101].

Common causes of medication errors as described by [101]:

- Look-alike containers
- Poor handwriting and look-alike drug names
- Oral orders misheard or understood
- Improper patient identification (in hospital and pharmacy)
- Improper drug storage
- Taking another's medication

While not all of these can be addressed by an automated recognition system, the *poor handwriting* can be checked. Since the automated system needs context, a prescription from the doctor should include a short description or reason for the medication in addition to the medication and dosage which is prescribed. The introduction of this system would only be used as a supplement to that of the pharmacist. Considering that such errors result in 1/12 of hospital admissions and 1/8 of emergency room visits [101] [3], according to the American Society of Health System-Pharmacists (ASHP) [3], a supplemental automated system is justified. Although it is important to note that the introduction of another technology also runs the risk of creating additional errors [98].

While no automated system can solve this problem completely, an assistive system can reduce suffering in a small number of subjects. It is proposed that a standard form be constructed to capture the basic conceptual information. Extra checking can be performed on those medications which might be hazardous to certain patients.

Chapter 9

Software System

A new enterprise software environment, built by the author from scratch using approximately 50,000 lines of well structured object oriented (OOP) code, facilitated the following processes: (i) truthing: the process of data entry of all form data (see Figure 9.2), (ii) reviewing: the process of correcting and viewing truthed data (see Figure 9.3), (iii) integration with any SQL compliant relational database management system (RDBMS) to organize and manage sets of images to be truthed and reviewed (see Figure 9.1), (iv) a scientific visualization tool capable of observing the application of all text extraction and recognition algorithms presented in this paper (see Figure 9.4), (v) a multi-threaded, multi-tier, cross-platform environment using transport layer security (TLS) over an object serialized TCP/IP layer to facilitate secure real-time truthing, reviewing and evaluation, and (vi) a series of batch operations involving image processing, handwriting recognition and form retrieval. In addition to the elaborate GUI, the system can perform combinatorial algorithm sequences on training and test decks and, finally, output extensive reports. The software is the result of strong engineering effort and successful completion of rigorous unit testing. The architecture consists of several diverse and scalable programming languages and database engines including $Java^{TM}$ [60] and $MySQL^{®}$ [90].

There are four interfaces to the software. Figure 9.1 shows the interface for TCP/IP network communications and between multiple clients to a designated server. There is also a frame in which the creation and management of sets is performed. Figure 9.2 illustrates the interface used by people to identify the locations and interpretation of words on PCR images. Figure 9.3 shows an interface used for the review and verification of words by an additional arbitrator. Figure 9.4 is an interface which allows the real-time analysis of pre-processing, binarization, post-processing, and recognition algorithms used during research.

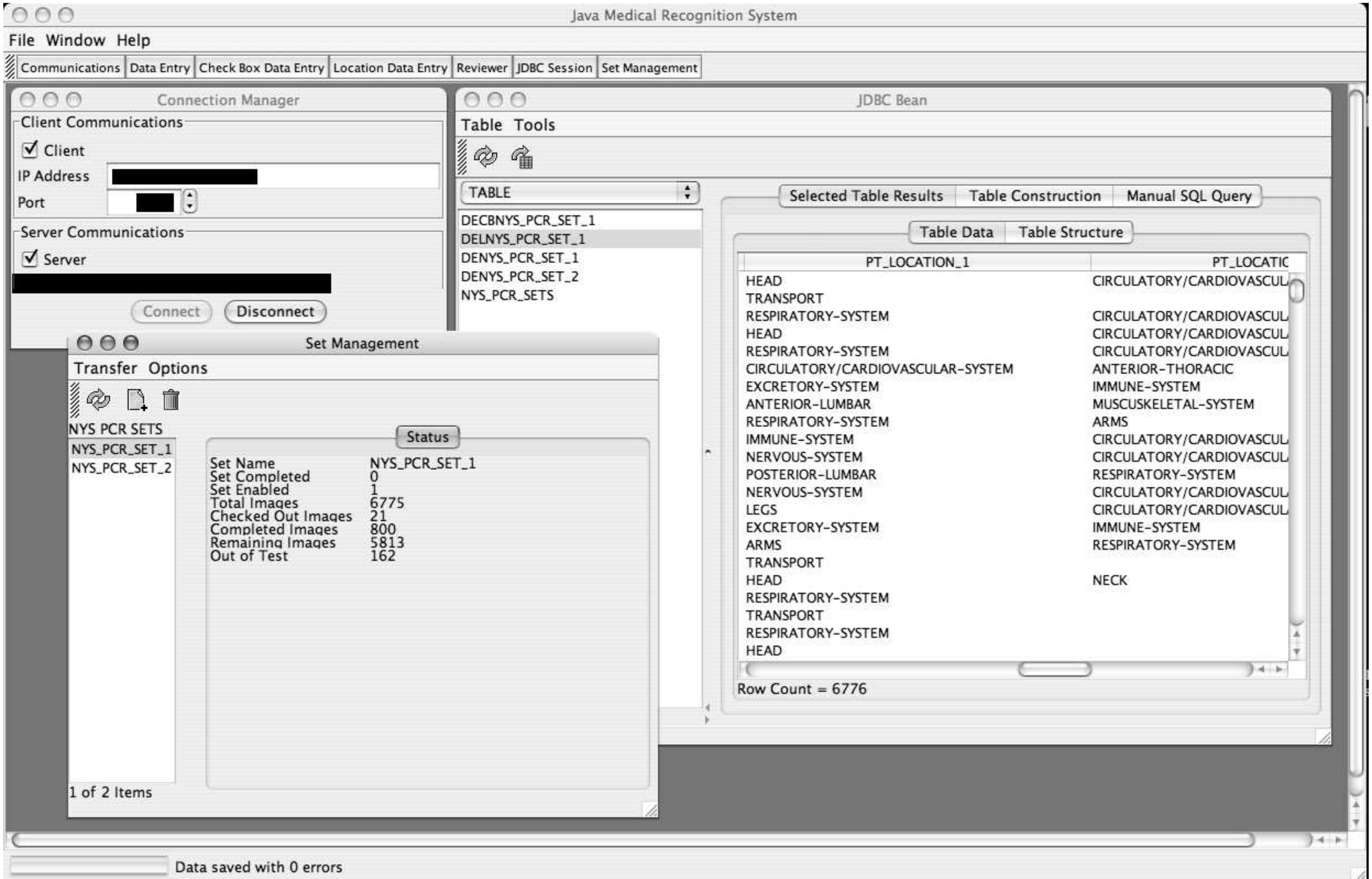


Figure 9.1: Network and Database Communications

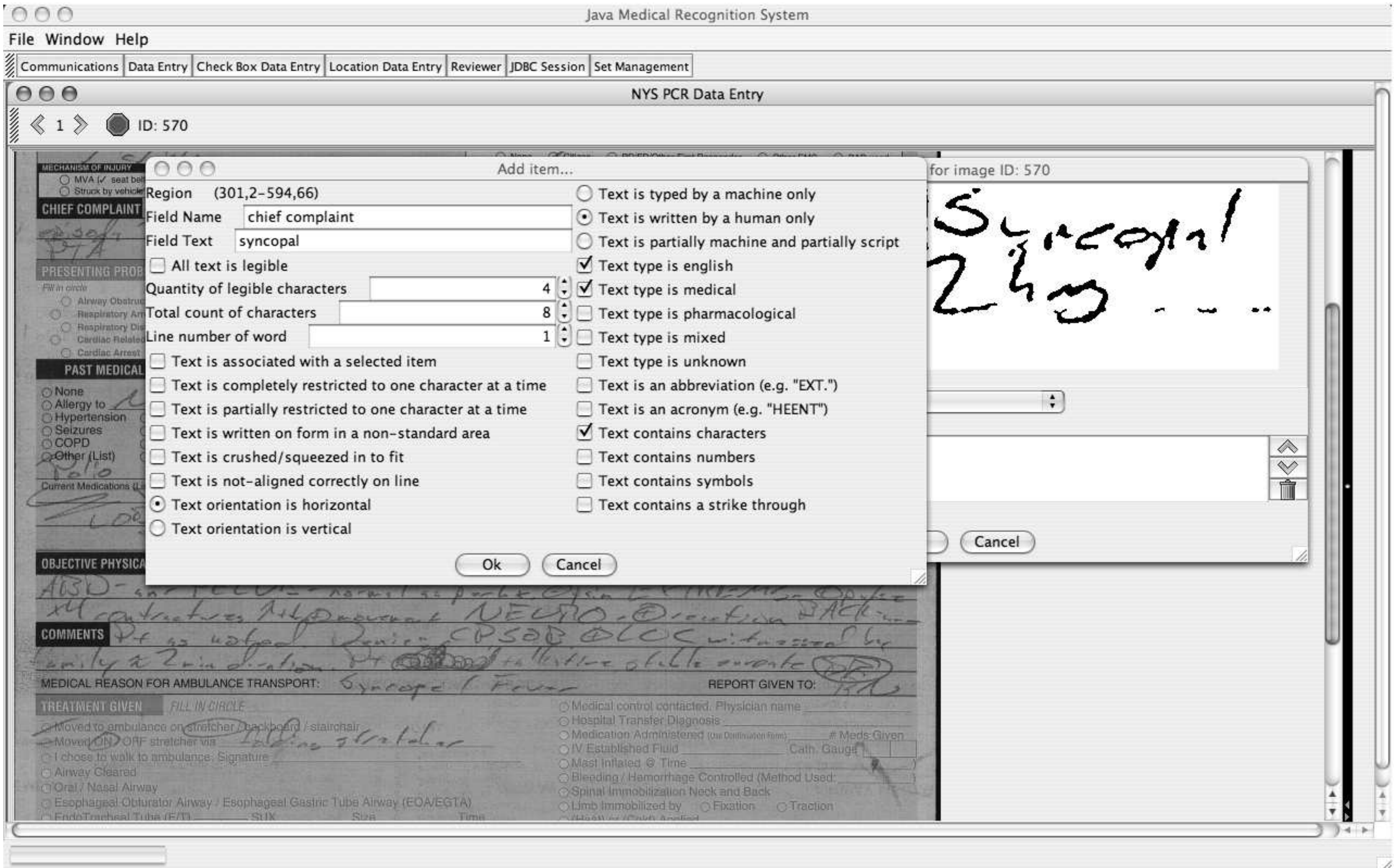


Figure 9.2: Truthing Interface

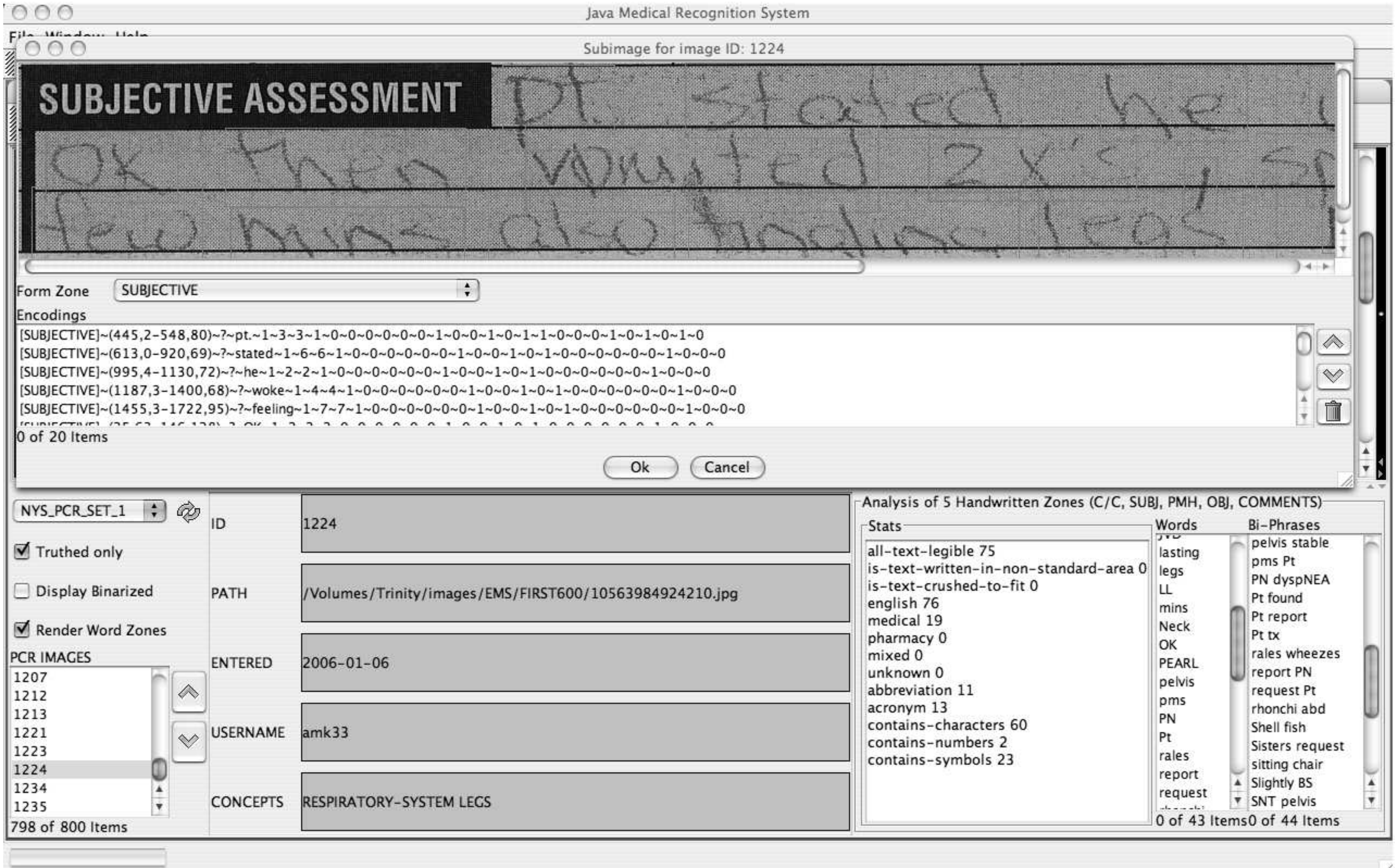


Figure 9.3: Reviewing Interface

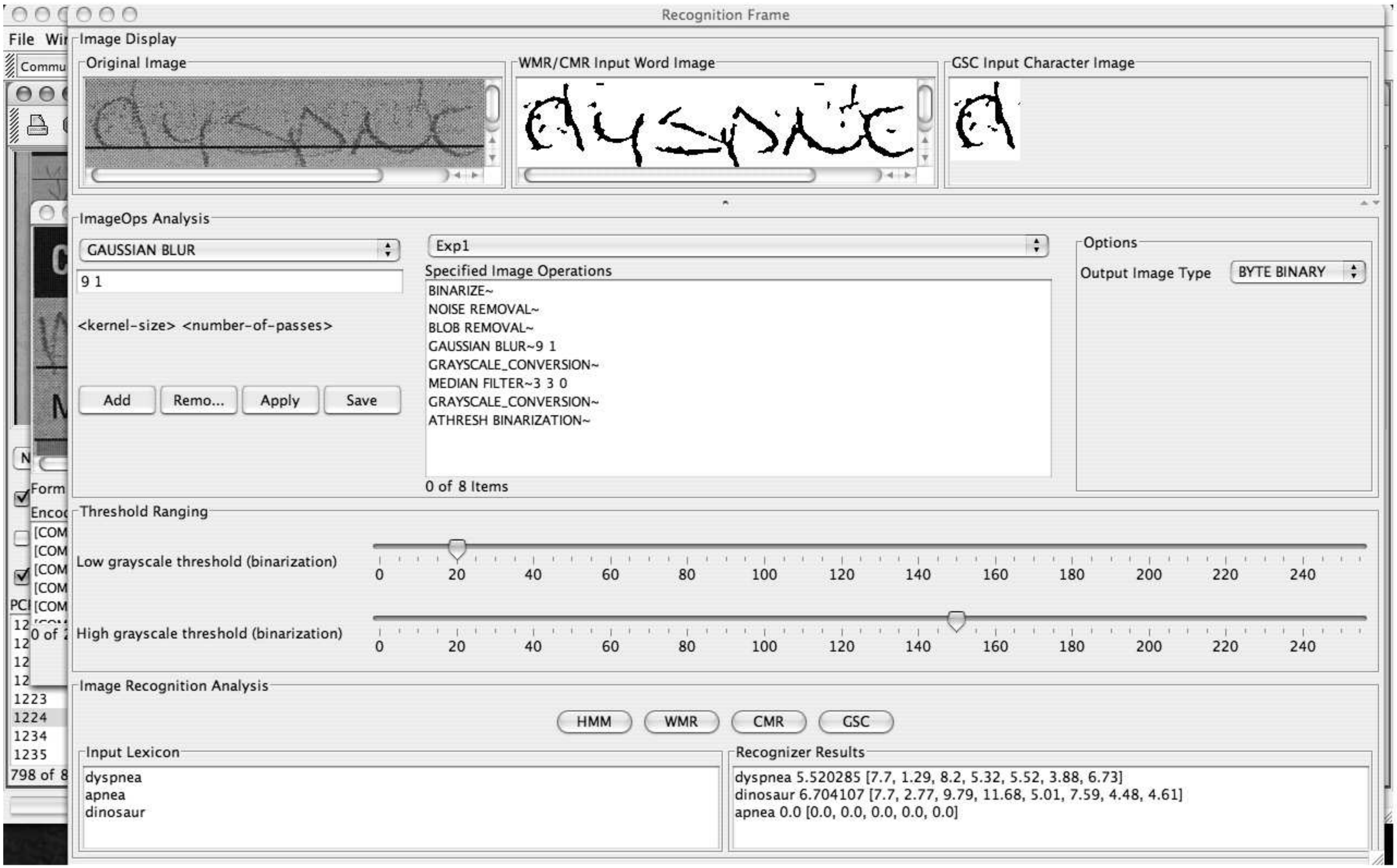


Figure 9.4: Recognition Interface

Chapter 10

Conclusions

10.1 Summary

This research describes the development of new algorithms by hybridizing handwriting recognition, information retrieval, image processing, natural language processing and latent semantic analysis.

Major contributions of this research are:

• The first search engine which operates on handwritten medical forms (see Chapter 7).

• A new lexicon reduction paradigm which can be applied to lexicon driven handwriting recognition algorithms (see Chapter 5).

• New metrics for evaluating the performance of lexicon reduction algorithms (see Chapter 4).

• A new binarization algorithm with comparative results against other such algorithms and post-processing techniques. A new strategy using sinusoidal waves has been introduced. This is the first algorithm that operates on carbon paper (see Chapter 3).

• Six practical applications of this research (e.g. health surveillance infrastructure) (see Chapter 8).

• Compliance with standard health informatics protocols established by the CDC [23] and

HL7 [55] (see Chapter 8).

• A highly evolved software system capable of analyzing these techniques (see Chapter 9).

Insights provided by these new algorithms:

• The binarization algorithm shows a non-boundary restrictive approach based on comparing the intensity of different regions of the image using a wave trajectory. Prior algorithms depend on such techniques as sliding windows, mean-variance comparisons and interpolations that are shown to under-perform. This new technique not only improves the performance of automated algorithms, but produces images which are enhanced to the human eye as well.

• The lexicon reduction paradigm shows that only a few characters from words are sufficient in determining a topic category. This is analogous to the human interpretation of the content of a form even if some of the characters cannot be read by a person. More specifically, this shows that the contextual acquisition of handwriting content can be accurately represented even with partial information. This is an improvement over algorithms which attempt to achieve perfect recognition as a prerequisite to knowledge representation. This also suggests that humans toggle between recognition and content just as the lexicon reduction algorithm bootstraps the same recognition process with an intermediate statistical interpretative step.

• The automated indexing of semantic content using partially recognized natural language encoded information, mimics the human ability to get the gist of the information on the form. These computations have resemblance to ontological frameworks used in anatomical information science. This is shown with the mapping of partial character recognition information to topic categories for recognition and retrieval of medical data.

A new paradigm for lexicon reduction and information retrieval in the complex situation of handwriting recognition of medical forms. The strategy is novel in their hybridization of linguistics, statistical modeling and handwriting recognition. A series of theoretical and practical recognition rates are provided as evidence. An improvement in raw recognition rate from $\sim$25% of the words on a PCR form to approximately $\sim$33% has been shown. A reduction in false accepts by $\sim$7%, a reduction in error rate by $\sim$10%-25% and a lexicon reduction from 32%-85% were also attained. In addition an information retrieval approach using the lexicon reduction technique showed an increase in document return of 50%. The addition of a category motivated query facilitates about 85% relevant searches at the recall position of 0.1 (see Table 7.1).

In addition, certain computational elements of bootstrapping are consistent with the human interpretation of information in unknown visual context. The training data used for the experiments can be considered incomplete due to the inherent complexities in the definition of a character. While the error rate is high for the machine, it is also high for the human and yet, both systems can still interpret the information.

Unfortunately, the error rate in the handwriting recognition still remains high. This is consistent with approximately 15% of those forms in which humans could not decipher at least 50% of their content. For all words across the remaining 85% of forms, approximately 10.5% of the words contained at least one human unrecognizable character. Examples of the difficulties in interpretation are evident in the taxonomy from Section 1.4. The main difficulties are (i) faded handwriting creating lost strokes, and (ii) shaky handwriting in mobile environments. The situations create error because of the inconsistencies with the training values. Furthermore, while the human still has a higher recognition rate, it does take longer for a human to interpret the handwriting versus reading more clean handwriting. This implies that higher level semantic reasoning is necessary to interpret medical handwriting.

10.2 Future Work

The following is a list of items that needs to be researched and developed in order to complete an operational system:

• *Electrocardiogram (ECG) Category Modeling:* In certain Advanced Life Support (ALS) cases, paramedics may tape an ECG sample printed from the ECG/defibrillation unit to the PCR form. This provides the temporal electrical nervous system firings at various positions in the heart which trigger the cardiac muscles. There are hundreds of such possible rhythms which indicate such things as heart attack or heart disease. This information could be used to assist with form context tagging.

• *Temporal Information Modeling:* All PCR forms contain temporal information involving arrival and departure from the scene, extrication duration and at least two vital signs, at two different times, for comparative purposes. It is conceivable that such information, would be useful in further restricting the possible categories. However, substantial timing information needs to be available to data mine such trends.

• *Form Registration:* Each state is likely to have its own form and, in the future, many different types of forms may be involved. To accommodate this, known formats need to be registered within the system. Only after the appropriate form template is determined can the recognition task begin.

• *Anchor Detection:* Before any recognition task is performed, anchor points on the form indicating the bounds of handwriting text locations must be identified. In addition, the detection and recognition of bubble sheet values, circled items, etc. must be determined.

• *Symbol Recognition:* Medical text often contains symbols that need to be detected, extracted and recognized. Unlike most characters in the English language, symbols can involve several strokes of various sizes and combinations. Simply treating symbols as additional character classes will intuitively degrade recognition performance.

• *Word Separation:* While form lines on the NYS PCR may assist with word separation tasks, other form templates may lack such anchor points. In addition, ambulance movement and emergency environments complicate the expected length and boundaries of words.

• *Word Spotting:* In this task, the input of a set of forms and a word to locate it should produce those forms without brute force recognition of all words. This would greatly improve the indexing ability of medical forms for search engines. This is due to the difficulty of the recognition task as opposed to analysis of ASCII text.

• *Writer Modeling:* From the practical perspective, it is important to improve the recognition rate of medical forms using all possible means. Part of this effort can involve registration of health care professionals handwriting against an identification value. Suppose that a standard health care ID is known and a sample of handwriting from that individual is known. Then models can be constructed based on the individuals writing. Although the emergency environment is expected to produce different handwriting for the same writer, nevertheless, this approach is expected to solve various performance and run-time issues.

• *Relevance Feedback:* Medical form retrieval performance may improve by incorporating human or machine feedback. A human could mark query results as relevant or irrelevant. The system could then construct a better representative query using terms from the marked documents. The topic categories from human indicated relevant documents could also be used to restrict the returned documents to the same or similar categories. Further details on relevance feedback approaches can be found in [102] [108].

10.3 Limitations

Chapter 3 specifies a binarization algorithm which was designed to operate with carbon mesh forms. However, not all forms or environments require this particular algorithm to be used. Therefore, the appropriate binarization algorithm would need to be determined for any registered form. Chapter 5 describes a lexicon reduction strategy that assumes that there is sufficient body of text. In other words, determining a category from a single word is not expected to work. In addition, a PCR deck that does not provide any cohesive phrases under a category implies either that there are no cohesive phrases or that the deck has to be larger. Chapter 7 provides a search engine that expects only two words as input due to the modeling of the cohesive phrases. Involving more than two words may result in an exponential increase in phrase computations depending on the requirements of a system.

Appendix A

Medical Ethics and Information Security

While the technology to design medical recognition systems is becoming a reality, it is also important to note the ethical responsibilities in such a task, particularly when related to epidemiological classification. The invention of such technology as a medical form search engine would not be made available to the public, for privacy and security concerns; instead, the design would be restricted to those medical personnel who are authorized to search for desired information. Therefore, as expected in systems that use this research, and including this research itself, it is imperative that patient confidentiality remain secured [19] [25] [28]. The best way to handle this is by blocking recognition or storage of any patient-specific information during the recognition phase, if such information is available. Note that the medical forms in this study have patient restricted information blanked out. The security of patient information is protected by the HIPAA Privacy Standard [33] [10].

In addition to patient security, there is also a need to protect all medical related inputs and computed outputs. To defend against such computed information from being acquired by an unauthorized party, various cryptographic security protocols must be integrated. However, cryptographic technologies, for information transfer and storage [112], and all publicly available cryptographic algorithms (excluding the impractical Vernam cipher [112]) are vulnerable to time. More specifically, information protected with the most secure of cryptographic algorithms today can likely be accessed using a brute force attack, using a rough estimate of 10-15 years; hence, the National Institute of Standards and Technology (NIST) approves such algorithms as the Advanced Encryption Standard (AES) with a 10-year expiration. To accommodate this situation, a migration step from one cryptographic system (e.g. AES) to a later one must be performed. If a brute force attack upon the public appears reasonably possible, then such a medical analysis system must be disabled until the data can be properly migrated to the latest cryptosystem in a secure and isolated environment. Further discussion on data mining in biomedical applications, as it pertains to the Terrorist Information Awareness (TIA) program developed by the Defense Advanced Research Projects Agency (DARPA) [30] [31] [32], is discussed by Chen, et al. [25].

Appendix B

Parallel Processing

Due to the complexity of the recognition task across a geographic region, various strategies in distributing the work load are necessary.

• *Region Level Parallel Processing:* Suppose New York State needed to recognize a vast number of forms. Automated sites should exist at each hospital rather than transporting all forms to a central location. Once a set of forms is scanned and recognized at the automated site, the data can then be more efficiently centralized (see Figure B.1).

• *Deck Level Parallel Processing:* Suppose that a region site has a great volume of medical forms. For example, it is expected that New York City will have a larger volume of medical documents than Buffalo. However, the time constraints in recognizing all forms within the region is constant. In order to reduce problems with a larger region falling behind, more machines can be used to distribute the decks. Once the system is trained, the recognition procedure for all subsequent forms can be handled independently.

• *PCR Level Parallel Processing:* Suppose that the Region and Deck level distribution is still inefficient. The next step is to parallel process the algorithm itself. This research allows for two additional breakdowns: (i) each of the five handwriting sections on the PCR image

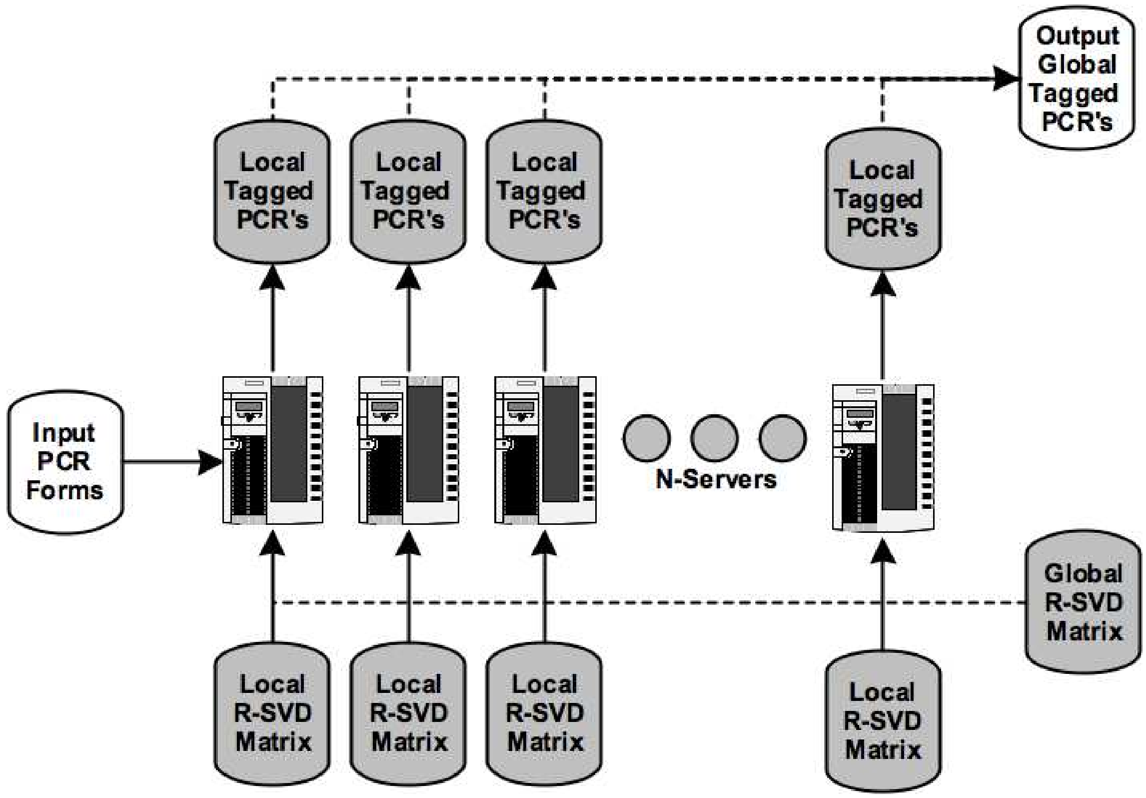


Figure B.1: Parallel Processing

can be binarized independently, and (ii) the initial recognition of confident characters for each word can be handled independently. These tasks can be appropriately distributed to additional machines with higher processing power. Once finished, these machines would report back to the central machine to connect the pieces and continue with the algorithm.

Appendix C

EMS Abbreviations List

Table C.1 contains a list of commonly used abbreviations used by healthcare professionals on PCR forms. They allow more information to fit within form boundaries as well as increase the speed of documentation during the rescue effort. The left column shows the abbreviation used and the right column indicates its respective meaning. General case-sensitivity for the abbreviations are shown.

ABBREVIATION	**MEANING**
#	number
(h)	hypodermic
(R)	rectal
1x	once
2	secondary / due to
2nd	second degree
a.s.	left ear
A.U.	both ears
abd.	abdomen
abdl.	abdominal
ABG	arterial blood gases
adm.	admitted
AEMT	advanced emergency medical technician
ALS	advanced life support
AM	before noon
amp.	ampule
amt.	amount
ant.	anterior

as lib.	as desired
ax.	axillary
B	black
b.i.d	twice daily
B.M.	bowel movement
B.S.	blood sugar
bilat	bilateral
bl.	blood
BGL	blood glucose level
BLS	basic life support
BP	blood pressure
C.	centigrade
c/c	chief complaint
c/o	complaining of
CA	cancer
CAO / COA	conscious alert oriented
cap.	capsule
CCU	critical care unit
CHF	congestive heart failure
ck	check
cl	chronic
cl.	chloride
cm	centimeter
cmp.	compound
CNS	central nervous system
CO2	carbon dioxide
con’t.	continue
COPD	chronic obstructive pulmonary disease
CPT	chest wall percussion
CS	cesarean section
CSF	cerebral spinal fluid
cu.	cubic
CVA	cerebrovascular accident (stroke)
D.O.A	dead on arrival
D.O.B	date of birth
D.T.’s	delirium tremens
D/C	discontinue
D/S	dextrose in saline
D/W	dextrose in water
dil.	dilute
disch.	discharge
dr.	dram

DSD	dry sterile dressing
DTR	deep tendon reflex
Dx	diagnosis
E.D.C.	estimated date of confinement
EMS	emergency medical services
E.S.	emergency service
ea.	each
EKG	electrocardiogram
emerg.	emergency
EMT	emergency medical technician
ETOH	alcohol
ext	extremities
F	fahrenheit
F.B.	foreign body
F.R.O.M	full range of motion
F.U.O	fever undetermined origin
FD	fire department
FHB	fetal heartbeat
fl.	fluid
FR	first responder
FRU	first responder unit
fx	fracture
G.B.	gall bladder
G.C.	gonococcal infection (gonorrhea)
G/W	glucose in water
gal.	gallon
gen.	general
GI	gastro intestinal
Gm.	gram
gr.	grain
GSW	gunshot wound
gtt.	drop
GU	genito-urinary
H2O	water
HEENT	head / eyes / ears / nose /throat
hs	hour of sleep
ht.	height
hx	history
I & O	intake and output
I.M.	intramuscular
ICU	intensive care unit
inc.	incontinent

incl.	include
incp.	incomplete
ing.	inguinal
inspr.	inspiration
int.	internal
invol.	involuntary
IV	intravenous
kg	kilogram
km	kilometer
KVO	keep vein open
(L)	left
L.	liter
lat.	lateral
lb.	pound
liq	liquids
LLQ	left lower quadrant
LMP	last menstrual period
LOC	loss of consciousness
LPM	liters per minute
LPN	licensed practical nurse
LRI	lower respiratory infection
LUQ	left upper quadrant
m	medicines
max.	maximum
med/surg	medical-surgical
meq	milliequivalent
MI	myocardial infarct (heart attack)
min.	minute
mm	milligram
mog	microgram
MVA	motor vehicle accident
n/g	nasal gastric
N/S	normal saline
N/V	nauseau/vomiting
Na	sodium
NaCL	sodium chloride
neg.	negative
NKA	no known allergies
NKDA	no known drug allergies
no.	number
norm.	normal
NPO	nothing by mouth

O2	oxygen
o.s.	left eye
o.u.	both eyes/ each eye
OB	obsterics
oint.	ointment
oz	ounce
p.	pulse
PD	police department
P.E.	physical exam
P.I.D.	pelvic inflammatory disease
P.O.	telephone order
p.r.n.	whenever necessary
PEARL	pupils equal and reactive to light
ped	pediatric
per	by
PERL	pupils equal/reactive to light
PM	afternoon
po	by mouth/orally
po	orally/by mouth
poss.	possible
post	post-operative
post part	post partum
post.	posterior
prep.	preparation
pt	patient
px.	physical
q	every
q.d.	every day
q.h.	every hour
q.l.d.	four times a day
q4h	every four hours
(R)	right
R.	respiration
R.O.M.	range of motion
R/O	rule out
reg.	regular
rehab	rehabilitation
RLQ	right lower quadrant
RUQ	right upper quadrant
RX	prescription
s.o.s.	once if necessary
sat.	saturated

sc.	subcutaneous
seco	second
sep.	separated
SNT	soft-not-tender
so.	solution
sob	short of breath
Sod	sodium
staph	staphylococcus
stat.	at once (immediately)
strep	streptococcus
subl.	sublingual
tab	tablet
TBA	to be admitted
tbsp.	tablespoon
temp	temperature
tid	three times daily
TPR	temperature pulse respiration
tsp	teaspoon
TX	transport
unr	unremarkable
V.O.	verbal order
v.s.	vital signs
vag.	vaginal
vasc.	vascular
via	by way of
vol.	volume
w.	white
W/N/L	within normal limits
wk.	week
WNL	within normal limits
wt.	weight
x1	once / applies to one
x2	twice / applies to two
x3	applies to three
x4	applies to four
y/o	years old
yo	years old
yrs	years

Table C.1: Handwriting abbreviations used on PCRs

Appendix D

The EDGE Project [36]

In the past, it has taken up to 6 weeks to identify any form of outbreak (e.g. whooping cough, SARS) or bio-terrorist attack. As a result, the New York State Department of Health is critically seeking practical, alternative methods for automating the data input of medical information. This urgency, however, is seconded by their primary goal of providing medical care. One such technology developed was a tablet-like device which could store all information digitally. This would, in theory, bypass the need for paper forms; particularly the NYS PCR. The EDGE device, developed by CUBRC, is one such instrument that has been deployed as a beta project [36].

The EDGE (Electronic Data Gathering for EMS) [36], shown in Figure D.1, is a hand-held computer platform developed by the Center for Transportation Research. It is designed for use by pre-hospital care providers (e.g., EMT's and paramedics) with the objectives to:

- Improve the timeliness, quality and quantity of data characterizing the pre-hospital care environment, particularly for cases associated with motor vehicle crash-related trauma.

Figure D.1: EDGE Device

• Provide EMS personnel with real-time support and information such as treatment protocols and prompts for required data.

• Provide a tool to improve pre-hospital care quality.

Western New York (WNY) has served as the beta test site for EDGE development and testing. As of this writing, EDGE units are being used by approximately thirty emergency response agencies, including 14 commercial and 16 volunteer EMS agencies. [36]

While in theory these devices were expected to perform well, some scientists believe they will not be practically ready for at least another decade. Emergency environments are conducive to chemicals, extreme cold and heated temperatures, sea and fire rescue, bodily fluids, as well as complex physical movement such as extrication, mass casualty incidents, terrorist and biochemical attack. During these 18 months, the units, while of excellent design and concept, have frustrated some health-care professionals and have, at times, broken down. Even one failure is unacceptable in the highly stressful rescue environment. Therefore there are concerns regarding the current practicality and cost.

While to some this technology does not appear to be able to replace paper documents yet, the need to capture real-time health data, which can lag up to four years behind, is strongly desired due to its potential value in epidemiological systems.

Bibliography

[1] Abdelmalek, N.N. Round Off Error Analysis for Gram-Schmidt Method and Solution of Linear Least Squares Problems. BIT Numerical Mathematics, Volume 11, Issue 4, pp. 345-367, 1971.

[2] Agarwal, A., Gupta, A. and Hussein, K. Bank Check Analysis and Recognition by Computers. Handbook of Character Recognition and Document Image Analysis. Eds. Bunke, H. and Wang, P.S.P. World Scientific Publishing Company, 1997.

[3] American Society of Health-System Pharmacists. http://www.ashp.org

[4] Alter O., Brown P.O., Botstein D. Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling. Proc Natl Acad Sci USA; 97:10101-06, 2000.

[5] Anton, H. Elementary Linear Algebra. Chapter 6: Inner Product Spaces and Chapter 7: Eigenvalues, Eigenvectors. John Wiley and Sons, Inc., 1994.

[6] Bahl, L.R., Jelinek, F., and Mercer, R.L. A Maximum Likelihood approach to Continuous Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence; 5(2), 1983.

[7] Bahlmann, C., Haasdonk, B., and Burkhardt, H. On-Line Handwriting Recognition with Support Vector Machines - A Kernel Approach. International Workshop on Frontiers in Handwriting Recognition, 2002.

[8] Bayer, T., Bohnacker, U. and Renz, I. Information Extraction from Paper Documents. Handbook of Character Recognition and Document Image Analysis. Eds. Bunke, H. and Wang, P.S.P. World Scientific Publishing Company, 1997.

[9] Beigi, H. Processing, Modeling and Parameter Estimation of the Dynamic On-Line Handwriting Signal. Proceedings World Congress on Automation, 1996.

[10] Berman, J.J. Confidentiality Issues for Medical Data Miners. Artificial Intelligence Med. 26(1-2):25-36, 2002.

[11] Berry M.W. Large-scale sparse singular value computations. International Journal of Supercomputer Applications; 6:13-49, 1992.

[12] Berry M.W., Dumais S.T., and Obrien G.W. Using Linear Algebra for Intelligent Information-Retrieval. Siam Review; 37:573-95, 1995.

[13] Black, P.E., ed. ”Levenshtein Distance”. Algorithms and Theory of Computation Handbook; CRC Press LLC, from Dictionary of Algorithms and Data Structures, NIST, 1999.

[14] Blum, J.R., Rosenblatt, J.I. Probability and Statistics. Chapter 4: Random Variables and Their Distributions; Chapter 6: Expectations, Moment Generating Functions, and Quantiles. W.B. Saunders Company. USA, 1972.

[15] Blumenstein, M., Verma, S. A Neural Based Segmentation and Recognition Technique for Handwritten Words. IEEE International Conference on Neural Networks, 1998.

[16] Brown, M.K., and Ganapathy, S. Preprocessing Techniques for Cursive Word Recognition. Pattern Recognition, Vol. 16 No. 5, 1983.

[17] Byun, H., and Lee, S.W. Applications of Support Vector Machines for Pattern Recognition: A Survey. Lecture Notes in Computer Science; Springer, 2002.

[18] Caesar, T., Gloger, J.M., and Mandler, E. Using Lexical Knowledge for the Recognition of Poorly Written Words. Third International Conference on Document Analysis and Recognition. Volume 2. pp. 915-918, 1995.

[19] Callahan, J.C. Ethical Issues in Professional Life. USA. Oxford University Press, Inc., 1988.

[20] Carbonnel, S. and Anquetil, E. Lexicon Organization and String Edit Distance Learning for Lexical Post-Processing in Handwriting Recognition., IWFHR-9,p.462-467, 2004.

[21] Chu-Carroll, J., and Carpenter, B. Dialogue Management in Vector-Based Call Routing. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics. pp. 256-262, 1999.

[22] Chu-Carroll, J., and Carpenter, B., Vector-Based Natural Language Call Routing. Computational Linguistics. Vol. 25, No. 3, pp. 361–388, 1999.

[23] Centers for Disease Control (CDC). National Electronic Disease Surveillance system (NEDSS). Public Health Information Network (PHIN).

[24] Chen, M.Y., Jundu, A., and Zhou, J. Off-Line Handwritten Word Recognition Using a Hidden Markov Model Type Stochastic Network. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994.

[25] Chen, H., Fuller, S.S., Friedman, C., Hersh, W. Medical Informatics. Knowledge Management and Data Mining in Biomedicine. Springer Science+ Business Media, Inc, 2005.

[26] Cho, S.B., and Kim, J.H. Applications of Neural Networks to Character Recognition. Pattern Recognition, 1991.

[27] Cho, S.B. Neural-Network Classifiers for Recognizing Totally Unconstrained Handwritten Numerals. IEEE Transactions on Neural Networks, 1997.

[28] Churchill, T. Principles and Practice of Public Health Surveillance; 2nd Edition. USA. Oxford University Press, Inc., 2000.

[29] Coonan, K.M., M.D. Medical Informatics Standards Applicable to Emergency Department Information Systems: Making Sense of the Jumble. Informatics Standards for ED Systems. University of Utah School of Medicine, 2004.

[30] Defense Advanced Research Project Agency. Total Information Awareness Program (TIA) System Description Document (SDD), Version 1.1, July 2002.

[31] Defense Advanced Research Project Agency. Information Awareness Office. Report to Congress Regarding the Terrorist Information Awareness Program: In Response to Consolidated Appropriations Resolution. Pub.L. No 108-7, Division M, 111(b) Detailed Information, May 2003.

[32] Defense Advanced Research Project Agency. Office of the Inspector General, Information Technology Management. Terrorist Information Awareness Program. D-2004-033, December 2003.

[33] Department of Health and Human Services. http://www.hhs.gov/ocr/hipaa/, 2006.

[34] Deerwester, S., Dumais, S.T., Furnas, G.Q., Landauer, and, T.K., Harshman, R. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science. 41(6):391-407, 1990.

[35] Public Health and Safety: Disease Surveillance. CUBRC. http://www.cubrc.org/, 2006.

[36] EDGE Device. CUBRC. http://www.cubrc.org/, 2006.

[37] Fagan, J. The Effectiveness of a Non-Syntactic Approach to Automatic Phrase Indexing for Document Retrieval. Journal of the American Society for Information Science, 40: 115-132, 1989.

[38] Favata, J.T. Offline General Handwritten Word Recognition Using an Approximate BEAM Matching Algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI); 23 (9); pp1009-1021, 2001.

[39] Feng, S.L. and Manmatha, R. Classification Models for Historic Manuscript Recognition. Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR), 2005.

[40] Gader, P.D., Keller, J.M., Krishnapuram, R., and Chiang, J.H. Neural and Fuzzy Methods in Handwriting Recognition, 1997.

[41] Garcia A.E. Large-Amplitude Nonlinear Motions in Proteins. Phys Rev Lett; 68:2696-99., 1992.

[42] Gatos, B., Pratikakis, I., and Perantonis, S.J. Adaptive Degraded Document Image Binarization. The Journal of the Pattern Recognition Society. Elsevier Ltd., 2005.

[43] Gilmour, I. Nonlinear Model Evaluation: Shadowing, Probabilistic Prediction and Weather Forecasting. PhD/DPhil Thesis, University of Oxford, 1999.

[44] Golub, G.B., Van Loan, C.E. Matrix Computations. 2nd Edition. John Hopkins University Press, 1989.

[45] Goodman, K.S. Reading: A Psycholinguistic Guessing Game. Language and Literacy: The Selected Writings of Kenneth S. Goodman. pp.33-43, 1967.

[46] Govindaraju, V., and Xue, H. Fast Handwriting Recognition for Indexing Historical Documents. First International Workshop on Document Image Analysis for Libraries (DIAL), 2004.

[47] Govindaraju, V., Slavik, P., and Xue, H. Use of Lexicon Density in Evaluating Word Recognizers. IEEE Trans PAMI, Vol. 24, No.6, pp. 789-800. 2002.

[48] Greengrass, E. Information Retrieval: A Survey. National Security Agency, Report Number TR-R52-02-96, 1997.

[49] Guillevic, D., and Suen, C.Y. Cursive Script Recognition: A fast reader scheme. Proceedings of the Second International Conference on Document Analysis and Recognition (ICDAR), 1993.

[50] Guillevic, D., Nishiwaki, D., and Yamada, K. Word Lexicon Reduction by Character Spotting. Seventh International Workshop on Frontiers in Handwriting Recognition. Amsterdam, 2000.

[51] Hatami, Safar., Hosseini, R., Kamarei, and M., and Ahmadi, H. Wavelet Based Fingerprint Image Enhancement. IEEE International Symposium on Circuits and Systems (ISCAS), 2005.

[52] He, J.Z.; Du, M.H.; Pei, S.W.; and Wan, Q., Face Recognition based on Spectroface and Uniform Eigenspace SVD for One Training Image per Person," Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on , vol.8, no.pp. 4842- 4845 Vol. 8, 18-21, 2005.

[53] Hersh, W., Price, S., and Donohoe, L. Assessing Thesaurus-Based Query Expansion Using the UMLS Metathesaurus. American Medical Informatics Association (AMIA) Symposium, 2000.

[54] Hersh, W.R. Information Retrieval: A Health and Biomedical Perspective. 2nd Edition. Springer-Verlag, New York, Inc. USA, 2003.

[55] Health Level (HL7) Group. Health Level Seven; An Application Protocol for Electronic Data Exchange in Healthcare Environments, Version 2.2. Technical Report, Health Level Seven, Inc. USA, 1994.

[56] Hsu, D.S., and Huang, W.M., Thakor, N.V. StylPen: On-line Adaptive Canceling of Pathological Tremor for Computing Pen Handwriting. IEEE Transactions on Biomedical Engineering, 1998.

[57] Hu, J. Brown, M.K., and Turin, W. HMM Based Online Handwriting Recognition. IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI), 1996.

[58] Huang, L. A Survey on Web Information Retrieval Technologies, 2000.

[59] JAMA: A Java Matrix Package. Version 1.0.2. National Institute of Standards and Technology (NIST). http://math.nist.gov/javanumerics/jama/, 2006.

[60] $Java^{TM}$ Programming Language. Sun Microsystems, Inc. http://www.java.sun.com/, 2001+.

[61] Jones, K.S. A Statistical Interpretation of Term Specificity and its Application in Retrieval. Journal of Documentation, 28(1):11-20, 1972.

[62] Jones, K.S. and Willet, P. Readings in Information Retrieval, San Francisco: Morgan Kaufmann, 1997.

[63] Kamel, M., Zhao, Extraction of Binary Character/Graphics Images from Grayscale Document Images. CVGIP: Graphics Models Image Processing; 55 (3), 1993.

[64] Kaufmann, G., Bunke, H., and Hadorn, M. Lexicon Reduction in an HMM-Framework Based on Quantized Feature Vectors. Proc. International Conference on Document Analysis and Recognition (ICDAR) 97; Vol. 2 , 18-20, pp. 1097-1101, 1997.

[65] Kim, G., and Govindaraju, V. Bank Check Recognition using Cross Validation between Legal and Courtesy Amounts. International Journal on Pattern Recognition and Artificial Intelligence, 1996.

[66] Kim, G., and Govindaraju, V.: A Lexicon Driven Approach to Handwritten Word Recognition for Real-Time Applications. IEEE Trans. PAMI 19(4): 366-379, 1997.

[67] Koerich, A.L., Sabourin, R., and Suen, C.Y. Fast Two-Level HMM Decoding for Large Vocabulary Handwriting Recognition. International Workshop on Frontiers in Handwriting Recognition, 2004.

[68] Koerich, A.L., Sabourin, R., and Suen, C.Y. Large Vocabulary Off-Line Handwriting Recognition: A Survey. Pattern Analysis and Applications, 2003.

[69] Koerich, A.L.; Leydier, Y.; Sabourin, and R.; Suen, C.Y. A Hybrid Large Vocabulary Handwritten Word Recognition System using Neural Networks with Hidden Markov Models. IWFHR-8; 6-8, 2002.

[70] Landauer, T.K., Foltz, P.W., and Laham, D. Introduction to Latent Semantic Analysis. Discourse Processes, 25, pp. 259-284, 1998.

[71] Langville, A. and Meyer, C.D. A Survey of Eigenvector Methods of Web Information Retrieval, 2003.

[72] Langville, A. and Meyer, C.D. The use of the Linear Algebra by Web Search Engines, 2004.

[73] Larson, Hostetler, and Edwards. Calculus with Analytic Geometry. Chapter 13 : Section 13.9. Fifth Edition. D.C. Heath and Company. USA, 1994.

[74] Lawson, C.L. and Handson, R.J. Solving Least SquaresProblems. Prentice-Hall, 1974.

[75] Leedham, G., Varma, S., Patankar, A., and Govindaraju, V. Separating Text and Background in Degraded Document Images - A Comparison of Global Thresholding Techniques for Multi-Stage Thresholding. Proceedings Eighth International Workshop on Frontiers of Handwriting Recognition, 2002.

[76] Liao, P.S., Chen, T.S., and Chung, P.C. A Fast Algorithm for Multilevel Thresholding. Journal of Information Science and Engineering, 2001.

[77] Liu, C.L., and Marukawa, K. Global Shape Normalization for Handwritten Chinese Character Recognition: A New Method. International Workshop on Frontiers of Handwriting Recognition, 2004.

[78] Luhn, H. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development, 1: 309-317, 1957.

[79] Madhavanath, S. Govindaraju, and V., Ramanaprasad, V. Reading Handwritten US Census Forms. Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR), 1995.

[80] Madhvanath, S. The Holistic Paradigm in Handwritten Word Recognition and its Application to Large and Dynamic Lexicon Scenarios. University at Buffalo Computer Science and Engineering. Ph.D. Dissertation, 1997.

[81] Madhvanath, S., Krpasundar, V., and Govindaraju, V. Syntactic Methodology of Pruning Large Lexicons in Cursive Script Recognition. Journal. The Journal of the Pattern Recognition Society. Pattern Recognition 34 (2001); Elsevier Science, 2001.

[82] Marti, U.V., and Bunke, H. Using a Statistical Language Model to Improve the Performance of an HMM-based Cursive Handwriting Recognition Systems. World Scientific Series in Machine Perception and Artificial Intelligence Series, 2001.

[83] McDermott, D. Artificial Intelligence Meets Natural Stupidity. ACM SIGART Bulletin. Issue 57. pp.4-9, 1976.

[84] MEDLINE. http://www.medlineplus.gov

[85] Milewski, R. and Govindaraju, V. Medical Word Recognition Using a Computational Semantic Lexicon. Eighth International Workshop on Frontiers in Handwriting Recognition. Canada, 2002.

[86] Milewski, R. and Govindaraju, V. Handwriting Analysis of Pre-Hospital Care Reports. IEEE Proceedings. Seventeenth IEEE Symposium on Computer-Based Medical Systems (CBMS), 2004.

[87] Milewski, R. and Govindaraju, V. Extraction of Handwritten Text from Carbon Copy Medical Forms. Document Analysis Systems (DAS). Springer-Verlag, 2006.

[88] Mitchell, T.M. Machine Learning. Chapter 4: Artificial Neural Networks. The McGraw-Hill Companies, Inc., 1997.

[89] Moler, C., and Morrison, D.: Singular Value Analysis of Cryptograms. American Mathematical Monthly 90 pp. 7887, 1983.

[90] $MySQL^{®}$ Relational Database Management System (RDBMS). MySQL AB. http://www.mysql.com/, 2001+.

[91] Nakai, M., Akira, N., Shimodaira, H., Sagayama, and S. Substroke Approach to HMM-Based On-Line Kanji Handwriting Recognition. Sixth International Conference on Document Analysis and Recognition, 2001.

[92] Nelson, S.J. Schopen, M., Savage, A.G. Schulman, J.L., and Arluk, N. The MeSH Translation Maintenance System: Structure, Interface Design, and Implementation. National Library of Medicine (NLM). National Institutes of Health (NIH). Medical Subject Headings (MeSH). Proceedings of the 11th World Congress on Medical Informatics, 2004.

[93] Niblack, W. An Introduction to Digital Image Processing. Englewood Cliffs, N.J. Prentice Hall, 1986.

[94] Niyogi, D., Srihari, S. and Govindaraju, V. Analysis of Printed Forms. Handbook of Character Recognition and Document Image Analysis. Eds. Bunke, H. and Wang, P.S.P. World Scientific Publishing Company, 1997.

[95] National Library of Medicine. PubMed Stop List.

[96] Oh, I.S., and Suen, C.Y. Distance Features for Neural Network-Based Recognition of Handwritten Characters. International Journal on Document Analysis and Recognition.,1998.

[97] Okuda, T., Tanaka, E., and Kasai, T. A Method for the Correction of Garbled Words based on the Levenshtein Distance. IEEE Transactions on Computers, Col. C-25, No. 2, 1976.

[98] Oren, E., Shaffer, ER., and Guglielmo, BJ. Impact of Emerging Technologies on Medication Errors and Adverse Drug Events. American Journal of Health Syst Pharm. 60:1447-1458, 2003.

[99] Otsu, N. A Threshold Selection Method from Gray-Level Histogram. IEEE Transactions on System Man Cybernetics, Vol. SMC-9, No. 1, 1979.

[100] Porter, M.F. An Algorithm for Suffix Stripping. Program, 14: pp. 130-137, 1980.

[101] Prescription Drug Errors Increasing. National Health Review. Vegetus Publications 1995 and Gale Group 2004.

[102] Raghavan, P. Information Retrieval Algorithms: A Survey. Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithm, 1997.

[103] Rapaport, William J., and Ehrlich, Karen. A Computational Theory of Vocabulary Acquisition. Lucja M. Iwanska, and Stuart C. Shapiro (eds.), Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language (Menlo Park, CA/Cambridge, MA: AAAI Press/MIT Press): 347-375, 2000.

[104] Richards J.A., Remote Sensing Digital Image Analysis. New York: Springer-Verlag, 1993.

[105] Rijsbergen, C.J. van, Robertson, S.E. and Porter, M.F. New Models in Probabilistic Information Retrieval. London: British Library, 1980.

[106] Romo T.D., Clarage J.B., Sorensen D.C., Phillips G.N., Jr. Automatic identification of discrete substates in proteins: singular value decomposition analysis of time-averaged crystallographic refinements. Proteins; 22:311-21, 1995.

[107] Rosse, C. Kumar, A. Leonardo, Mejino, J.L.V. Cook, D.L., Detwiler, L.T. and Smith, B. A Strategy for Improving and Integrating Biomedical Ontologies. Proceedings of AMIA Symposium 2005, Washington DC, pp. 639-643, 2005.

[108] Ruthven, I., and Lalmas, M. A Survey on the Use of Relevance Feedback for Information Access Systems. The Knowledge Engineering Review. Cambridge University Press. Volume 18 Issue 2: pp. 95-145, 2003.

[109] Salton, G. Introduction to Modern Information Retrieval. New York. McGraw-Hill, 1983.

[110] Salton, G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, 1989.

[111] Sauvola, J. Seppanen, T. Haapakoski, and Pietikainen, M. Adaptive Document Binarization. In International Conference on Document Analysis and Recognition, Volume 1, 1997.

[112] Schneier, B. Applied Cryptography. John Wiley & Sons, Inc., 1996.

[113] Schurmann, J., et al. Document Analysis-From Pixels to Contents. Processing IEEE Vol. 80 No. 7, 1992.

[114] Seeger, M., and Dance, C. Binarising Camera Images for OCR. Xerox Research Center Europe. 6th International Conference on Document Analysis and Recognition, 2001.

[115] Shi, Z. and Govindaraju, V. Character Image Enhancement by Selective Region-Growing. Pattern Recognition Letters, 17, 1996.

[116] Sinha, R.M.K., and Prasada, B. Visual Text Recognition through Contextual Processing. Pattern Recognition Vol 21, No 5. pp. 463-479, 1988

[117] Smith, B., and Ceusters, W. A Methodology for the Migration of Medical Terminologies to Electronic Health Records. Proceedings of AMIA Symposium 2005. pp. 704-708, 2005.

[118] Smith, B., Mejino J.L.V., Schulz, S., Kumar, A. and Rosse, C. Anatomical Information Science in A. G. Cohn and D. M. Mark (eds.), Spatial Information Theory. Proceedings of COSIT 2005 (Lecture Notes in Computer Science), Berlin/Heidelberg/New York: Springer, pp. 149164, 2005.

[119] Sonka, M., Hlavac, V., and Boyle, R. Image Processing, Analysis, and Machine Vision; 2nd Edition. PWS Publishing, 1999.

[120] Srihari, S.N., Hull, J.J., and Choudhari, R. Integrating Diverse Knowledge Sources in Text Recognition. ACM Transactions on Office Information Systems. Vol.1 No.1 pp. 68-87, 1983.

[121] Strang, S. Linear Algebra and its Applications, 2nd Edition, Academic Press, New York, 1980.

[122] Sauvola, J. Seppanen, T. Haapakoski, and Pietikainen, M. Adaptive Document Binarization. In International Conference on Document Analysis and Recognition, Volume 1, 1997.

[123] Suen, C.Y. N-gram Statistics for Natural Language Understanding and Processing. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 1, No.2, pp 164-172, 1979.

[124] Suen, C.Y., Nadal, C., Mai, T.A., Legault, R., Lam, L. Recognition of Totally Unconstrained handwritten numerals based on the Concept of Multiple Experts. Frontiers in Handwriting Recognition, C.Y. Suen, Ed., in Proc. Int. Workshop on Frontiers in Handwriting Recognition. Canada. pp. 131-143, 1990.

[125] Tappert, C.C., Suen, C.Y., and Wakahara, T. The State of the Art in On-Line Handwriting Recognition IEEE Trans. PAMI Vol. 12 No. 8, 1990.

[126] Tomai, C., Zhang, B., and Govindaraju, V. Transcript Mapping for Historic Handwritten Document Images. Proceedings Eighth International Workshop on Frontiers of Handwriting Recognition, 2002.

[127] Trefethen, L. and Bau, D. III. Numerical Linear Algebra. SIAM, 1997

[128] Trier, O.D. and Taxt, T. Evaluation of Binarization Methods for Document Images. IEEE Trans. PAMI, 17 (3): 312-31,; 1995.

[129] Unified Medical Language System (UMLS). National Library of Medicine (NLM). National Institutes of Health (NIH).

[130] Vinciarelli, A., Bengio, S., and Bunke, H. Offline Recognition of Unconstrained Handwritten Texts using HMMs and Statistical Language Models. IEEE Trans PAMI, 2004.

[131] The White House. The National Strategy for Homeland Security: Office of Homeland Security. Emergency Preparedness and Response. http://www.whitehouse.gov/homeland/book/, 2002.

[132] William, H., Buckley, C, Leone, TJ, and Hickam, D. OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. Proceedings of the ACM SIGIR, 1994.

[133] Wolf, C., Jolion, J.M., and Chassaing, F. Text Localization, Enhancement and Binarization in Multimedia Documents. 16th International Conference on Pattern Recognition (ICPR), 2002.

[134] Wolf, C., and Doermann, D. Binarization of Low Quality Text using a Markov Random Field Model. 16th International Conference on Pattern Recognition (ICPR), 2002.

[135] Western Regional Emergency Medical Services. Bureau of Emergency Medical Services. New York State (NYS) Department of Health (DoH). Prehospital Care Report v4.

[136] Wu, V. and Manmatha, R. Document Image Clean-Up and Binarization. Proc. SPIE Symposium on Electronic Imaging, 1998.

[137] Xu, L., Krzyzak, A., and Suen, C.Y. Method of Combining Multiple Classifiers and Their Applications to Handwriting Recognition. IEEE Transactions on Systems, Man, and Cybernetics. Vol.22, No. 3, 1992.

[138] Xue, H., and Govindaraju, V. Stochastic Models Combining Discrete Symbols and Continuous Attributes - Application in Handwriting Recognition. Proceedings of 5th IAPR International Workshop on Document Analysis Systems. pp. 70-81, 2002.

[139] Xue, H., and Govindaraju, V. On the Dependence of Handwritten Word Recognizers on Lexicons. IEEE Trans. PAMI, Vol. 24, No. 12, pp. 1553-1564, 2002.

[140] Yang, Y., and Chute, C.G. An Example-Based Mapping Method for Text Categorization and Retrieval. ACM Transactions on Information Systems. Vol. 12, No. 3, pp. 252-277, 1994.

[141] Yang, Y., and Yan, H. An Adaptive Logical Method for Binarization of Degraded Document Images. The Journal of the Pattern Recognition Society, 2000.

[142] Yanowitz, S.D. and Bruckstein, A.M. A New Method for Image Segmentation. Computer Vision Graphics and Image Processing Vol. 46 No. 1, 1989.

[143] Zimmermann, M. and Mao, J. Lexicon Reduction using Key Characters in Cursive Handwritten Words. Pattern Recog. Letters; Vol 20, pp. 1297-1304, 1999.