תקציר

בניגוד לאמונה הרווחת, התפתחות המחשבים והקטנת מימדיהם לא גרמה לנטישה של כתב-היד באופן מוחלט, והשימוש בדף ועט עודם נפוצים להעברת אינפורמציה בחיי היום-יום. המרת כתב-יד לצורה דיגיטלית טומנת בחובה יתרונות רבים. טקסט ממחושב לא רק מאפשר שיכפול יעיל אלא גם יכולות עיבוד נרחבות כגון, עריכה חיפוש, מיון, שיתוף וכיוצא בזה.

בעיית הזיהוי הממוחשב של כתב-היד בשפה הערבית הינה בעיה מאתגרת וזאת מאחר והשפה הערבית נכתבת בצורה מחוברת. לאחר תקופה ממושכת בה חוקרים התעניינו בזיהוי של כתב-יד לטיני וסיני ושבהם הצליחו להשיג אחוזי זיהוי גבוהים, בשנים האחרונות ניכרת התעניינות רבה בשפה הערבית.

בעוד שנדרשים ביצועי זמן אמת במערכות אשר מבצעות זיהוי כתב-יד מקוון, גישות קונבנציונליות של זיהוי טקסט בדרך כלל ממתינות עד לסיום הכתיבה בכדי להתחיל בתהליך הזיהוי ובניתוח הטקסט. דחית תהליך הזיהוי עד לסיום הכתיבה אינו מאפשר ניצול של פרק זמן הכתיבה לביצוע זיהוי, דבר אשר גורם להאטת המערכת ומונע מימוש יישומים מתקדמים כגון השלמה ותיקון איות אוטומטי תוך כדי כתיבה.

במסגרת עבודה זו אנו מציעים שיטה לזיהוי כתב-יד בשפה הערבית בזמן אמת. אנו ממחישים את היכולת לבצע תהליכי סגמנטציה וזיהוי אשר דורשים זמן עיבוד ממושך, תוך כדי ניצול זמן הכתיבה.

תהליך הסגמנטציה הינו מבוסס זיהוי, כלומר, הוא משתמש במידת הדימיון בין הקטעים השונים של משיכת העט על מנת להחליט אם קטע נתון מייצג אות כלשהי ובהתאם לכך מחליט היכן ממוקמות נקודות הסגמנטציה. מערכת הסגמנטציה מכילה שלושה שלבים עיקריים. בשלב הראשון, אשר מתבצע תוך כדי הכתיבה, המערכת מנסה לזהות קטעים אנכיים המתבסס על תכונות מורפולוגיות. בתוך קטעים אלה מועמדות נקודות סגמנטציה. נקודות הסגמנטציה משרות חלוקה של משיכת העט למספר קטעים, ובעזרת המערכת לזיהוי אותיות, קטעים אלה מקבלים ציונים אשר מסמלים את מידת הדמיון בינם לבין האותיות הקיימות במסד הנתונים. בשלב השני, נקודות סגמנטציה מסוננות ומתבצע תהליך של חישוב מחדש של ציוני הדימיון. בשלב השלישי, המערכת בוחרת את קבוצת נקודות הסגמנטציה הסופיות בעזרת מספר אלגוריתמי בחירה.

ממערכת זיהוי האותיות נדרש זמן תגובה קצר ביותר מאחר והסגמנטציה מתבצעת בזמן אמת, תוך כדי כתיבה. לחישוב המרחק בין הצורות השונות, המערכת משתמשת במטריקת "מרחק מובילי החול" (earth mover's distance). מחקרים רבים הראו כי קיימת מידת התאמה גבוה בין מטריקה זו למרחק התפיסתי. הזיהוי המהיר מתבצע על-ידי שיבוץ יעיל של התבניות למרחב האדוות (wavelets), אשר בו מתאפשר חישוב מקורב מהיר של מטריקה זו ובנוסף מאפשר חיפוש התבניות הדומות ביותר הוא בצורה יעילה, מאחר ומרחב זה משמר את אי-שוויון המשולש.

אנו מראים כי המידע אותו המערכת מצליחה לדלות תוך כדי התקדמות הכתיבה מאפשר צמצום משמעותי של גודל המילון הפוטנציאלי אשר בו יש לחפש בתהליך מאוחר יותר של זוהוי מדויק יותר של מילים בשיטה ההוליסטית.

בעקבות מחקר זה, נכתבו שני מאמרים מדעיים, בשיתוף עם דר' ראיד סעאבנה, והתקבלו לפרסום בשני כנסים בינלאומיים בנושא זיהוי תבניות וזיהוי כתב יד.

אוניברסיטת תל-אביב

הפקולטה להנדסה ע"ש איבי ואלדר פליישמן

בית הספר לתארים מתקדמים ע"ש זנדמן-סליינר

סגמנטציה וזיהוי כתב יד מקוון בשפה הערבית בזמן אמת

חיבור זה הוגש כעבודת מחקר לקראת התואר "מוסמך אוניברסיטה" בהנדסת חשמל

על ידי

ג'ורג' קור

העבודה נעשתה בבית הספר להנדסת חשמל

במחלקה למערכות

בהנחיית פרופ' דנה רון וד"ר ראיד סעאבנה

אלול תשע"ד

אוניברסיטת תל-אביב

הפקולטה להנדסה ע״ש איבי ואלדר פליישמן

בית הספר לתארים מתקדמים ע״ש זנדמן-סליינר

סגמנטציה וזיהוי כתב יד מקוון בשפה הערבית בזמן אמת

חיבור זה הוגש כעבודת מחקר לקראת התואר ״מוסמך אוניברסיטה״ בהנדסת חשמל

על ידי

ג׳ורג׳ קור

אלול תשע״ד