### DataSet

### מטרה:

הכנת Dataset אופטימלי לפרויקט, שיאפשר ייצוג מדויק של רצפי תנועות יד לאורך זמן, תוך שמירה על גמישות מבנית ונגישות גבוהה לשימוש ישיר בתהליך האימון של המודל.כל דוגמה מייצגת רצף של סימוני ידיים לאורך זמן, כאשר כל סימון יד הוא רשימה של 21 נקודות תלת־ממדיות (x, y, z) שמייצגות את מיפוי היד בפריים אחד מתוך סרטון קצר.

מטלות:

1. **פורמט הdataset** -לעצב פורמט ל-dataset שיתמוך בכך שכל דוגמה יכולה להיות באורך שונה. כלומר, מספר הפריימים משתנה, ולכן מספר שורות ה-landmarks שונות, כאשר כל שורה מכילה את כל 21 סימוני היד עבור פריים אחד.

2.**יצירת דוגמאות חדשות** - פיתוח תוכנית הקלטה אינטראקטיבית שתאפשר איסוף ויצירה מהירה של דוגמאות חדשות. הכלי יאפשר הקלטת דוגמאות באורכים משתנים (מספר פריימים שונה) עם שמירה אוטומטית של כל דוגמה ב-dataset, תוך התאמה מלאה למבנה הייחודי של הנתונים.

3. **שיפור הdataset** - הוספת מנגנון לבדיקה מיידית של כל דוגמה חדשה לאחר ההקלטה, כך שתוצג כסרטון בהילוך איטי יחד עם תחזיות של מודל ראשוני על כל פריים. מנגנון זה יאפשר לזהות דוגמאות שהמודל מתקשה בהן, ולהוסיף אותן ל-dataset לצורך שיפור האימון העתידי. בנוסף, תתווסף אפשרות לבחירה האם לשמור כל דוגמה חדשה ב-dataset או לא.

4. **טעינת הdatset לאימון** - יצירת מחלקה שאחראית על טעינת הקבצים והכנתם בפורמט המתאים, על מנת לאפשר חיבור חלק ותקני של הדאטה לתהליך האימון של המודל.

פורמט הdataset:

**בעיה** -בניגוד ל־Datasets רגילים המבוססים על דוגמאות סטטיות (למשל שורה אחת ב־CSV לכל תמונה), הדאטה בפרויקט זה מורכב מדוגמאות דינמיות. כל דוגמה היא רצף של תנועות יד מתוך סרטון קצר, כלומר סדרת פריימים עם סימוני יד (landmarks) לאורך זמן.

עקב כך, לא ניתן לייצג את כל הדאטה בקובץ CSV אחד, מאחר שמספר השורות (הפריימים) שונה מדוגמה לדוגמה, מה שהופך את השימוש בפורמט אחיד למסובך ואף לא ישים.

**פתרון/תוצאה -** כדי להתמודד עם האתגר, הוחלט לשמור כל דוגמה כקובץ CSV נפרד בתיקיית data.  
כל קובץ מייצג דוגמה אחת, ומכיל שורות לפי מספר הפריימים שבה, כאשר כל שורה כוללת את סימוני היד של פריים יחיד בפורמט הבא:  
 hand-side,label,id,x0,y0,z0,...,x20,y20,z20  
 (צד היד, התווית,מספר פריים , וקואורדינטות תלת־ממד של 21 נקודות ביד).

שיטה זו מאפשרת גמישות מלאה באורך הדוגמאות, ניהול ברור של הנתונים, וטעינה פשוטה וישירה של הדאטה לתוך תהליך האימון של המודל.

יצירת דוגמאות חדשות:

**בעיה** -לא עמד לרשותנו Dataset חיצוני שמכיל רצפים של תנועות יד בפורמט הרצוי (landmarks מפריימים בווידאו), ולכן לא הייתה דרך לאסוף דוגמאות מותאמות לפרויקט בצורה מסודרת ואפקטיבית. נדרש כלי שיאפשר להקליט דוגמאות בצורה מהירה, אינטראקטיבית ויעילה, תוך שמירה אוטומטית במבנה המתאים.

**פתרון/תוצאה** - פותח כלי איסוף ייעודי שמפעיל את המצלמה, מזהה את תנועות היד בזמן אמת, ומאפשר הקלטת דוגמאות לפי קטגוריה (אבן/נייר/מספריים) באמצעות לחיצה על מקשי קיצור. במהלך ההקלטה נאספים סימוני היד (landmarks) מכל פריים, ולאחר עצירה נשמרים הסימונים באופן אוטומטי כקובץ CSV נפרד בתיקיית הדאטה, במבנה אחיד המתאים לטעינה לאימון.

כלי זה איפשר בנייה מהירה ויעילה של Dataset מותאם אישית לצרכי הפרויקט.

שיפור הdataset:

**בעיה -** לאחר איסוף דוגמאות ראשוני, לא הייתה דרך לדעת באופן מדויק באילו סוגי דוגמאות המודל מתקשה לזהות נכון את התנועות. למשל, זוויות מסוימות של היד בכל סימן או מצבים אחרים שגורמים לשגיאות בזיהוי. מכיוון שזה קריטי לשפר את ביצועי המודל, נדרש מנגנון שיאפשר לזהות בזמן אמת את המקרים בהם המודל נכשל, כדי לאסוף דוגמאות חדשות דומות ולהעשיר את ה־Dataset לצורך אימון עתידי איכותי יותר.

**פתרון / תוצאה -** ללאחר אימון ראשוני, הוספנו לכלי הקלטת דוגמאות חדשות אפשרות להצגת הדוגמה שהוקלטה בהילוך איטי, כאשר על כל פריים מוצג חיזוי המודל. בנוסף, ניתנת למשתמש האפשרות לבחור האם לשמור את הדוגמה או לבטל אותה.

כך ניתן לזהות במדויק דוגמאות שהמודל מתקשה לחזות (למשל זוויות מסוימות או מצבים מורכבים), לאסוף דוגמאות נוספות בסגנון זה באופן יזום, ולהשתמש בהן בעתיד לאימון מחודש. הדבר שיפר את איכות ה־Dataset ואת ביצועי המודל בחיזוי בזמן אמת.

טעינת הdataset לאימון:

**בעיה** - כל דוגמה ב־Dataset מורכבת מרצף של סימוני יד (landmarks) לאורך מספר פריימים, כאשר אורך הרצף משתנה מדוגמה לדוגמה. מכיוון שמודלים ללמידה עמוקה מצפים לקבל קלטים באורך אחיד בתוך כל מקבץ (batch), לא ניתן להזין ישירות את כל הדוגמאות באורכים שונים לתהליך האימון.

**פתרון / תוצאות** - יצרנו מחלקה ייעודית שמטפלת בטעינת הקבצים מה־Dataset ובחיבורם לתהליך האימון. המחלקה משתמשת בטכניקת ריפוד (padding), שבה כל רצף מוזן באפסים עד לאורך הרצף הארוך ביותר במקבץ, וכך כל הדוגמאות מתאחדות לצורת טנזור אחיד.  
 פונקציית הריפוד מותאמת אוטומטית כחלק מתהליך יצירת ה־DataLoader, ומאפשרת להזין את כל הדוגמאות למודל בצורה יעילה, ללא אובדן מידע, תוך שמירה על מבנה הרצפים והאורך היחסי שלהם.