The Andrew and Erna Viterbi Faculty of ELECTRICAL & COMPUTER ENGINEERING

SIPL SIGNAL AND IMAGE PROCESSING LAB

TECHNION Israel Institute of Technology

דו"ח סיכום פרויקט מיוחד:

הערכת פארקים באמצעות ראייה ממוחשבת

Characterizing People's Activities and Group Sizes in Videos of Public Open Spaces

מבצעים:

שהם גרינבלט **Shoham Grunblat**

ליעד מרדכי **Liad Mordechai**

מנחה: **אורי בריט** **Ori Bryt**

סמסטר רישום: אביב תשפ"ד

תאריך הגשה: מרץ, 2025

P 7843-2-24

תודות

ראשית נרצה להודות לאורי בריט על ההנחיה הצמודה והמועילה והתמיכה המקצועית לאורך הפרויקט ולשיר גרביץ מהפקולטה לארכיטקטורה ובינוי ערים על שיתוף הפעולה הפורה והעזרה הרבה.
כמובן נודה גם למעבדה על הקצאת המשאבים ועל המשוב המעולה והמלמד בכל שלבי הפרויקט, ובפרט לנמרוד פלג ויאיר משה על הייעוץ והקניית מעטפת מאפשרת וגמישה לעבודה.

תוכן עניינים

1 רקע 6

2 סקר ספרות 7

3 תמצית פרויקט הגדרת גודל קבוצה..... 9

3.1 שיטת העבודה..... 9

3.2 פרטי האלגוריתם 10

4 תוצאות ומסקנות..... 13

5 סיכום והמשך העבודה..... 15

6 רשימת מקורות 22

7 רקע 16

8 תמצית פרויקט זיהוי פעולות 19

רשימת איורים

איור 1 :הגדרת מרחקים ע״פ Hall 7
איור 2: אלגוריתם גילוי וסיווג הקבוצות 11
איור 3: תוצאת האשכול לקבוצות 12
איור 4: אלגוריתם הסינון ומעקב היציבות..... 13
איור 5: מסגרת בוחן- מה עושה הילדה בורוד? 16
איור 6: ההבדל בין תיוג וידאו לתיוג תמונות..... 16

תקציר

מרחבים ציבוריים, ובפרט פארקים, הם חלק משמעותי וחשוב באורח החיים העירוני. בשל גידול האוכלוסיה והתרחבות הבנייה העירונית, במיוחד עקב מגפת הקורונה, תכנון ושימוש אופטימלי במרחב הציבורי הזמין הם קריטיים עבור שיפור איכות החיים.

ההתפתחויות הטכנולוגיות בתחומי הבינה המלאכותית והראייה הממוחשבת מספקות הזדמנות לייעל ולשפר את תהליך התכנון החכם של המרחב על ידי ניתוח מעמיק של סטטיסטיקות כגון אופן, תדירות, סוג ומשך השימוש במרחב הציבורי כמו גם התפלגות אוכלוסיה ומאפיינים נוספים.

הפרויקט שלנו הוא חלק מסדרת פרויקטים שמטרתם לפתח כלים לצורך התכנון החכם על ידי שימוש בסרטונים ממצלמות אבטחה של פארקים. בחלק א׳ של הפרויקט, התמקדנו בסיווג גודל ויציבות של קבוצה לאורך זמן מתוך קטעי וידיאו באמצעות אלגוריתם המשתמש ברשת הראייה הממוחשבת המאומנת מראש YOLOv8 לזיהוי עצמים ובאשכול מבוסס צפיפות עם DBSCAN. חלק ב׳ של הפרויקט מתמקד בזיהוי וסיווג של סוגי פעילויות מתוך קטעי הוידאו באמצעות רשת נוירונים מתאימה.

בהינתן הנתונים המוגבלים והקשים לניתוח שהיו לרשותנו, עליהם נפרט בהמשך, הצלחנו להגיע לתוצאות טובות ממה שציפינו. כפי שנראה, האלגוריתם שבנינו מצליח לאשכל בצורה טובה קבוצות של אנשים במרחב ולאפיין את מיקומן במרחב, גודלן, משך הזמן שבו הן קיימות ומאפיינים שימושיים נוספים, תוך הבחנה והפרדה בין קבוצות סמוכות וסינון רעשים כגון קבוצות לא יציבות ומעברים ארעיים של אנשים אלו על פני אלו.
בחלק ב׳ של הפרויקט, לאחר תיוג אינטנסיבי וממושך ובניית רשת כללית, תוכנת התיוג של הפקולטה הושבתה וכל המידע המתויג שלנו אבד. אי לכך, טרם הצלחנו לממש את מטרת הפרויקט אך נוכחנו בהיתכנות הרעיון ובאופן הפעולה הנדרש.

Abstract

Public spaces, particularly parks, are a vital and significant component of urban life. Due to population growth and urban development, especially in the wake of the COVID-19 pandemic, optimal planning and utilization of available public spaces have become critical for improving quality of life.

Technological advancements in artificial intelligence and computer vision present an opportunity to optimize and enhance intelligent space planning through in-depth analysis of statistics such as the manner, frequency, type, and duration of public space usage, as well as population distribution and other characteristics.

Our project is part of a series of works aimed at developing tools for intelligent planning through the analysis of security camera footage from parks. In part A, we focused on classifying group size and stability over time from video segments using an algorithm that combines the pre-trained YOLOv8 computer vision network for object detection with density-based clustering using DBSCAN. Our follow-up project focuses on identifying and classifying types of activities with a Neural Network.

Given the limited and challenging data at our disposal, which we will detail later, we achieved results that exceeded our expectations. As we will demonstrate, our algorithm successfully clusters groups of people in space and characterizes their spatial location, size, duration of existence, and other useful attributes, while distinguishing between adjacent groups and filtering noise such as unstable groups and temporary overlapping of people.

In Part B of the project, after intensive and prolonged tagging and the construction of a general network as proof-of-concept, the faculty's tagging software was shut down, and all our tagged data was lost. Consequently, we have not yet been able to achieve the project's goal, but we have confirmed the feasibility of the idea and the required mode of operation.

1 רקע – זיהוי גודל קבוצה

גילוי קבוצות מתבצע באופן נאיבי ע"י אלגוריתם k-means [1] , אשר בהינתן מספר מוגדר של קבוצות, מוצא אותן בתמונה. לחלופין, כאשר לא מוגדר מספר הקבוצות, ניתן להשתמש באלגוריתם KNN על מנת לגלות קבוצות בגודל רצוי.

האלגוריתם איננו חף מבעיות: הוא לא עמיד בפני רעש, מצריך ניחוש מקדים של מספר הקבוצות, תוצאתו תלויה באתחול, וזמן ריצתו ארוך. על מנת להתמודד עם בעיות אלו, הצענו להשתמש באלגוריתם אשכול מבוסס צפיפות בנוכחות רעש- DBSCAN.

לאחר זיהוי וסיווג אנשים במסגרת באמצעות YOLOv8, נעשה שימוש ב-DBSCAN על מנת לאשכל את האנשים המזוהים לקבוצות ולאחר מכן אלגוריתם מותאם שבנינו מסנן, מסווג ועוקב אחר הקבוצות והדינמיקה שלהן לאורך זמן.

2 סקר ספרות – זיהוי גודל קבוצה

בעיית גילוי העצמים מתוך תמונות בודדות כמו גם מתוך סרטוני וידיאו היא בעיה פתורה שהביצועים שלה הולכים ומשתפרים כל העת ולכן לא נתמקד בה, ונתייחס אליה כקופסה שחורה. הכלי הנפוץ ביותר בתעשייה הוא הרשת המאומנת מראש YOLO [2] אשר מגלה עצמים ומוציאה אותם כפלט המורכב מקואורדינטות תחימה וסיווג רך לקטגוריות עצמים באמצעות הסתברויות התאמה לכל קטגוריה.

בעבודה שלנו נעשה שימוש בגרסה 8 של האלגוריתם, שהייתה העדכנית ביותר בזמן ביצוע הפרויקט, אם כי בעת כתיבת שורות אלו גרסה 10 כבר זמינה לשימוש - מה שמצביע על ההתקדמות המהירה והמתמדת בטכנולוגיה בתחום והאפשרויות שהיא מציעה.

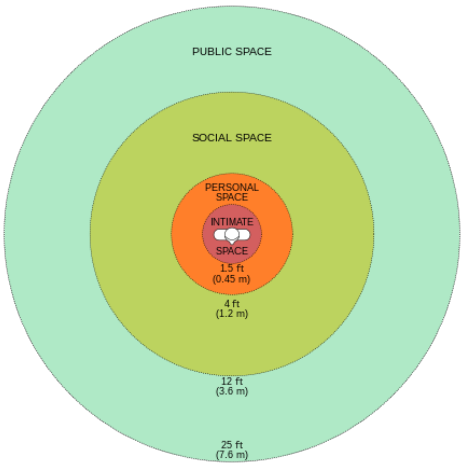
הבעיה המרכזית באלגוריתם שלנו היא בעיית האשכול.

מטבע הפרויקט, הבעיה הראשונה עמה היה עלינו להתמודד הייתה הגדרות יבשות של מהי קבוצה.

לשם כך היה עלינו לשאול שאלות רבות, שלחלקן הייתה תשובה חד משמעית וחלקן שאלות הנתונות לפרשנות ונידונות בהרחבה בספרות המקצועית בתחום בינוי הערים.

על מנת להגדיר היטב את הבעיה, ההנחות וההגדרות שהגדרנו הן כלהלן:

- גודל קבוצה מתחלק לארבע קטגוריות: אדם יחיד, שני אנשים, שלושה עד שישה אנשים, ושבעה אנשים ומעלה.
- קבוצה יציבה היא קבוצה הקיימת באותו מקטע של המסגרת לאורך פרק זמן מינימלי של עשר שניות ומזוהה בלפחות ארבע דגימות מסגרת במהלך פרק הזמן ובתנאי שקטגוריית הגודל שלה לא משתנה בפרק הזמן.
- מרחקים בין אנשים מתחלקים לארבעה סוגי מרחקים, כמפורט באיור 1: מרחק אינטימי, מרחק אישי, מרחק חברתי ומרחק ציבורי [3].



איור 1 :הגדרת מרחקים ע"פ *Hall*

לאחר שהגדרנו מהי קבוצה, ניגשנו לסקור את נושא האשכול עצמו.

מסקירת העבודות הקודמות בנושא, מצאנו גישות שונות שכל אחת מהן מציעה יתרונות וחסרונות משלה ומתאימה לנסיבות הייחודיות של המקרה בו דנה העבודה.

כך למשל, בעבודתם של Chang ושות׳ [4] העוסקת בזיהוי דינמיקה של קבוצות ממצלמות אבטחה של בתי סוהר, הוצעה גישה לאשכול באמצעות עץ פורש מינימלי(MST), אך גישה זו לא מתמודדת היטב עם זוויות הצילום התת-אופטימליות של הדאטה שלנו המעוותות את מטריקת המרחק ויחסי השכנות, ולא משליכה לגיאומטריה תלת מימדית.

בעבודתם של Jens ושות׳ [5], העוסקת בזיהוי גודל קבוצה מתוך תמונות סטריאו, הוצע אלגוריתם ייחודי אשר מתבסס על קידוד ידני של איזורים טיפוסיים במרחב כדי להתמודד עם בעיית האוריינטציה של אנשים קרובים, אך הקידוד הידני הופך את הפתרון למאוד ספציפי ומקשה על הרחבתו לבעיה שלנו, ובנוסף האלגוריתם לא מתייחס לניטור הקבוצות לאורך מימד הזמן אלא לכל מסגרת בנפרד.

3 תמצית פרויקט הגדרת גודל קבוצה

3.1 שיטת העבודה

הקלט שלנו כולל צילומי וידאו ממצלמות אבטחה של שני פארקים ציבוריים באור יהודה לאורך שלושה ימים שונים. מתוך הקלט מחולצת סדרת מסגרות אותה האלגוריתם מנתח.
מודל YOLOv8, רשת עצבית מתקדמת לזיהוי עצמים, משמש לזיהוי ומיקום אנשים בכל פריים. זהו מודל מאומן מראש על מאגר הנתונים COCO, המאפשר לו לזהות מגוון רחב של עצמים. עם זאת, עבור פרויקט זה נלקחים בחשבון רק זיהויים התואמים למחלקת "אדם". המודל פולט תיבות תוחמות, ציוני אמינות ותוויות מחלקה לכל עצם שזוהה.

לאחר זיהוי האנשים במסגרת, התיבות התוחמות מועברות לאלגוריתם DBSCAN[6] לצורך אשכול. כפי שצוין, אלגוריתם DCSCAN נבחר בשל יכולתו למצוא אשכולות בעלי צורה שרירותית ולסנן רעש, מה שהופך אותו למתאים לתרחישים בעולם האמיתי עם מבנים קבוצתיים מגוונים ודינמיים. האלגוריתם מקבץ עצמים על סמך קרבה מרחבית יחסית של התיבות התוחמות שלהם, ובכך מזהה קבוצות.

עבור כל קבוצה שזוהתה מחושבים מספר מאפיינים:
<u>מרכז הקבוצה</u> :המרכז הגיאומטרי של תיבות הגבול בתוך הקבוצה.
<u>רדיוס הקבוצה</u> :המרחק המרבי ממרכז הקבוצה לכל חבר, המייצג את התפשטות המרחבית של הקבוצה.
<u>גודל הקבוצה</u> :מספר האנשים בתוך הקבוצה, המשמש לסיווג הקבוצה.

האלגוריתם גם עוקב אחר שינויים בהרכבי הקבוצות לאורך זמן. על ידי שיוך קבוצות מזוהות בין מסגרות, המערכת יכולה לנתח את הדינמיקה של התנהגות קבוצתית, כגון איחוד, פיצול או דפוסי תנועה.

3.2 פרטי האלגוריתם

שפת התכנות בה השתמשנו למימוש היא פייתון, וזאת בשל הגמישות שלה והתמיכה הרחבה בעיבוד תמונה ולמידה עמוקה.

היישום משתמש בספריות פייתון ייעודיות, כולל OpenCV לעיבוד וקריאת וידאו, PyTorch לטיפול במודלים של למידה עמוקה ו-SciKit-Learn לאשכול.

<u>עיבוד וידאו</u>

תחילה, הסרטון מומר למסגרות באמצעות OpenCV. האלגוריתם דוגם מסגרת אחת בכל שנייה, וכך מבצע דילול של הנתונים המקוריים כדי להפחית את העומס החישובי תוך שמירה על רזולוציה מספקת לזיהוי שינויים בקבוצות.

כל מסגרת עוברת דרך YOLOv8, המזהה את מיקומי האנשים ומפיק את התיבות התוחמות שלהם.

<u>זיהוי עצמים</u>

פלט הזיהוי כולל את התיבות התוחמות של האנשים, ציוני אמינות (confidence scores) ותוויות המחלקות. הפלטים מעובדים כדי להתמקד באנשים בלבד, והתוצאות נשמרות להמשך עיבוד.

יתרונותיו העיקריים של YOLOv8:

- מהירות גבוהה: מסוגל לעבד תמונות בזמן אמת, מה שהופך אותו אידיאלי לניתוח וידאו.
- דיוק משופר: מציע איזון מצוין בין מהירות לדיוק.
- גמישות: ניתן להתאים את המודל למשימות שונות כגון זיהוי עצמים, הקטעה וסיווג תמונות.

<u>אשכול עם DBSCAN</u>

השלב הבא כולל אשכולות של אנשים לתוך קבוצות באמצעות DBSCAN, כמתואר באיור 2 **Error! Reference source not found.**.

לאחר הפעלת מודל YOLOv8 על כל מסגרת, מתקבלות קואורדינטות של התיבות התוחמות. DBSCAN משמש לאשכול קואורדינטות אלו לקבוצות על בסיס פרמטרים מוגדרים של קרבה מרחבית. האלגוריתם מצליח לזהות קבוצות בצורות מגוונות, תוך התחשבות באי סדרים במבנה האנושי ובנוכחות של רעש.

תהליך האשכול:

עבור כל אדם שזוהה DBSCAN, בוחן את הסביבה שלו ברדיוס ϵ.

אם מספר מינימלי של נקודות (MinPts) נמצא בתוך רדיוס זה, נוצר אשכול חדש.

האלגוריתם ממשיך להרחיב את האשכול כל עוד ניתן להוסיף נקודות נוספות בהתאם לקריטריונים של צפיפות.

פרמטרים חשובים:

- ϵ: מרחק השכנות המקסימלי. זהו היפרפרמטר שקבענו לכל מצלמה ופארק בנפרד על פי ניסוי וטעייה על סדרת אימון.

- MinPts: מספר הנקודות המינימלי הנדרש ליצירת אשכול. בפרויקט זה נקבע ל-1 כדי לזהות גם אנשים בודדים בתור קבוצה.

Algorithm 1 Object Detection and Video Processing Algorithm

```
function DetectObjects(image)
    Load YOLOv8 model
    Run detection on image
    filter: person class only
    Extract bounding boxes (x1, y1, x2, y2)
end function
function ClusterObjects(boxes, epsilon)
    Apply DBSCAN clustering with eps = ϵ
    for each cluster do
        Compute center (center_x, center_y) and radius r
        Determine group size
    end for
end function
function ProcessVideo(park, cam, epsilon)
    Save frames from video in 25-sample distances
    for each frame f in frames do
        Run DetectObjects(f)
        Cluster objects using ClusterObjects
        Detect time using DetectTime
        Map objects to segments using DetectMapping
        Save results to CSV
    end for
end function
```

איור 2: אלגוריתם גילוי וסיווג הקבוצות

באיור 3 ניתן לראות את תוצאת האשכול במסגרת לדוגמה.

איור 3: תוצאת האשכול לקבוצות

<u>ניתוח קבוצות</u>

לאחר זיהוי הקבוצות, האלגוריתם מחשב עבור כל קבוצה את המרכז הגיאומטרי, הרדיוס והגודל (מספר האנשים). חישובים אלו מאפשרים סיווג ראשוני של הקבוצות לפי גודל.

האלגוריתם מבצע עיבוד נוסף כדי לשפר את איכות הזיהוי ולסנן תוצאות שגויות:

- מעקב אחר קבוצות: קבוצות מזוהות לאורך מספר מסגרות רצופות כדי להבטיח יציבות ולמנוע זיהויים מקריים.
- סיווג גודל: הקבוצות מסווגות לקטגוריות גודל (יחיד, זוג, 3-6 אנשים, 7+ אנשים) על סמך מספר האנשים בכל אשכול.
- סינון לפי משך זמן: רק קבוצות שזוהו לאורך זמן העובר את הסף התחתון שהגדרנו כ-10 שניות נחשבות כתקפות.

<u>מעקב וניטור התנהגות</u>

באמצעות מעקב אחרי הקבוצות לאורך זמן, ניתן לנתח דינמיקות התנהגותיות כמו התקרבות והתרחקות של אנשים בקבוצה, איחוד או פיצול של קבוצות. מעקב זה מתבצע על ידי מעקב אחרי התיבות התוחמות בין מסגרות וקישור קבוצות בין זמני המדידה.

בנוסף, קיימת אפשרות להגדיר כללים לזיהוי פעילויות ספציפיות או התנהגויות חריגות, כגון שינויים פתאומיים בגודל הקבוצה או תנועות לא רגילות. אלגוריתם זה מוצג באיור 4.

Algorithm 2 Stable Group CSV Parsing Algorithm

```
Define class Group for storing group properties
Initialize active_groups and frame_groups
for each frame in sorted frame_groups do
    Update existing groups' frame activity
    for each current_group in frame do
        if current_group matches existing group (by size, segment, and distance) then
            Update group properties (center, radius, etc.)
            Reset frames_active for matched group
        else
            Create new Group and add to active_groups
        end if
    end for
    for each group in active_groups do
        if group inactive for > window frames then
            Deactivate group
        end if
    end for
end for
Write stable groups with at least 4 detections and >10s duration to output_file
```

איור 4: אלגוריתם הסינון ומעקב היציבות

4 תוצאות ומסקנות – זיהוי גודל קבוצה

בשלב התוצאות נתקלנו בבעיה חדשה הנובעת מסוג הנתונים איתם עבדנו: הנתונים דלילים וישנם קטעי זמן מתים רבים בהם יש מיעוט גילויים או שאין גילויים כלל. ניתן לייחס זאת לעובדה שקטעי הוידיאו נאספו לאורך כל היום בסביבה בה הפעילות מרוכזת באופן טיפוסי סביב שעות מסוימות במהלך היום.

לכך נוספת העובדה שאיכות הנתונים הקשתה על אלגוריתם הזיהוי לגלות עצמים מלכתחילה, מה שהקטין משמעותית את מאגר הנתונים השימושיים שלנו שניתן לאמת והפך אחוזים רבים מאורך הסרטונים לריקים מבחינתנו.

בסך הכל הוערכו 178 קבוצות, ונמצאו 17 טעויות, מה שהוביל לשיעור דיוק של 89.5%. רוב הטעויות התרחשו כאשר אנשים התקרבו זה לזה ואז התרחקו בפרק זמן קצר, מה שיצר תרחישים מאתגרים שבהם אנשים היו קטנים מדי או רחוקים מדי מהמצלמה, דבר שהגביל את יכולתו של האלגוריתם YOLO לזהות אותם בצורה מדויקת.

לאור מיעוט העבודות הקודמות בנושא המשתמשות בסרטוני וידיאו בתור נתונים, השוואת תוצאות תהיה חסרת משמעות, אך גם בהשוואה לתוצאות שהושגו על נתונים של תמונות בדידות זהו דיוק תחרותי.

מעבר לשימוש הטבעי שלשמו בוצע הפרויקט– תכנון חכם של מרחבים ציבוריים- שימושים נוספים יכולים להיות בתחום הבטחוני לצורך ניתוח וניטור בזמן אמת של מצלמות אבטחה בבתי כלא או איזורים נפיצים, בתחום מניעת

האלימות בבתי ספר ומרחבים ציבוריים, ואפילו בתחום הספורט הקבוצתי ככלי לניתוח אסטרטגיות משחק ופיזור שחקנים על המגרש במהלך המשחק.

שיפורים משמעותיים שיכולים להתבצע הם שימוש בגרסאות עדכניות ומשופרות של אלגוריתם הזיהוי YOLO, לו נוספות יכולות דיוק וניתוח מתקדמות יותר בכל גרסה, שימוש בבינה מלאכותית על מנת לבצע חיזוי וניתוח נתונים מעמיק בהסתמך על מידע נוסף על האיזור והזמן, למשל באמצעות הוצאת מאפיינים ע"י טרנספורמר, שיפור המיקום והזווית של המצלמות על מנת לקבל נתונים נוחים ואיכותיים יותר, איסוף נרחב יותר של נתונים ממצלמות במקומות אחרים ובסוגים סביבה שונים על מנת ליצור יכולת הכללה לאלגוריתם וכיוונים רבים נוספים.

הצלחת הפרויקט ככלי ראשוני מצביעה על כדאיות בפיתוח והעמקת יכולות הכלי גם הלאה.

5 סיכום והמשך העבודה – זיהוי גודל קבוצה

פרויקט זה בא להציג אלגוריתם למציאת גודל קבוצה מתוך סרטוני וידיאו, שנבנה כחלק מסדרת כלים המפותחים במעבדה לצורך ניתוח ותכנון חכם של מרחבים ציבוריים.

הבעיות העיקריות שעמדו בפנינו היו הגיאומטריה של המצלמות, איכות הנתונים והדינמיות וחוסר הודאות של מיקומי הקבוצות, גדליהן ומספרן בכל מסגרת.

פתרון הבעיה הורכב מכמה שלבים:

ראשית, על מנת להתגבר על איכות הנתונים, השתמשנו בגרסה המתקדמת ביותר של אלגוריתם זיהוי העצמים YOLO שהייתה זמינה בשוק בעת תחילת העבודה.

לאחר עיבוד מקדים וסינון ראשוני, העברנו את הזיהויים שיצאו כפלט מ-YOLO לאלגוריתם האשכול שלנו.

השתמשנו באלגוריתם המבוסס על DBSCAN- אשכול מבוסס צפיפות בהינתן רעש- אשר הקנה לנו גמישות בכל הנוגע לקבוצות. את הפלט של אלגוריתם האשכול הזנו לאלגוריתם הסינון והמעקב שלנו כדי לזהות את הקבוצות לאורך זמן ולנתח את הדינמיקה והיציבות שלהן.

התוצאות הטובות אליהן הגענו מראות שגלום פוטנציאל בפיתוח כלים מסוג זה ושילוב אלגוריתמיקה מעולמות הבינה המלאכותית וניתוח הנתונים בתכנון אדריכלי וסביבתי חכם, כמו גם בתחומים רבים אחרים בחיי היום יום.

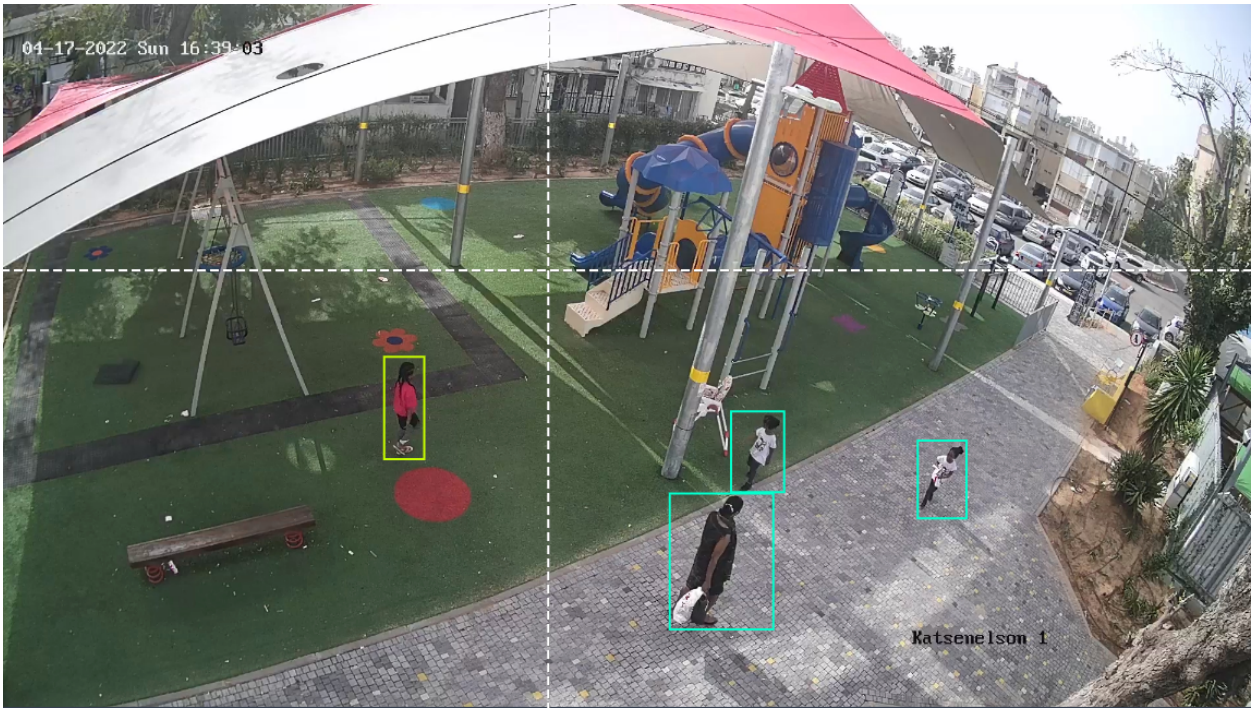
עם זאת, ראינו שישנו הרבה מקום לשיפור והצענו דרכים רבות לעשות כך.

בפרויקט ההמשך, אנו עובדים על פיתוח כלי משלים לכלי שפותח בפרויקט זה, אשר ישמש לסיווג של פעולות מתוך אותם הסרטונים בהם השתמשנו. כלי זה משתמש ברשת נוירונים המאומנת על הסרטונים הללו בתוספת תגיות המסווגות את העצמים המזוהים ע״י YOLO לאחד משמונה סוגי פעילות טיפוסיים, כגון ריצה, הליכה או משחק, וזאת על מנת לנתח ולהבין את אופן השימוש של האוכלוסיה בפארק הציבורי ולאפשר תכנון מסתגל ונכון יותר לאור זאת.

6 רקע – זיהוי פעילויות

כאשר אנו ניגשים למשימת זיהוי הפעילויות מתוך וידאו, הבעיה הראשונה שעומדת בפנינו היא בעיית המשכיות התנועה.

איור 5באיור 5 נציג כמקרה בוחן את המסגרת הבאה, הלקוחה מתוך הסרטונים איתם עבדנו:



איור 5: מסגרת בוחן- מה עושה הילדה בורוד?

על פניו, נראה שהילדה עומדת. רגליה צמודות וגופה במנח סטטי. ואכן, אילו מדובר היה בזיהוי פעולות מתוך תמונות בודדות, הרשת הבסיסית הייתה מתייגת את פעולת הילדה בתור עמידה.

רשת מתוחכמת יותר הייתה לוקחת בחשבון מספר מסגרות קדימה ומספר מסגרות אחורה, והייתה מסיקה שהילדה דווקא אינה עומדת אלא הולכת.

אך האם זה מספיק?

פרט מידע שיתגלה לנו בצפייה בקטע הוידאו יספר לנו שלא, שכן הילדה בכלל רודפת אחרי כדור שבעטה קדימה לפני כמה עשרות מסגרות, ובעוד כמה מסגרות היא תתאחד איתו ותמשיך לשחק בו, כמתואר באיור 6.

איור 6: ההבדל בין תיוג וידאו לתיוג תמונות

סוגיה נוספת שעולה כאשר הנתונים הם סרטוני וידאו היא אנשים אשר מבצעים מספר פעולות בזו אחר זו, ובפרט פעולות המכילות זו את זו.

נתבונן למשל באיור 7, בו אנו רואים אב ובנו אשר רצים על הדשא הסינתטי.

מבחינת דפוס פעולתם, אין עוררין על כך ששניהם רצים, ואכן מפתה לתייג אותם בתור רצים. גם התבוננות בוידאו לא תסתור את העובדה הזו.

אבל לשם מה נועדה הריצה? הרי לפני שהחלו לרוץ, שיחקו ביחד באיזור המגלשות, וריצתם קצרה ומתבצעת מהמגלשה אל הנדנדה, שם ימשיכו לשחק.

האם לפיכך פעולת הריצה היא אכן ריצה, או שהיא חלק מתהליך המשחק שלהם?

איור 7: מה עושים האב והילד?

התשובה לסוגיה זו נגזרת ממטרת העל של הפרויקט- ניתוח דפוסי השימוש בפארק. כאשר אנו מתעדים פעולה של פרט בפארק, אנו צריכים לקחת בחשבון את מטרת הפעולה והאופי שלה, ולכן ריצה בין מתקנים תיחשב חלק מפעולת המשחק, בעוד שהקטגוריה "ריצה" תתייחס רק לאנשים שהגיעו לפארק על מנת לרוץ ומבצעים ריצה כפעילות גופנית.

באותו אופן, באיור 8 נוכל לראות ילדה יושבת על הנדנדה ושתי ילדות יושבות על מדרגות המגלשה.

רשת נאיבית הייתה מתייגת את פעולת הילדה שעל הנדנדה בתור משחק, אך מהתבוננות בוידאו לאורך זמן אנו יכולים לראות שהילדה אינה מתנדנדת אלא יושבת מבלי לזוז, ומהות הפעולה שהיא מבצעת כרגע היא ישיבה.

איור 8: האם הילדות יושבות או משחקות?

הגדרות אלו והדקויות שהן מכניסות לתהליך הן חלק מהותי מהתיוג והאימון של הרשת וחשיבותן להצלחת המודל גדולה, לכן השקענו זמן רב בתכנון ובחידוד ההגדרות תוך כדי התיוג.

7 סקר ספרות – זיהוי פעולות

זיהוי פעולות מבוסס וידאו נועד לזהות ולסווג באופן אוטומטי פעולות אנושיות ברצפי וידאו. משימה זו כוללת יישומים רבים, כולל מעקב, אינטראקציה בין אדם למחשב, ניתוח ספורט, ניתוח התנהגותי ונהיגה אוטונומית. המורכבות של נתוני וידאו, הכוללים מידע מרחבי וזמני, מחייבת אלגוריתמים מתוחכמים לזיהוי פעולות יעיל. למידה עמוקה התגלתה ככלי רב עוצמה להתמודדות עם אתגרים אלה, המאפשרת פיתוח מודלים חזקים ומדויקים.

גישות למידה עמוקה מוקדמות הסתמכו בעיקר על רשתות קונבולוציה(CNN) דו-ממדיות, שתוכננו במקור לזיהוי תמונות. שיטות אלה חילצו לעתים קרובות מאפיינים מרחביים ממסגרות בודדות ואז צירפו אותם באמצעות רשתות RNN או טכניקות pooling פשוטות. לדוגמה, סימוניאן וזיסרמן [7] הציעו ארכיטקטורת CNN דו-ערוצית, כאשר ערוץ אחד מעבד מסגרות RGB והשני מעבד זרימה אופטית, תוך לכידת מידע על התנועה ברצף המסגרות. עם זאת, שיטות אלה התקשו ללכוד ביעילות תלות זמנית לאורך רצפי וידאו ארוכים.

כדי למדל מידע זמני בצורה טובה יותר, הוצע שימוש ברשתות קונבולוציה תלת-ממדיות. רשתות אלה מרחיבות קונבולוציות דו-ממדיות לממד הזמני, ומאפשרות חילוץ סימולטני של מאפיינים מרחביים וזמניים. טראן ושות׳ [8] הציעו את C3D, ארכיטקטורת CNN תלת-ממדית חדישה שהציגה ביצועים מעולים על מדדי זיהוי פעולות.
בהמשך הוצגה רשת I3D (רשת קונבולוציה תלת מימדית מנופחת) [9] , אשר שיפרה עוד יותר את C3D על ידי ניפוח ארכיטקטורות CNN דו-ממדיות שאומנו מראש על ImageNet לרשתות תלת-ממדיות, תוך מינוף משקלים שאומנו מראש לביצועים משופרים. עם זאת, רשתות קונבולוציה תלת-ממדיות הן בעלות סיבוכיות חישובית גבוהה מאוד ודורשות זיכרון רב, מה שמגביל את יכולת ההרחבה שלהן לסרטונים ארוכים.

כדי לאזן בין דיוק ליעילות, פותחו גישות דו-ערוציות והיברידיות. שיטות אלה משלבות רשתות קונבולוציה דו-ממדיות ותלת-ממדיות, תוך מינוף חוזקותיהן של שתי הארכיטקטורות. פייכטנהופר ושות׳ [10] הציעו רשת ResNet, ששילבה זרמי RGB וזרימה אופטית. רשתות SlowFast הציגו ארכיטקטורה דו-נתיבית[11] , כאשר הנתיב האיטי לוכד סמנטיקה מרחבית בתדירות מסגרות נמוכה, והנתיב המהיר לוכד דינמיקה זמנית בתדירות מסגרות גבוהה. שיטות אלה השיגו ביצועים תחרותיים על מדדי זיהוי פעולות שונים.

לאחרונה, מודלים מבוססי טרנספורמר הותאמו לזיהוי פעולות וידאו. מודלים אלה משתמשים במנגנוני תשומת לב עצמית כדי ללכוד תלות ארוכת טווח ברצפי וידאו.
ארכיטקטורת ViViT (Video Vision Transformer) [12]מרחיבה את ארכיטקטורת Vision Transformer (ViT) לווידאו, תוך התייחסות למסגרות וידאו כרצף של תיקונים.
ארכיטקטורת TimeSformer [13] בוחנת עוד אסטרטגיות שונות ליישום תשומת לב עצמית לנתונים מרחביים-זמניים. מודלים מבוססי טרנספורמר אלה הראו תוצאות מבטיחות, במיוחד בלכידת תלות זמנית ארוכת טווח.

בחירת ייצוג הנתונים ואסטרטגיית הלמידה משפיעה באופן משמעותי על ביצועי מודלים של זיהוי פעולות. זרימה אופטית ומסגרות RGB הם ייצוגים נפוצים.

על אף ההתקדמות המשמעותית, זיהוי פעולות וידאו עדיין מתמודד עם מספר אתגרים. אלה כוללים:
תלות ארוכת טווח: לכידת תלות זמנית ארוכת טווח נותרה אתגר, במיוחד בסרטונים עם פעולות מורכבות ומשכי זמן ארוכים.
סיבוכיות חישובית: רשתות קונבולוציה תלת-ממדיות ומודלים מבוססי טרנספורמר הם יקרים מבחינה חישובית, מה שמגביל את יכולת השימוש בהם על מידע רב בזמן קצר.
נתונים מתויגים מוגבלים: רכישת מערכי נתונים גדולים של סרטונים מתויגים היא יקרה וגוזלת זמן. אצלנו הבעיה הזו גדולה אף יותר, מכיוון שהמידע שברשותנו מצומצם ונאסף בתנאים מאוד מסוימים. לכן יכולת ההכללה של המודל עומדת בסימן שאלה.

על מנת להתמודד עם בעיות אלו, בפרויקט שלנו בחרנו להתמקד בפעולות קצרות יחסית הנמשכות שניות עד דקות בודדות. רעיון שעלה כדי להתמודד עם בעיית הנתונים היה להשתמש במודל שאומן מראש על מאגרים גדולים ונפוצים, אך פתרון זה לא מתאים לבעיה שלנו מכיוון שבמאגרי זיהוי פעולות כלליים, הנתונים הם לרוב סרטונים קצרים בני כחמש שניות המתעדים אדם יחיד מבצע פעולה יחידה וברורה, ישירות מול המצלמה. אצלנו, לעומת זאת, התיעוד הוא ממצלמות אבטחה וכזה הוא כולל דמויות רבות המבצעות פעולות שונות זו מזו ואשר לעתים קרובות עוברות מפעולה לפעולה תוך כדי תנועה, ובנוסף נמצאות במרחקים וזוויות לא אידיאליים ביחס למצלמה, בהמשך להסבר שהעלינו בחלק א׳ של הפרויקט.

8 תמצית פרויקט הגדרת גודל קבוצה

פרויקט זה בא להציג מודל רשת נוירונים לזיהוי פעולות מתוך וידאו (Video Based Action Recognition).

העבודה על הפרויקט מחולקת לשלושה חלקים:

1. תיוג סוגי פעילויות- העלינו את סרטוני הוידאו ממצלמות האבטחה לתוכנת התיוג CVAT ותייגנו פעילויות של אנשים לאורך הסרטונים. הפעילויות התחלקו לשמונה קטגוריות:
   - ישיבה
   - עמידה
   - הליכה
   - ריצה
   - משחק
   - הליכה עם עגלה
   - רכיבה על כלי דו גלגלי
   - הליכה עם חיית מחמד

במהלך התיוג היה עלינו להפעיל שיקול דעת בהתאם להגדרות שהוחלט עליהן בהתבסס הן על הפן הארכיטקטוני והן על הפן הטכנולוגי.

2. עיבוד מקדים של הנתונים המתויגים ושל סט הבוחן- על מנת להעביר את הנתונים לרשת נוירונים, יש צורך להתאים בין הפלט של ה-cvat, הכולל אובייקטים המורכבים מרצפים של מסגרות ותיוג, לבין תיבות הגבול שמתאימות לפלט מודל yolo. בנוסף יש לבצע סגמנטציה מרחבית ולחלץ זמן מהמסגרת, כפי שעשינו בחלק א׳ של הפרויקט.
3. בניית רשת הנוירונים ואימונה.

בהתבסס על המחקר המקדים וסקר הספרות שערכנו, הכיוון הראשוני שבחרנו לרשת הנוירונים הוא רשת קונבולוציה אשר ידועה כמתאימה למידע חזותי.

9 רשימת מקורות

[1] M. Ahmed, R. Seraj and S. M. S. Islam, "The k-means Algorithm: a Comprehensive Survey and Performance Evaluation," *Electronics,* vol. 9, no. 1295, 2020.

[2] G. Jocher; A. Chaurasia; J. Qiu, "YOLOv8," Ultralytics, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.

[3] E. Hall, The Hidden Dimension, New York: Anchor Books, 1966.

[4] M. Chang, N. Krahnstoever and S. Lim, "Group Level Activity Recognition in Crowded Environments Across Multiple Cameras," in *Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010.

[5] K. Jens, A. Das and M. B. Kjaergaard, "Detecting Group Sizes and Human-centered Interaction Using 3D Stereo Vision Cameras," in *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2020.

[6] H. P. Kriegel, M. Ester, J. Sander and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of 2nd International Conference on Knowledge Discovery and data Mining (KDD-96)*, Munich, Germany, 1996.

[7] K. Simonyan and A. Zisserman, "Two-stream Convolutional Networks for Action Recognition in Videos," in *Advances in Neural Information Processing Systems*, 2014.

[8] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *IEE International Conference on Computer Vision*, 2015.

[9] J. Carreira, H. Agrawal, G. Landsberger, G. Toderici and C. Schmid, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[10] C. Feichtenhofer, A. Pinz and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[11] C. Feichtenhofer, H. Fan, J. Malik and K. He, "Slowfast Networks for Video Recognition," in *Proceedings of the IEEE international conference on computer vision*, 2019.

[12] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic and C. Schmid, "ViViT: A Video Vision Transformer," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, 2021.

[13] G. Bertasius, H. Wang and L. Torresani, "Space-time Transformer," in *roceedings of the IEEE/CVF international conference on computer vision*, Montreal, 2021.