DETECTING GENDER BIAS IN HUMAN ACTIVITY VIDEO DATASETS: A MULTI-COMPONENT VISUAL METRIC APPROACH

Mudalige T.N.

(IT21208294)

BSc (Hons) degree in Information Technology Specializing in Data Science

Department of Information Technology

Sri Lanka Institute of Information Technology

April 2025

DETECTING GENDER BIAS IN HUMAN ACTIVITY VIDEO DATASETS: A MULTI-COMPONENT VISUAL METRIC APPROACH

Mudalige T.N.

(IT21208294)

Dissertation submitted in partial fulfillment of the requirements for the Bachelor of BSc (Hons) degree in Information Technology Specializing in Data Science

Department of Information Technology

Sri Lanka Institute of Information Technology

April 2025

DECLARATION

I declare that this is my own work and this dissertation1 does not incorporate without acknowledgement any material previously submitted for a degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology, the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:  Date: 2025.04.11

The above candidate has conducted research for the bachelor’s degree Dissertation under my supervision.

Signature of the Supervisor: Date:

ABSTRACT

As computer vision systems increasingly influence areas like surveillance, sports analytics, and interactive media, addressing fairness and gender bias in video datasets become critical. This research proposes a framework to detect and quantify gender bias in human activity videos, focusing on activities such as sports and fitness.

A multi-component metric was developed to evaluate bias across five dimensions: **Size Bias**, **Centering Bias**, **Screen Time Bias**, **Embedding Bias**, and **Motion Bias**. These capture differences in visual prominence, positioning, visibility, feature similarity, and movement dynamics across genders. Features were extracted using **YOLOv8**, **MediaPipe**, and the **SlowFast** model.

Bias scores were computed per video and aggregated using three methods, **directional mean**, **magnitude sum**, and a **PCA-weighted score**, then normalized to a consistent range. The results revealed non-uniform gender representation across activity classes, with some showing male-dominant framing and others female-oriented motion styles.

This metric provides an interpretable way to analyze gender representation, aiding dataset audits and promoting fairness in vision-based AI systems.

Keywords: gender bias, video datasets, human activity recognition, fairness metrics, computer vision

ACKNOWLEDGEMENT

I would like to express my sincere appreciation to my supervisor, Dr. Prasanna S. Haddela, and co-supervisor, Ms.Thisara Shayamalee, for their expert guidance, constructive feedback, and consistent support throughout this research. Their contributions were critical in shaping the methodology and maintaining academic rigor during each phase of the project.

I also wish to acknowledge the collaborative input of my research group members, whose discussions and assistance added valuable perspectives to the development of this work.

Furthermore, I am grateful to the academic and administrative staff of the Sri Lanka Institute of Information Technology (SLIIT) for their support and professionalism. I extend my particular thanks to the SLIIT Library for providing timely access to essential research materials and resources.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBRIVIATIONS

Abbreviation	Full Form
AI	Artificial Intelligence
HAR	Human Activity Recognition
PCA	Principal Component Analysis
YOLO	You Only Look Once (Object Detection Model)
CV	Computer Vision
FPS	Frames Per Second
Z-score	Standard Score (Statistical Normalization Method)
GPU	Graphics Processing Unit
CSV	Comma-Separated Values

1. INTRODUCTION

1.1. Background

Computer vision technologies are increasingly embedded in applications such as surveillance, sports analytics, virtual fitness, and human-computer interaction. These systems often rely on large-scale video datasets to train models for human activity recognition (HAR), where individuals are automatically classified based on their actions. However, growing evidence suggests that such datasets may carry embedded gender biases that affect how individuals are represented and interpreted by these models [1], [2].

Gender bias in visual datasets can manifest in several ways, including disproportionate representation, stereotypical associations between gender and activity, and differences in how subjects are visually presented. For instance, variations in bounding box size, camera centering, visibility duration, and movement patterns can all contribute to biased representations. These patterns not only reflect existing societal stereotypes but also risk perpetuating them through automated systems. Research has shown that even advanced models, such as vision-language frameworks and activity classifiers, tend to associate specific activities with particular genders irrespective of the actual content [1].

While previous efforts have addressed bias in image-based vision systems, the complexity of video data introduces additional considerations such as temporal patterns, motion dynamics, and continuous framing. These aspects are especially relevant in activity recognition, where motion and pose are central to the classification task. Studies have identified gender-related artifacts in well-known datasets, indicating that bias can be encoded both in low-level features such as color and lighting, and in higher-level representations like pose and subject positioning [2], [3].

This research aims to systematically measure and quantify gender bias in human activity video datasets using a set of multi-component visual metrics. By analyzing spatial, temporal, and representational features of individuals across activities, this work contributes to ongoing efforts to promote fairness and transparency in computer vision systems.

1.2. Problem Statement

The advancement of human activity recognition (HAR) has been driven by large-scale video datasets and powerful deep learning models. However, a critical yet underexplored issue in this domain is the presence of gender bias embedded in these datasets. This bias may manifest through unequal visual treatment of male and female subjects in terms of bounding box size, central positioning, screen time, motion portrayal, and representation in learned embeddings. When left unaddressed, such biases can lead to skewed model behavior, unfair outcomes, and reinforcement of gender-based stereotypes.

Existing methods for bias detection in computer vision often focus on classification tasks or static imagery. They rarely account for the dynamic, temporal, and spatial complexities of video data, particularly in human activity contexts where movement, pose, and framing evolve across frames. As a result, there is a lack of quantifiable tools that can evaluate gender bias in activity videos using both visual and motion-based features.

This research addresses this gap by developing a composite metric system capable of measuring five distinct forms of visual bias. The goal is to provide a structured, interpretable, and scalable approach for identifying how gender bias is represented and propagated in human activity video datasets. The absence of such tools currently limits efforts to audit dataset fairness and design interventions that promote equity in vision-based systems.

1.3. Research Gap

Gender bias in artificial intelligence has been widely studied in domains such as facial recognition and image classification. However, video datasets used for human activity recognition have not received equivalent scrutiny, despite their growing role in applications ranging from surveillance to health monitoring. These datasets may encode structural biases that affect how gender is represented, leading to imbalances in visibility, framing, or behavioral interpretation [4].

Most prior work on bias mitigation in computer vision has focused on static imagery or individual attributes such as facial features. These approaches rarely capture the spatial and temporal complexities inherent in video data, particularly when human motion and poses are critical for recognition. As a result, important dimensions of gender bias such as bounding box prominence, centering patterns, motion characteristics, and screen presence remain underexplored [4], [2].

Additionally, while several methods exist to assess fairness at the model level, there is a lack of unified metrics to quantify bias at the dataset level, especially across multiple visual dimensions. This creates a gap in both the diagnosis and auditing of bias within the data pipelines that support vision-based activity recognition systems. Without structured tools to identify these issues, biases may propagate unnoticed into real-world applications [2], [3].

This research aims to bridge this gap by developing a multi-component metric system capable of capturing diverse forms of gender bias in human activity video datasets. The proposed framework integrates spatial, motion, and embedding-based features to provide a quantitative, interpretable assessment of bias distribution across gender categories.

1.4. Research Objectives

Main Objective:

To develop a comprehensive metric-based framework that detects and quantifies gender bias in human activity video datasets using multi-dimensional visual and motion-based features.

Sub-Objectives:

- To define and implement five component-level metrics representing distinct bias dimensions: Size Bias, Centering Bias, Screen Time Bias, Embedding Bias, Motion Bias.
- To extract spatial and motion features from video frames using:
    - YOLOv8 for person detection and bounding box features
    - MediaPipe for pose estimation and motion vectors
    - SlowFast model for activity-related embeddings
- To normalize and aggregate metric scores using statistical techniques such as directional averaging, magnitude summation, and PCA-based weighting.
- To assess and compare gender bias across different activity categories within the dataset.
- To build an interpretable evaluation tool that can support dataset audits and fairness analyses.

1.5. Scope and Limitations

Scope:

This research focuses on identifying and quantifying gender bias in video datasets used for human activity recognition. The study is limited to binary gender classification (male and female) as provided by the dataset labels. The analysis includes videos categorized by activity class and evaluates each video based on visual and motion-based characteristics.

The framework covers five distinct components of visual bias: size, centering, screen time, motion, and embedding similarity. Feature extraction techniques include object detection (YOLOv8), pose estimation (MediaPipe), and action recognition embeddings (SlowFast). The final metric outputs enable video-level and dataset-level interpretation of bias in terms of both direction and intensity.

The framework is intended to be dataset-agnostic and can be extended to any structured video dataset containing human activity footage and gender annotations.

Limitations:

- The study is constrained to datasets labeled with binary gender only. It does not account for non-binary or ambiguous gender identities due to a lack of corresponding annotations.
- The metric design assumes that visual prominence (such as bounding box size or screen time) correlates with representational bias. This assumption may not fully capture the nuances of all cultural or contextual interpretations.
- The results are influenced by the performance of underlying models (YOLOv8, MediaPipe, SlowFast). Inaccuracies in detection or pose estimation may introduce noise into the metric calculations.

Despite these limitations, the proposed metric framework offers a scalable and interpretable approach for quantifying gender bias in human activity video datasets and contributes to broader efforts in algorithmic fairness.

1.6. Literature Review

1.6.1. Bias in Machine Learning

Machine learning systems often reflect and amplify the biases embedded in their training data. These biases may arise due to unequal representation, labeling inconsistencies, or cultural assumptions, and they can lead to discriminatory behavior in downstream applications. Bias can manifest across sensitive attributes such as gender, race, or geography and is particularly concerning in systems that influence social outcomes like hiring, healthcare, or surveillance.

Studies have highlighted that even seemingly neutral tasks, such as object detection or activity recognition, can internalize spurious correlations related to protected attributes [4]. This raises the need for frameworks that can identify, measure, and mitigate these biases before models are deployed.

1.6.2. Gender Bias in Video and Vision Datasets

In computer vision, gender bias has been well-documented in image classification and facial recognition systems. However, video datasets, particularly those used for human activity recognition, have received comparatively less attention. Videos introduce temporal and spatial elements that can encode additional forms of bias, such as motion patterns, screen time, and subject framing.

Tools such as REVISE have emerged to support dataset-level bias analysis, helping researchers uncover gender and object-based imbalances in visual data [5]. These imbalances can influence model predictions and exacerbate unfair treatment if not appropriately accounted for.

1.6.3. Existing Metrics and Their Limitations

Fairness and bias mitigation in machine learning have been active areas of research, especially in applications involving sensitive attributes such as gender or race. Several statistical metrics have been introduced to measure bias at the model output level, including Demographic Parity, which requires equal outcome distribution across groups, and Equalized Odds, which ensures equal error rates for distinct groups [6], [7]. Other metrics like Statistical Parity Difference and Bias Amplification aim to quantify the degree to which model decisions or learned representations reflect or intensify pre-existing biases in the data [6].

In the domain of computer vision, tools such as REVISE have been proposed to assess image datasets for bias. REVISE focuses on object co-occurrence, scene diversity, and stereotypical labeling, primarily within static image datasets [5]. Similarly, tools like FairFace and Fairlearn provide diagnostic interfaces to examine representation fairness in facial recognition or classification tasks.

However, these metrics and tools exhibit several limitations when applied to video-based human activity recognition:

- They are typically designed for static images or structured tabular data, lacking the capacity to analyze temporal dynamics, motion features, or spatial positioning across video frames.
- Metrics such as equalized odds or demographic parity are model-centric and focus on output disparities, rather than evaluating representational bias at the dataset level.
- Few tools offer component-level interpretability, making it difficult to isolate and analyze specific contributors to visual bias, such as bounding box prominence or pose behavior.

Furthermore, existing evaluations often do not consider the interplay between spatial (e.g., position, size) and temporal (e.g., motion, visibility duration) attributes, which are crucial in video analysis. There remains a lack of integrated, interpretable, and modular bias metrics tailored for video datasets that encode complex human activities.

This research addresses these shortcomings by introducing a multi-component metric framework that quantifies gender bias in human activity videos through visual and motion-based analysis. It is designed to be interpretable, dataset-level, and extensible, helping bridge the current gap in bias assessment methodologies for video data.

1.6.4. Summary of Insights

The review of existing literature highlights that while machine learning fairness metrics such as statistical parity, equalized odds, and demographic parity are well established for classification tasks, they are often limited to model outputs and do not adequately assess bias within the dataset itself. Tools such as REVISE [5] have emerged to diagnose bias in image datasets, but they lack temporal and motion-related analysis capabilities essential for video content.

Research shows that gender bias in visual datasets can manifest in several ways, including imbalanced representation, framing disparities, and activity-gender stereotypes. However, few studies have extended these observations to video-based human activity recognition, where spatial and temporal elements introduce additional dimensions of representational bias.

Existing bias metrics and fairness tools are image-centric or designed for structured data. They fall short in handling complex motion cues, pose dynamics, and visibility variation that are central to video-based datasets. Moreover, they typically do not provide component-level interpretability, making it difficult to pinpoint the source and nature of bias.

These gaps underscore the need for a multi-dimensional, interpretable, and video-specific framework to assess gender bias in activity datasets. The insights gathered form the foundation for the methodology proposed in this study, which combines spatial, motion, and embedding-based features to generate meaningful and actionable bias scores.

2. METHODOLOGY

2.1. Overview of Approach

This study adopts a structured methodology to quantify gender bias in human activity video datasets through the development of a multi-component visual metric framework. The approach comprises four sequential stages: dataset preparation, feature extraction, bias metric computation, and score aggregation.

Dataset Preparation

The dataset utilized in this research consists of 500 annotated videos, each categorized by activity type and labeled by the perceived gender of the subject (male or female). The dataset spans a range of physical activities and serves as the foundation for metric computation.

Feature Extraction

To capture spatial, temporal, and semantic characteristics from each video, three established computer vision models are employed:

- **YOLOv8** is used to detect persons within frames and extract bounding box dimensions and positional data.
- **MediaPipe Pose** provides pose keypoints for estimating motion vectors and evaluating stylistic dynamics.
- **SlowFast** is used to generate high-level video embeddings that capture semantic action representations.

Bias Metric Computation

Five independent component metrics are defined to capture different forms of visual gender bias:

- **Size Bias**: Average bounding box area, normalized by frame dimensions.
- **Centering Bias**: Mean Euclidean distance from the frame center.
- **Screen Time Bias**: Total frame count during which a subject is visible.

- **Embedding Bias**: Cosine similarity between video embeddings and precomputed gender-specific centroids.
- **Motion Bias**: Comparison of motion vectors to gender-based motion profiles derived from pose data.

Score Aggregation and Normalization

Each bias metric is normalized and aggregated to compute:

- A **directional score** indicating bias orientation (male or female),
- A **magnitude score** representing the overall strength of bias, and
- A **PCA-weighted score** to emphasize components based on data-driven variance.

This methodology facilitates a systematic, interpretable, and replicable assessment of gender bias at both the individual video level and across the entire dataset.

2.2. Dataset Description

This research utilizes the HAA500 (Human-Centric Atomic Action) dataset [8], a curated video dataset designed for fine-grained human activity recognition. HAA500 consists of over 591,000 labeled frames covering 500 distinct atomic action classes. Unlike coarsely annotated action datasets, HAA500 provides granular action labels that distinguish between similar yet context-specific movements.

The dataset is structured around human-centric motion, with an emphasis on minimizing irrelevant frames and maximizing the clarity of movement. Each video captures a specific atomic action performed by a subject, making the dataset highly suitable for spatial and temporal analysis of pose, motion, and framing. HAA500 includes detailed annotations with a high average pose detectability of 69.7 percent, which enhances its applicability for pose estimation and motion bias analysis.

In this study, a gender-labeled subset of five hundred videos was selected from the HAA500 dataset. Each video is annotated with the activity class and perceived gender (male or female), forming the basis for the computation of gender bias metrics. The diversity of actions and framing styles in HAA500 provides a robust testbed for evaluating visual and motion-based biases across activity types.

2.3. Selection of Features for Equation Building

The components selected for measuring gender bias in human activity video datasets are based on their empirical relevance in visual bias literature and their practical measurability using existing vision models. Each metric captures a distinct mode of representation bias, allowing for a comprehensive and interpretable analysis.

2.3.1. Size Bias

Bounding box size is a strong proxy for visual prominence. Prior work has shown that larger object regions draw more visual attention and can reinforce biased associations when subjects of one gender consistently appear more prominently in the frame. This phenomenon has been linked to unequal representation in tasks such as captioning and object detection [9].

2.3.2. Centering Bias

Positioning within the center of the frame increases perceptual salience and is often used to guide viewer attention. Gender-based disparities in centering have been observed in image and video datasets, often reflecting unconscious framing choices. Such center bias can amplify unequal representation of subjects, especially in surveillance or media analytics contexts [2].

2.3.3. Screen Time Bias

Screen time measures visibility duration and has been linked to both narrative dominance and training data influence. If one gender appears more frequently or for longer durations, this imbalance can influence both human viewers and model learning. This is critical in video contexts, where exposure frequency can implicitly convey importance [4].

2.3.4. Embedding Bias

Embedding bias examines how semantically similar a video is to gender-based activity patterns. This reflects not only visible content, but also higher-order associations captured by deep models. Such biases in feature space can lead to skewed model interpretations, as observed in pretrained captioning and vision models [9].

2.3.5. Motion Bias

Motion bias captures gender-based variations in pose dynamics and body movement patterns across video sequences. These differences may reflect behavioral tendencies, camera framing decisions, or performance styles, and can be encoded as stylistic motion features using pose trajectories.

Studies have shown that pose-based representations, such as joint angles and motion trajectories, are effective not only for activity classification but also for predicting gender from movement. For example, Kastaniotis et al. proposed a real-time gait-based gender recognition system using pose estimation from depth images. Their method demonstrated that angular motion descriptors derived from skeletal keypoints could successfully distinguish male and female motion patterns in diverse gait sequences [10].

This evidence reinforces the inclusion of motion vectors as a component in the gender bias metric framework.

2.4. Feature Extraction

To compute gender bias metrics across visual and motion dimensions, this study utilizes three deep learning models that provide complementary feature sets. Each model is selected based on its capability to capture specific visual characteristics relevant to activity recognition and human-centric video analysis.

2.4.1. Size and Centering Features (YOLO)

YOLO (You Only Look Once) is a widely adopted object detection framework designed for real-time performance and high spatial accuracy. In this study, YOLOv8 is employed to detect individuals in video frames and extract the bounding box area (for size bias) and bounding box center coordinates (for centering bias). YOLO predicts object locations in a single forward pass, enabling efficient per-frame analysis of spatial positioning and scale [11].

Figure 1 :Example of YOLO-based bounding box detection illustrating size and centering features

Figure 2 :Example of YOLO-based bounding box detection illustrating size and centering features

2.4.2. Motion Features (MediaPipe)

Motion bias is quantified by analyzing pose-based movement across video frames. For this, the MediaPipe Pose solution is used, which offers high-fidelity estimation of 33 key points per frame, including body, face, and hand joints. These pose vectors are converted into normalized motion trajectories, allowing the system to detect stylistic or dynamic differences by gender [12].

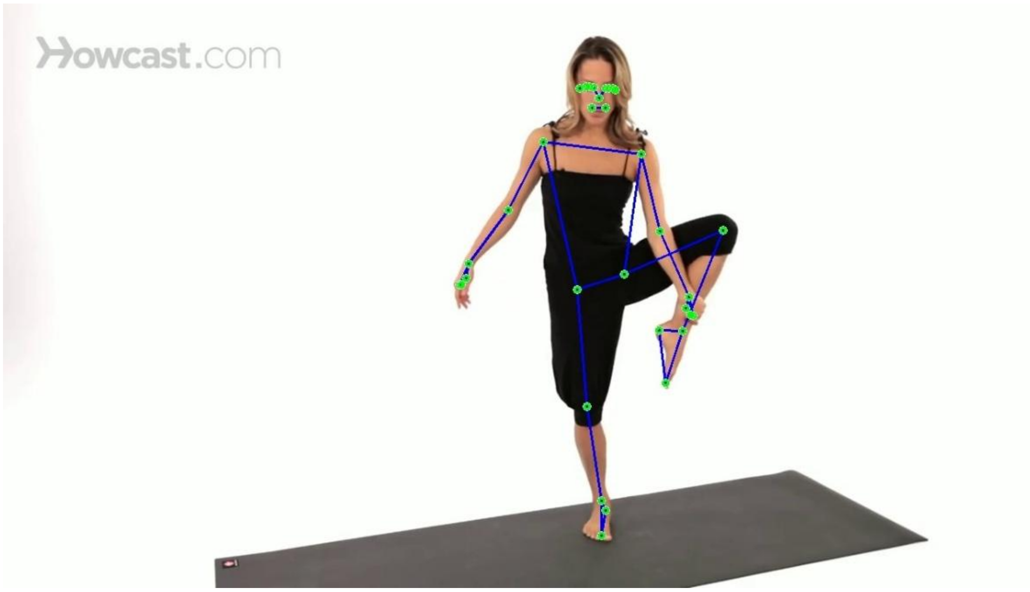

Figure 3 :Example of pose keypoint extraction using MediaPipe

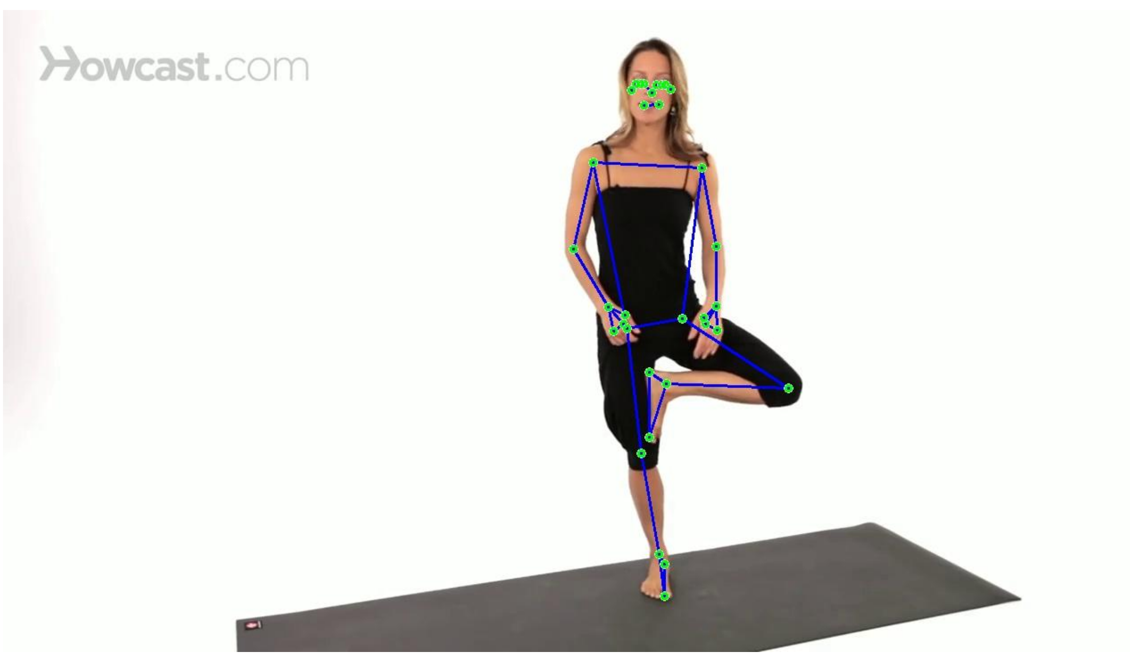

Figure 4 :Example of pose keypoint extraction using MediaPipe

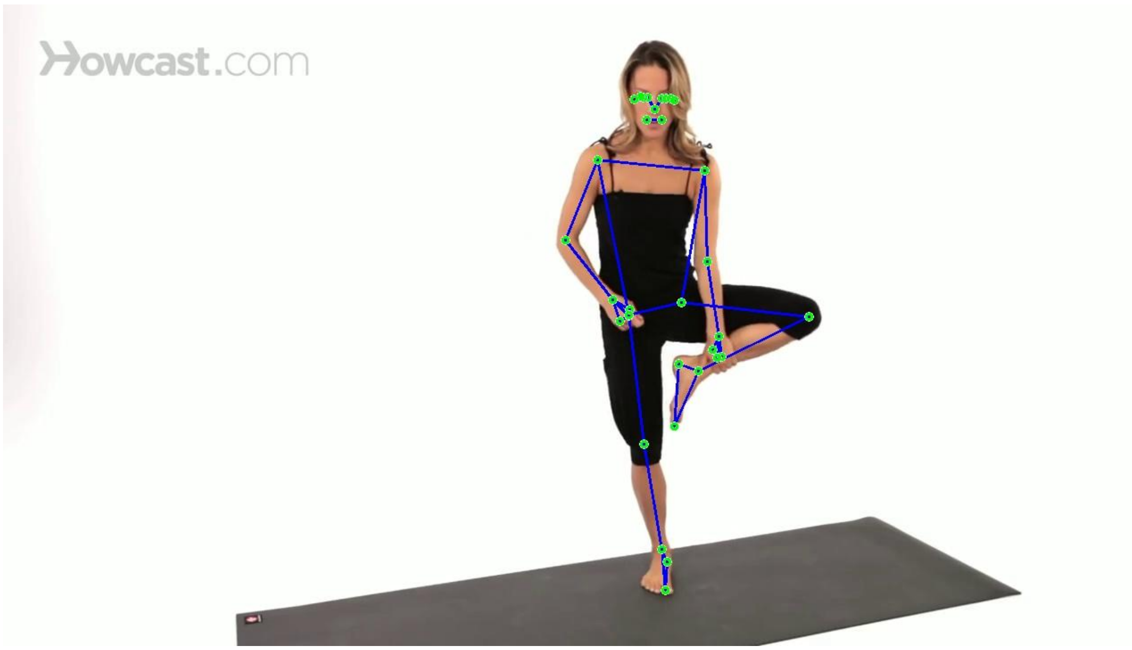

Figure 5 : Example of pose keypoint extraction using MediaPipe

2.4.3. Embedding Features (SlowFast)

To represent the semantic content of each video, the SlowFast model is used to extract high-level action embeddings. SlowFast operates by processing video at two temporal resolutions: a slow pathway for semantic context and a fast pathway for motion detail. This dual-stream architecture allows for robust feature learning from human-centric actions, which is critical for computing embedding bias using gender-specific centroid similarity [13].

```
embed_df.describe()
```

mbed_9	...	embed_2294	embed_2295	embed_2296	embed_2297	embed_2298	embed_2299	embed_2300	embed_2301	embed_2302	embed_2303
.000000	...	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000
.057707	...	0.193259	0.226268	0.012247	0.137003	0.275690	0.122027	0.077852	0.231381	0.122972	0.040282
.085807	...	0.429458	0.465742	0.044923	0.167788	0.481372	0.471572	0.101661	0.368995	0.146957	0.137827
.000000	...	0.000000	0.000042	0.000000	0.000000	0.000000	0.000015	0.000000	0.000000	0.000008	0.000000
.005608	...	0.006992	0.020240	0.001573	0.011418	0.011971	0.003936	0.011432	0.003624	0.024813	0.004012
.025679	...	0.023103	0.055136	0.003559	0.073746	0.062565	0.009783	0.041142	0.050223	0.070734	0.011679
.072255	...	0.164725	0.193045	0.008098	0.205081	0.303008	0.022280	0.103764	0.314849	0.162661	0.034549
.610676	...	4.033778	4.600695	0.780630	0.987513	2.902510	4.521287	0.771984	3.046377	1.075049	2.626102

Figure 6 : SlowFast embeddings

2.5. Bias Metric Components

To capture different dimensions of gender bias, five individual bias metrics were computed per video. Each metric targets a unique visual or motion-related attribute and is later combined into composite bias scores.

2.5.1. Size Bias

Purpose:

To measure the visual prominence of a subject based on their relative size within the video frame.

Inputs:

- x_1, y_1, x_2, y_2: Bounding box coordinates (from YOLO)
- W, H: Frame width and height

Computation:

1. Bounding box area:

$$A_{bbox} = (x_2 - x_1) * (y_2 - y_1)$$

2. Frame area:

$$A_{frame} = W * H$$

3. Size ratio per frame:

$$Rsize = \frac{A_{bbox}}{A_{frame}}$$

4. Final score (video-level mean with gender sign):

$$Size\ Bias = mean(Rsize) * \begin{cases} +1 : if\ gender\ =\ male \\ -1 : if\ gender\ =\ female \end{cases}$$

2.5.2. Centering Bias

Purpose:

To assess how centrally the subject is framed within the video.

Inputs:

- Bounding box center: $x_c\ =\ \frac{(x1+x2)}{2}$, $y_c = \frac{(y1+y2)}{2}$
- Frame center: $f_x\ =\ \frac{W}{2}$, $f_y\ =\ \frac{H}{2}$

Computation:

1. Euclidean distance to center:

$$d\ =\ \sqrt{(x_c\ +\ f_x)^2 + (y_c\ +\ f_y)^{\ 2}}$$

2. Normalized centering score:

$$D_{norm} = 1 - \frac{d}{(0.5 * \sqrt{W^2 + H^2})}$$

3. Final score:

$$Centering\ Bias = mean(D_{norm}) * \begin{cases} +1 : if\ gender = male \\ -1 : if\ gender = female \end{cases}$$

2.5.3. Screen Time Bias

Purpose:

To measure the duration of on-screen visibility for individuals in each video, while incorporating gender directionality. This metric reflects whether a video's subject contributes more to male-leaning or female-leaning representation based on screen exposure.

Inputs:

- f_v: Number of detected frames for a given video
- T_{total}: Total number of detected frames across all videos in the dataset

Computation:

Final score:

$$Screen\ Time\ Bias = \frac{f_v}{T_{total}} * \begin{cases} +1 : if\ gender = male \\ -1 : if\ gender = female \end{cases}$$

2.5.4. Embedding Bias

Purpose:

To evaluate the semantic similarity of each video to male and female action embedding centroids.

Inputs:

- V: Video embedding vector (SlowFast)
- C_{male}, C_{female}: Gender-specific embedding centroids

Computation:

1. Normalize embedding:

$$V' = \frac{V}{||V||}$$

2. Cosine distances:

$$d_m = cosine(V', C_{male}) \quad , \quad d_f = cosine(V', C_{female})$$

3. Final metric:

$$Embedding\ Bias = d_f - d_m$$

2.5.5. Motion Bias

Purpose:

To capture gender differences in movement patterns using normalized pose dynamics.

Inputs:

- P_t : Pose keypoints at time t (from MediaPipe), normalized by frame size
- C_{motion_male}, C_{motion_female} : Gender-specific motion centroids

Computation:

1. Frame-level motion vector:

$M = mean(|P_{t+1} - P_t|)$; across all frames

2. Cosine distances:

$dm = cosine(M, C_{motion\_male})$, $dm = cosine(M, C_{motion\_female})$

3. Final score:

$Motion\ Bias = d_f - d_m$

2.6. Normalization of Bias Components

Purpose

Each of the five individual bias components, Size Bias, Centering Bias, Screen Time Bias, Embedding Bias, and Motion Bias are computed on different scales and units. Without normalization, these raw values would contribute unequally to any aggregated metric, disproportionately emphasizing those with higher variance or broader ranges. For instance, centering bias values may range from −0.9 to +0.9, while screen time bias could span only a narrow band such as −0.01 to +0.01. This imbalance would distort downstream interpretations, particularly in methods like Principal Component Analysis (PCA), where variance magnitude directly affects feature weighting.

Normalization Approach

To ensure fair contribution from each component, all five metrics were standardized using Z-score normalization via *StandardScaler*. This method transforms each feature to have:

- A mean of 0
- A standard deviation of 1

Mathematically, each value x is transformed as:

$$x_{std} = \frac{x - \mu}{\sigma}$$

Where μ is the mean and σ is the standard deviation of the respective bias component. This scaling ensures that all metrics contribute equally in terms of statistical variance, regardless of their original range or distribution.

2.7. Metric Aggregation

Once the five individual bias components were normalized using Z-score standardization, three composite metrics were derived per video to capture different perspectives of bias: directionality, intensity, and data-driven weighting. These aggregated scores allow for holistic interpretation and comparison of gender bias across activities.

2.7.1. Directional Score

Purpose:

To summarize the overall gender tilt of a video by averaging its normalized bias components.

Computation:

Let B_1, B_2, B_3, B_4, B_5 represent the standardized values of the five bias components. The directional score is computed as:

$$Directional\ Score = \frac{1}{5}\sum B_i$$

Interpretation:

- Positive scores indicate male-leaning bias
- Negative scores indicate female-leaning bias
- Values close to 0 imply balanced or neutral representation

This metric is simple and interpretable, offering a general view of bias polarity.

2.7.2. Magnitude Score

Purpose: Measures the overall strength of bias, regardless of direction.

Computation:

$$\text{Magnitude Score} = \sum_{i=1}^{5} |B_i|$$

Interpretation:

- Higher scores reflect stronger or more extreme bias
- A value of 0 would indicate complete neutrality across all dimensions

This metric is useful when direction is less important than the extent of distortion in representation.

2.7.3. PCA-Weighted Score

To derive a data-driven bias score that emphasizes the components with the highest variance across the dataset.

Computation:

Principal Component Analysis (PCA) was applied to the standardized metrics, and the first principal component was extracted:

$$PCA\ Score = w_1\mathrm{B1} + w_2\mathrm{B2} + w_3\mathrm{B3} + w_4\mathrm{B4} + w_5\mathrm{B5}$$

Where w_i are the PCA weights derived from component variance.

Rescaling for Interpretability:

Since PCA scores are unbounded, the output values were rescaled to the range [-1,+1] using MinMax normalization:

$$PCA - Weighted\ Score = -1 + 2 * \frac{PCA\ Score - min(x)}{max(x) - min(x)}$$

Interpretation:

- +1: Highest male-leaning bias
- -1: Highest female-leaning bias
- 0: Balanced representation

This score captures the most statistically meaningful pattern of bias present in the data, weighted by natural variability.

2.8. Tools and Technologies

This research integrates multiple modern tools, libraries, and models from the fields of computer vision, deep learning, and data analysis. The following tools were selected based on their performance, open-source availability, and suitability.

Computer Vision Models

- YOLOv8 (You Only Look Once, Version 8)

  Used for high-speed person detection and bounding box extraction. Its real-time object detection capabilities were leveraged to compute size and centering biases with high spatial accuracy.

- MediaPipe Pose

  Provided dense pose estimation with 33 keypoints per frame, enabling the analysis of skeletal movement and motion dynamics. It was instrumental in calculating motion bias through frame-to-frame joint tracking.

- SlowFast Network (from PyTorchVideo)

  A state-of-the-art model for action recognition, used to extract semantic embeddings from each video. These embeddings were compared to gender-based centroids to calculate embedding bias.

Data Handling and Processing

- Pandas

  Used extensively for dataset management, feature aggregation, and bias score computation.

- NumPy

  Provided fast matrix operations and distance calculations, including cosine similarity and vector normalization.

- scikit-learn

  Used for normalization (MinMaxScaler), cosine distance metrics, and Principal Component Analysis (PCA) for dimensionality reduction in the aggregated score.

Development and Execution Environment

- Google Colab

  Served as a cloud-based development platform, offering GPU acceleration for SlowFast and YOLOv8 inference, and simplified dataset management via Google Drive integration.

This technology stack enabled a modular, efficient, and reproducible implementation of the multi-component gender bias detection framework.

3. RESULTS AND DISCUSSION

3.1. Bias Scores Across Activity Categories

Table 3.1 presents a comprehensive summary of gender bias scores calculated across all activity categories using three distinct metrics:

- **Mean Directional Bias Score** (Z-score standardized average of components): Indicates the direction of bias (negative = female, positive = male).
- **Bias Magnitude Score** (Sum of absolute Z-scores): Reflects the overall intensity of bias regardless of gender.
- **PCA-Weighted Bias Score** (Scaled principal component score): A composite measure that weights each metric by its contribution to variance.

These scores enable a multidimensional understanding of how gender bias manifests in different activity types. Activities such as *yoga_cat*, *yoga_bridge*, and *yoga_dancer* exhibit strong female-leaning biases across all three metrics, while *tennis_serve*, *badminton_underswing*, and *tennis_backhand* demonstrate strong male-oriented profiles.

Activity Category	Mean Directional Bias Score	Bias Magnitude Score	PCA-Weighted Bias Score
yoga_cat	-0.919092	5.979994	-0.327748
yoga_bridge	-0.770948	5.222263	-0.233092
yoga_dancer	-0.734299	4.585500	-0.256013
yoga_updog	-0.724523	5.460519	-0.162032
yoga_triangle	-0.646654	5.135346	-0.170824
gym_ride	-0.563478	5.040088	0.054926
gym_run	-0.507807	4.611012	0.116533
yoga_tree	-0.416410	4.293056	-0.000368
gym_squat	-0.226094	4.770275	-0.039488
gym_plank	-0.222724	5.389683	0.150607

gym_lift	-0.055650	3.628947	0.220343
gym_lunges	0.064973	4.297306	0.282633
backward_roll	0.095107	3.165132	0.341159
gym_push	0.185132	3.781344	0.399479
diving_jump	0.187934	2.451491	0.329341
diving_rotate	0.201408	2.396463	0.375535
gym_pull	0.293334	4.135528	0.451711
tennis_forehand	0.483019	4.008871	0.508117
golf_swing	0.502071	3.284349	0.494390
badminton_overswing	0.538724	3.831396	0.545243
backflip	0.546839	3.270509	0.559145
badminton_serve	0.603412	4.643824	0.599659
tennis_backhand	0.665701	4.115206	0.618860
badminton_underswing	0.675600	4.481038	0.627961
tennis_serve	0.744424	4.377540	0.645459

Table 1 :Activity-Wise Bias Scores Across Metrics

3.2. Gender Bias Trends in Activity Classes

To explore how gender bias manifests across different physical activity categories, the combined bias scores were analyzed at the category level. This section interprets the results using three key metrics derived from standardized bias components: **Directional Bias Score**, **Bias Magnitude Score**, and **PCA-Weighted Bias Score**.

3.2.1. Directional Bias Score

Figure 3.1 presents the average **directional bias** for each activity class, computed as the mean of standardized bias components. A positive score indicates a male-leaning bias, whereas a negative score signals a female-leaning tendency.

Notably, activities such as *tennis_serve*, *badminton_underswing*, and *tennis_backhand* exhibit strong male bias, while categories such as *yoga_cat*, *yoga_bridge*, and *yoga_dancer* reveal pronounced female bias. This dichotomy reflects the influence of traditional gender participation patterns in sports versus wellness-focused domains like yoga.

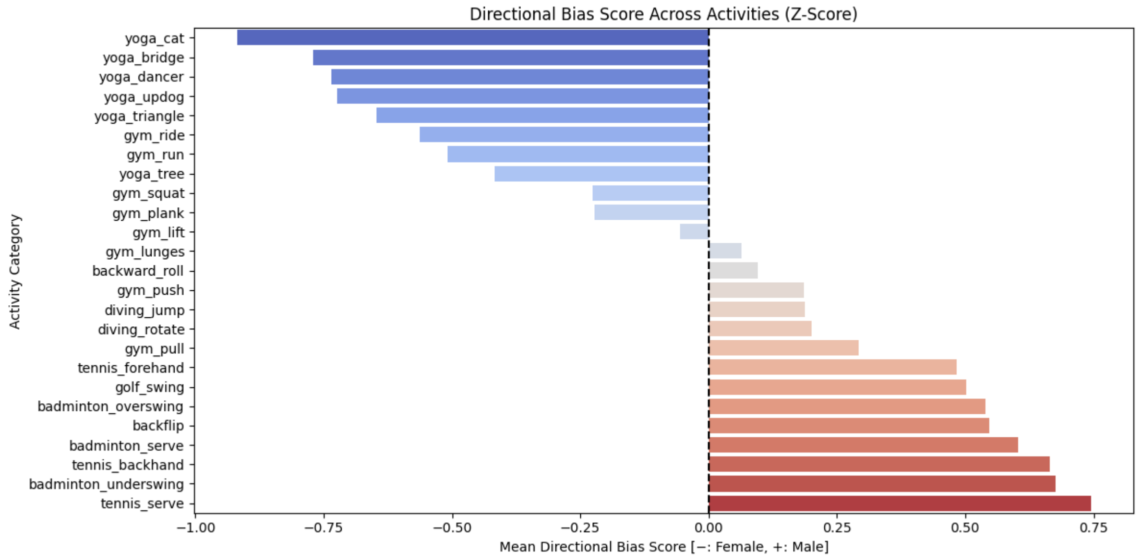

Figure 7 :Average directional bias score by activity category. Positive scores indicate male bias; negative scores indicate female bias.

3.2.2. Bias Magnitude Score

To assess the overall strength of gender bias regardless of direction, the sum of absolute standardized component values was computed (Figure 3.2). A higher magnitude implies stronger deviation from gender neutrality, even if the activity is evenly distributed between genders.

Categories such as *yoga_cat*, *yoga_bridge*, and *gym_plank* top the list in bias magnitude, revealing not just directional skew but significant visual or motion-based disparities between genders. Conversely, mid-range activities like *gym_push* and *diving_jump* show moderate bias intensities.

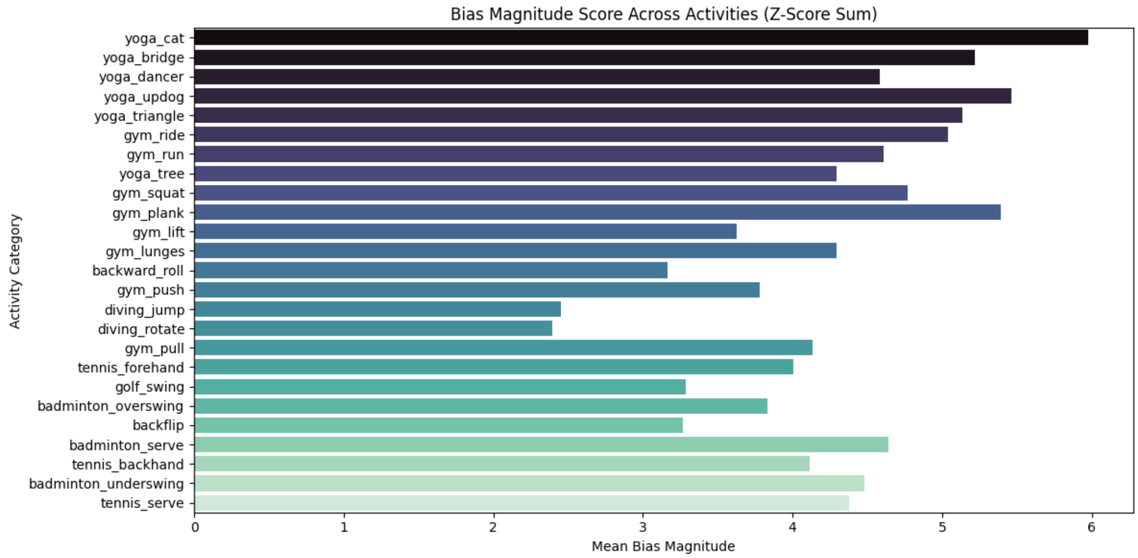

Figure 8 :Average bias magnitude by activity category. Higher values indicate stronger visual representation bias across components.

3.2.3. PCA-Weighted Bias Score

The PCA-weighted score captures the most explanatory dimension of variation in bias using Principal Component Analysis (Figure 3.3). Here, dimensional weights emphasize metrics contributing most to global bias variability.

Results show that male-dominant sports such as *tennis_serve*, *badminton_underswing*, and *backflip* have high positive PCA-weighted scores, while female-associated categories like *yoga_cat*, *yoga_dancer*, and *yoga_bridge* are positioned strongly in the negative direction. Interestingly, some categories such as *gym squat* show mild directional bias but a stronger PCA-weighted leaning, suggesting nuanced gender distinctions captured through motion and embedding-based features.

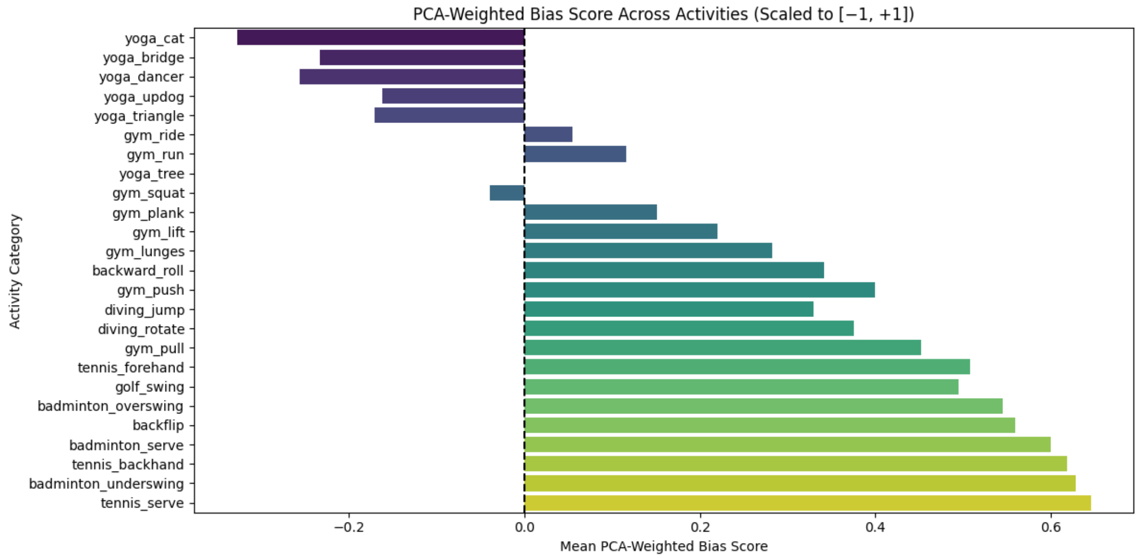

Figure 9 :PCA-weighted bias scores, highlighting categories where deeper visual features skew toward a particular gender.

Interpretation Summary

- **Yoga categories** consistently show strong female bias across all three views, aligning with demographic trends in practice participation.
- **Racket and combat sports** lean strongly toward male bias, reflecting visual, semantic, and motion differences tied to performance styles.
- **PCA-weighted scores** provide sharper differentiation in classes with mixed gender representation, highlighting the utility of multivariate aggregation.

3.2.4. Gender Composition and Bias Correlation

To understand the interplay between gender distribution and visual bias, activity categories were examined in terms of their gender representation and corresponding bias metrics. Table 3.2 presents the top ten **male-dominant** and **female-dominant** categories based on the number of videos labeled for each gender.

In the female-dominant set, categories such as *yoga_dancer*, *yoga_cat*, and *yoga_bridge* demonstrated extreme gender skew, with *yoga_dancer* containing no male samples. These categories also ranked among the most female-biased in both directional and PCA-weighted metrics, indicating strong visual, spatial, and motion characteristics more typical in female-labeled samples. The alignment between representation imbalance and bias scores is particularly evident in yoga-based classes, which also recorded the highest bias magnitude scores, suggesting consistently distinctive features across components.

Conversely, male-dominant categories such as *tennis_serve*, *badminton_underswing*, and *backflip* were associated with high positive values in both directional **and** PCA-weighted bias scores. These results reflect not only demographic skew but also substantial visual and motion-specific emphasis toward male-labeled samples. The strong correlation in these categories indicates that participation imbalance amplifies measurable bias in both component and composite metrics.

However, not all categories with strong bias exhibited large gender disparities. For example, *gym_plank* and *gym_push* displayed moderate gender counts but still produced noticeable directional and PCA-weighted biases. These cases suggest that content-level factors such as pose movement, camera framing, or screen prominence contribute independently to visual bias beyond gender frequency alone.

🔵 Top Male-Dominant Categories			🔴 Top Female-Dominant Categories		
Category	**M**	**F**	**Category**	**M**	**F**
backflip	18	2	yoga_dancer	0	20
tennis_serve	17	3	yoga_cat	2	18
tennis_backhand	17	3	yoga_bridge	2	18
badminton_underswing	16	4	gym_squat	3	17
badminton_serve	16	4	yoga_triangle	3	17

badminton_overswing	15	5	yoga_updog	4	16
golf_swing	15	5	yoga_tree	5	15
gym_pull	14	6	gym_ride	6	14
tennis_forehand	14	6	gym_run	7	13
gym_push	13	7	gym_plank	9	11

Table 2 :Most gender-dominant activity categories based on video counts (Top 10 per gender

3.3. Interpretation and Significance

The integration of component-level bias metrics provides deeper insights into which features, such as spatial prominence, motion dynamics, or semantic embeddings, contribute to perceived gender biases in video-based datasets.

Key Takeaways:

- Bias is feature-dependent:

The use of multiple metrics revealed that some activities (e.g., *gym_plank*, *gym_push*) displayed relatively balanced gender representation but still exhibited bias in motion or pose-centric components. This suggests that bias can emerge from stylistic framing, not just frequency.

- Directionality ≠ Magnitude:

While some activities may lean slightly toward one gender, the magnitude metric emphasizes whether this tilt is visually and semantically strong. Thus, a small directional score can still correspond to strong bias if variation across metrics is high.

- PCA-weighted bias generalizes variance:

The use of PCA assigns weights to each metric based on how much they contribute to inter-video variance. Activities with high PCA scores (e.g., *tennis_serve*, *badminton_serve*) show not only directional tilt but also feature-distinct clustering, suggesting stronger model-learnable bias.

- Correlation with representation:

Bias scores correlate with gender distribution, but not always linearly. Some categories with relatively balanced gender count still exhibit high composite bias, implying that qualitative visual encoding (like pose expressivity or centrality) influences perception and algorithmic weighting more than raw counts.

Implications:

The findings emphasize the need for balanced dataset construction—not only in terms of quantity but also in visual and semantic diversity. Bias in video datasets can propagate through deep models trained on such inputs, ultimately affecting action recognition outcomes, surveillance applications, and content recommendation systems.

Moreover, the proposed framework, which combines interpretable metrics with variance-aware aggregation, offers a generalizable approach for bias diagnosis and remediation in computer vision pipelines.

3.4. Limitations

While the proposed framework provides a systematic and interpretable approach to measuring gender bias in human activity video datasets, several limitations should be acknowledged to contextualize the scope and applicability of the findings.

- **Dataset-Specific Generalizability**

The analysis was conducted using a single video dataset with a fixed structure and predefined activity classes. Although the results are internally consistent, they may not generalize across datasets that vary in cultural context, resolution, or collection methodology. Cross-dataset validation is necessary to confirm the broader applicability of the framework.

- **Binary Gender Constraint**

The current methodology operates under a binary gender classification (male/female), which, while simplifying the computational model, does not account for non-binary, transgender, or other gender-diverse representations. This limitation reflects broader dataset labeling practices and underscores the need for more inclusive annotations in future work.

- **Proxy Indicators of Bias**

Some bias components such as screen time, centering, and bounding box size—are used as proxies for representational prominence. These indicators may not fully capture the sociocultural dimensions of bias or viewer perception. Nevertheless, they provide useful, quantifiable signals that can be aggregated for pattern detection and comparison.

- **Label and Detection Fidelity**

The reliability of bias metrics is directly tied to the accuracy of person detection, pose estimation, and gender labeling. Although robust models (YOLOv8, MediaPipe, SlowFast) were used, errors in key point extraction or gender assignment can introduce noise, particularly in low-light or occluded scenes.

- **Component Independence Assumption**

Each metric was computed independently, assuming orthogonality across spatial, motion, and embedding dimensions. However, real-world biases may involve interdependencies, for instance, motion dynamics may correlate with framing style. While independence simplifies analysis, future extensions could explore feature interactions or multivariate coupling.

Conclusion

Despite these limitations, the framework provides a transparent, extensible foundation for identifying and analyzing visual gender bias in activity recognition datasets. Rather than delivering definitive judgments, the metrics serve as diagnostic tools to guide further exploration, dataset auditing, and fairness-oriented model improvements.

4. CONCLUSION AND FUTURE WORK

4.1. Summary of Findings

This study proposed a multi-component framework to detect and quantify gender bias in human activity video datasets. Five individual metrics (Size Bias, Centering Bias, Screen Time Bias, Embedding Bias, and Motion Bias) were designed to capture diverse visual, spatial, and dynamic dimensions of representation. These component scores were normalized using Z-score standardization and further aggregated into three composite bias metrics: directional mean, magnitude, and a PCA-weighted score.

The empirical evaluation revealed significant gender bias across activity categories. Female-dominant classes, such as *yoga_dancer* and *yoga_cat*, consistently showed strong negative bias scores, while sports-centric categories like *tennis_serve* and *badminton_serve* displayed male-leaning tendencies. Importantly, some classes exhibited substantial bias despite balanced gender counts, highlighting the influence of framing, pose, and motion style in visual representation.

The framework not only enabled dataset-wide auditing but also facilitated bias pattern recognition at the category level, offering a replicable method for fairness analysis in video-based machine learning pipelines.

4.2. Contributions

The key contributions of this research are as follows:

- **Novel Bias Metric Design:**
  Introduced a modular, interpretable set of visual bias components tailored for human activity recognition videos.
- **Multi-Metric Aggregation Strategy:**
  Developed a three-tier bias scoring approach (directional, magnitude, PCA-

weighted) to capture both directionality and strength of bias, enhancing robustness and interpretability.

- **Empirical Analysis of Activity-Specific Bias:**
  Provided detailed quantitative and visual insights into how bias varies across activities, correlating scores with gender composition and visual framing styles.
- **Scalable Framework for Dataset Auditing:**
  Created a reproducible pipeline integrating YOLOv8, MediaPipe, and SlowFast models, enabling scalable feature extraction and bias computation.

4.3. Future Work Recommendations

To enhance the scope, fairness, and societal relevance of this work, several directions are recommended:

- **Beyond Binary Gender:**
  Extend the methodology to support multi-gender or non-binary annotations, allowing for more inclusive bias detection.
- **Human-Centric Validation:**
  Incorporate perceptual studies or expert annotation to align computational metrics with perceived fairness and social impact.
- **Cross-Dataset Benchmarking:**
  Apply the framework across multiple datasets to evaluate generalizability and dataset-specific bias artifacts.
- **Real-World Model Impact Testing:**
  Investigate how dataset bias metrics influence downstream models in tasks like activity recognition, recommendation, or behavioral prediction.

This work offers a first step toward a deeper, quantifiable understanding of gender bias in activity recognition datasets and provides a scalable foundation for building more equitable computer vision systems.

5. References

[1] M. G. M. R. K. N. e. a. Ali Abdollahi, "GABInsight: Exploring Gender-Activity Binding Bias in Vision-Language Models," *Frontiers in Artificial Intelligence and Applications,* 2024.

[2] D. Z. A. W. e. a. Nicole Meister, "Gender Artifacts in Visual Datasets," in *IEEE International Conference on Computer Vision (ICCV)*, Paris, France, 2023.

[3] Y. H. T. B.-M. G. Choon-Boon Ng, "Recognizing Human Gender in Computer Vision: A Survey," in *Pacific Rim International Conference on Artificial Intelligence*, Kuching, Malaysia, 2012.

[4] K. Q. I. C. K. K. G. P. Q. N. K. H. O. R. Zeyu Wang, "Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, 2020.

[5] A. N. O. R. Angelina Wang, "REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets," in *European Conference on Computer Vision (ECCV)*, Glasgow, UK, 2020.

[6] R. B. L. F. V. N. L. R. M. P. e. a. Tiago Pagano, "Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods," *Big Data and Cognitive Computing,* vol. 7, no. 1, 2023.

[7] P. R. A. C. Francois Hu, "Parametric Fairness with Statistical Guarantees," *arXiv preprint arXiv:2310.20508*, 2023.

[8] C.-h. W. H.-r. Y. Y.-W. T. C.-K. T. Jihoon Chung, "HAA500: Human-Centric Atomic Action Dataset with Curated Videos," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada, 2021.

[9] Z.-Y. D. T. W. A. C. N. P. Haoyi Qiu, "Gender Biases in Automatic Evaluation Metrics: A Case Study on Image Captioning," *arXiv.org*, 2023.

[10] I. T. G. E. S. F. Dimitrios Kastaniotis, "Gait-based Gender Recognition Using Pose Information for Real Time Applications," in *International Conference on Digital Signal Processing*, Santorini, Greece, 2013.

[11] K. J. S. K. S. A. S. R. Upendra Kumar Dwivedi, "An Overview of Moving Object Detection Using YOLO Deep Learning Models," in *IEEE International Conference on Data Technologies (ICDT)*, Dehradun, India, 2024.

[12] C. e. a. Lugaresi, "MediaPipe: A Framework for Building Perception Pipelines," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, USA, 2019.

[13] H. F. J. M. K. H. Christoph Feichtenhofer, "SlowFast Networks for Video Recognition," in *IEEE International Conference on Computer Vision (ICCV)*, Seoul, South Korea, 2019.

[14] S. B. D. M. Venkata Naresh Mandhala, "Need of Mitigating Bias in the Datasets using Machine Learning Algorithms," in *International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, Chennai, India, 2022.

[15] B. S. Mostafa M. Mohamed, "Normalise for Fairness: A Simple Normalisation Technique for Fairness in Regression Machine Learning Problems," *arXiv.org*, 2022.