The Datasets Behind Human Action Recognition in Computer Vision Research

by

Bhavya Gutha

(Under the Direction of Ari Schlesinger)

Abstract

While datasets are foundational to the development of action recognition systems, data work is often marginalized in favor of model innovation. In this thesis, I studied how data work is portrayed by action recognition researchers via the introduction and discussions of datasets containing videos of humans in action recognition research publications. I analyzed the use of 967 unique action recognition datasets across 1,320 publications. These publications were collected from three top computer vision conferences—CVPR, ICCV, and ECCV. This study finds that human subjects are frequently abstracted through metadata and discussed with minimal contextual detail. Emphasis is typically placed on attributes like scale, novelty, and technical difficulty, while demographic diversity and cultural specificity are overlooked. Additionally, the absence of a standardized definition of "action" limits the ability to consistently design and evaluate these datasets. These patterns reflect broader cultural values within computer vision that privilege technical innovation over human-centered considerations.

Index words: [Action Recognition, Video Datasets, Dataset Representation]

The Datasets Behind Human Action Recognition in Computer Vision Research

by

Bhavya Gutha

B.A., University of Georgia, 2023
B.S., University of Georgia, 2023

A Thesis Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Master of Science

Athens, Georgia

2025

The Datasets Behind Human Action Recognition in Computer Vision Research

by

Bhavya Gutha

Major Professor:	Ari Schlesinger
Committee:	Jin Sun Kimberly Van Orman

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
August 2025

Dedication

To Amma, Nanna, and Aditya for their unwavering love and support. To my dog, Spark, for being the sweetest. To my friends for always having my back.

Acknowledgments

I am deeply grateful to my committee members who have supported and guided me through the MSAI program. I would like to express my heartfelt thanks to Dr. Schlesinger for her encouragement and thoughtful feedback—she has challenged me to think more critically and has made me a better writer and scholar. I am also sincerely thankful to Dr. Sun for his patience, steady mentorship, and consistently thoughtful advice. Last, but certainly not least, I am especially grateful to Dr. Van Orman, whose teaching and guidance have profoundly shaped my experience in the program.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Computer vision is a field of artificial intelligence focused on teaching computers to see, so they can understand and interact with people and environments. We want computers to see people so they can respond appropriately to human behavior, which enhances safety and accessibility in real-world applications. Some of the core tasks in this field are image classification, object detection, and pose estimation. To understand how computers see we investigate datasets, particularly what they contain and how they are built.

Human action recognition is a subfield of computer vision focused on recognizing and understanding what people are doing in videos. Many of its applications are in high stakes scenarios, where malfunctioning in the algorithm could result in life threatening consequences. This includes applications like self-driving cars, where one of the things its programming should constantly monitor and identify is pedestrian activity, lest someone is grievously injured. Another application is patient monitoring in healthcare facilities, like recognizing if an elderly patient has fallen and ensuring there is appropriate follow through for their safety. For these algorithms to perform reliably and equitably, they must be trained on datasets that have diverse representations of humans. This research investigates how publications describe the video datasets used for human action recognition, with attention to how they frame and justify their data choices.

1.1 Motivation for the Study

Despite the critical role that datasets play in advancing action recognition, data work is often undervalued and rarely published as a standalone contribution [65, 66]. While this issue has been broadly acknowledged in computer vision, little attention has been paid to how it manifests specifically in the context of action recognition. This thesis addresses that gap by examining how datasets are introduced and described in action recognition literature. These practices, I argue, reflect broader cultural tendencies within the field, where model development is routinely prioritized over data creation and curation. When data work is marginalized, it shapes not only how datasets are built, but how they—and the people within them—are represented in scholarly conversations. Understanding these patterns is essential for recognizing how the field's values and priorities are reinforced through research practices, and for promoting greater transparency and accountability in how datasets are used.

1.2 Research Objective

One of the objectives of this research is to analyze how action recognition papers in top computer vision conferences introduce and discuss the datasets containing videos of humans, as well as what is excluded from the conversation. In line with the research motivations, the analysis primarily centers around the discussion of the humans present in the data. Furthermore, the justifications for using or creating a dataset is examined for human-centric reasoning. As part of this research, the analyzed papers are compiled into a corpus, and the identified datasets are organized into a repository.

1.3 Research Questions

The question that this research attempts to answer in accordance with the research motivations and objectives is:

- How do action recognition researchers describe and introduce video datasets containing humans?
- What aspects of datasets do researchers emphasize? What do they overlook?
- What do all of these things reveal about the priorities of action recognition researchers?

1.4 Thesis Structure

The remainder of this thesis is organized as follows: Chapter 2 provides a literature review of relevant topics. Chapter 3 outlines the research methodology. Chapter 4 presents the results from the analysis of action recognition publications. Chapter 5 interprets these results to examine the underlying values in action recognition research. Chapter 6 discusses the limitations of this study and suggests directions for future work. Finally, Chapter 7 offers concluding remarks.

Chapter 2

LITERATURE REVIEW

This research is informed by a growing body of scholarship that examines the values embedded in computer vision practices, particularly in the development of datasets and algorithms. It focuses on a specific subfield: human action recognition, which aims to detect and interpret human activity in video data. This technology is increasingly used in high-stakes contexts, where algorithmic errors can have life-threatening consequences. For instance, self-driving cars must reliably identify pedestrian movement to prevent serious harm, while healthcare systems may use action recognition to monitor patients—such as detecting when an elderly person has fallen—to ensure timely intervention. Because datasets form the foundation of these models, this study analyzes how action recognition papers published in top computer vision conferences introduce and describe the datasets containing videos of humans, as well as what is left unaddressed. The goal is to understand what dataset qualities action recognition researchers—and the field at large—tend to prioritize or overlook.

2.1 Values Shape Data Collection

In 1996, Friedman and Nissenbaum identified three categories of biases in computer systems: preexisting, technical, and emergent [24]. Preexisting bias is deeply rooted in societal institutions, practices, and attitudes. It can infiltrate systems through explicit and conscious channels, such as deliberate design choices or policies, as well as through implicit and unconscious means, such as cultural norms or stereotypes that influence decision-making without awareness. This type of bias often originates from the environment in which the system is developed, reflecting the biases of the designers, the datasets they choose, and the broader societal context. Technical bias "arises from technical constraints or technical considerations", like algorithms that are developed without the context of social meaning. Meanwhile, emergent bias only becomes apparent after a system is deployed and users begin to interact with it. This type of bias often arises as a result of changes in "societal knowledge, population, or cultural values". For example, a system designed with data that lacks diversity in gender, race, or socio-economic background may perform poorly for underrepresented groups, leading to unintended discriminatory outcomes. These emergent biases can be particularly insidious because they may not be evident during the development phase but become problematic as the system scales and interacts with a broader user base.

Effectively identifying and mitigating these forms of bias requires addressing their root cause: the data on which systems are built. As the foundational layer of models, datasets shape every stage of system behavior. Issues at this stage can lead to data cascades—a term coined by Sambasivan et al. [65] to describe "compounding events causing negative, downstream effects from data issues, resulting in technical debt." In response, a growing body of scholarship have been studying publications in computer vision, as well as machine learning as a whole, to understand what values the fields prioritize and the expense this come with [4, 55, 65, 66, 67]. These works have been calling for greater author reflexivity, and in the context of dataset duration, this entails transparency about the decisions made during dataset construction and how those decisions are shaped by the positionality of all involved actors—dataset creators, annotators, and the individuals represented in the data. Acknowledging these positionalities is essential for understanding who is included in a dataset—and who is left out—and for anticipating how systems trained on such data may perform when applied to underrepresented or excluded groups [55, 66, 67].

The 2021 paper by Scheuerman et al., *Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development* [66], greatly influenced this research . In this work, the authors investigate "what dataset documentation communicates about the underlying values of vision data and the larger practices and goals of computer vision as a field". They compiled a corpus of computer vision datasets and examined how dataset creators described the curation process—both in terms of content and framing—to uncover what these practices suggest about how datasets are valued within the field. They found that computer vision dataset authors generally value efficiency, universality, impartiality, and model work over care, contextuality, positionality, and data work. These preferences affect not only the kind of data collected, but also the quality. For instance, annotation procedures are typically framed as objective, with quality control designed to enforce consistency among annotators—yet this pursuit of universality often suppresses the cultural or situated context of human experience that inevitably shapes their interpretation of actions and how that is represented in the data. Further, dataset creators rarely reflect on their own positionality or on how their choices influence what the dataset comes to represent.

Similar to Scheuerman et al.'s (2021) [66] work in computer vision, Birhane et al. (2022) [4] conducted research on the values encoded in machine learning research by analyzing 100 highly cited machine learning papers. Their study identified dominant values in research papers, such as performance, generalization, quantitative evidence, efficiency, building on past work, and novelty. They found that machine learning research prioritizes contributions to the specific research community, with minimal attention to how the work serves broader societal needs. Further, there is a lack of ethical reflection amongst the publications, where they frequently omit discussions on the potential negative impacts of the research. Many papers neither explicitly address these impacts nor provide reasons for not doing so. In fact, the publications frequently failed to explain why these applications are necessary or deserving of further development. Additionally, Birhane et al. conduct a quantitative analysis of the affiliations and funding sources of these research papers, revealing a significant increase in the involvement of big tech companies. This trend suggests a growing influence of these companies in technological research, which is likely to encourage new technologies that more likely to serve corporate means rather than aiding the average person.

Many scholars advocate for reflexive thinking to encourage awareness amongst researchers of how the assumptions they operate under shape their work—especially in dataset construction [55, 66, 67, 68]. Miceli et al. [55] propose reflexive documentation, a practice in which researchers actively reflect on their decisions during the research process and keep records of these reflections. This includes examining why particular choices are made, what assumptions underlie those decisions, and how these choices might influence the resulting data or research outcomes. According to Miceli et al., reflexive documentation makes visible the hierarchies and worldviews driving dataset creation. This practice encourages critical reflection rather than treating design choices as neutral or self-evident. Consciously examining and articulating the rationale behind their choices compels researchers to confront and question these underlying assumptions and help them view data as things with human influence rather than impartial investigations. By requiring researchers to justify choices they made while operating under implicit assumptions, they are more likely to consider alternative perspectives and seek a broader range of data sources. Documenting and sharing the reasoning behind their decisions helps researchers provide a clearer understanding of how the data was curated under which value systems.

A commonly proposed practice is for researchers to report the demographics of both authors and annotators [66, 67]. This information helps situate the backgrounds of annotators in relation to the data they produce, which is important given that domain expertise can significantly shape annotation quality. Annotators with relevant knowledge may interpret data differently than those without. For researchers, positionality statements [66] offer a way to reflect on how their identities, perspectives, and values influence their research decisions and data collection. These self-disclosures can increase transparency by making visible what may assumptions shape the dataset's construction.

However, as Schlesinger et al. [68] caution, such disclosures are not without risk. One major concern is the potential for reinforcing existing biases—particularly against women and scholars from underrepresented racial or ethnic backgrounds. In a research culture where systemic inequities persist, revealing demographic information could expose authors from marginalized communities to increased scrutiny or cause their work to be unfairly devalued. There is a legitimate fear that reviewers or readers might, consciously or unconsciously, discount the expertise or credibility of researchers based on gender or ethnicity, ultimately undermining the goals of equity and inclusion that such transparency aims to support.

Meanwhile, some scholars—such as Chen and Sundar [11]—propose that transparency in data construction can also be enhanced by involving users directly and giving them access to a snapshot of the underlying data driving the systems they interact with. The premise is that when users better understand how a system functions, they are more equipped to appropriately calibrate their trust in it. This transparency helps mitigate automation bias—a tendency to trust algorithmic outputs even when they are clearly incorrect or inconsistent.

2.2 Model Performance Reflects Dataset Values

The lack of reflexivity in dataset construction has material consequences. In their landmark study *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* [6], Buolamwini and Gebru exposed significant racial and gender disparities in commercial facial analysis systems. They found that these systems consistently performed better on male faces than female faces, better on lighter-skinned faces than darker-skinned ones, and worst of all on darker-skinned women. Their audit of two widely used facial analysis datasets—Adience and IJB-A—revealed stark imbalances: 86.2% of Adience and 79.6% of IJB-A consisted of images of lighter-skinned individuals, while darker-skinned women made up only 7.4% and 4.4% of those datasets, respectively. Notably, IJB-A is a government-sponsored dataset intended to be geographically diverse, and Adience is a gender classification dataset.

These findings echo Friedman and Nissenbaum's [24] framework of preexisting, technical, and emergent biases in computer science. The skewed dataset composition reflects preexisting bias. The disproportionately poor performance of facial recognition systems on darker-skinned women exemplifies emergent bias. Together, these compounding biases contribute to data cascades, which, as stated previously, are "compounding events causing negative, downstream effects from data issues, resulting in technical debt" [65].

Despite being the foundation of machine learning models—and the fact that poor-quality data can lead to serious downstream harms—data work is often undervalued and de-prioritized [65, 66]. It is frequently viewed as auxiliary rather than as a core contribution worthy of recognition. As one interviewee in Sambasivan et al.'s study put it, "Everyone wants to do the model work, not the data work." Model development is more closely associated with academic prestige and career advancement, whereas data-centric research typically struggles to gain visibility unless paired with novel model contributions [65, 66]. Scheuerman et al. (2020), in *How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis* [67], further illustrate this point by showing how race and gender are frequently operationalized in image datasets without clear definitions or justifications. Their analysis reveals a widespread lack of transparency in labeling decisions—further evidence that data work is not held to the same reflective or methodological standards as modeling efforts.

2.3 Where This Thesis is Situated

For the most part, the existing literature analyzing datasets in computer vision focuses on image datasets, particularly those for facial recognition. This study distinguishes itself from existing research by focusing on how datasets for action recognition containing videos of humans are represented in literature. Inspired by Scheuerman et al. (2021) [66], this research investigates how action recognition researchers introduce and describe datasets containing videos of humans. The datasets that were identified from these papers were compiled into a repository. Special attention was paid to publications that introduce datasets. The goal is to identify the values and priorities of the action recognition community, and of computer vision more broadly.

Chapter 3

METHODOLOGY

This chapter has four sections; paper corpus, analysis, repository construction, and limitations. The paper corpus section collects a corpus of papers on action recognition published in the top three computer vision conferences, listed in Section 3.1. The analysis section discusses the methods used to analyze the introduction and description of datasets in these papers. The repository construction section details how the datasets containing video data of humans found in the papers were collected into a repository. Finally, the limitations section discusses the limits of my methodology and how rigor and quality was ensured.

3.1 Paper Corpus

To obtain a corpus of papers for analysis, I used the ranking provided by the Core Conference Ranking Database [1] to find the top computer vision conferences. An A* ranking is the highest ranking possible, and the only three that meet this criterion are the following conferences: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE International Conference on Computer Vision (ICCV), and European Conference on Computer Vision (ECCV). After searching “action recognition” (quotation marks included) in all three conferences, I exported the resulting papers into three Microsoft Excel spreadsheets with respect to each conference. The inclusion of the quotation marks was to ensure retrieval of only publications that specifically focus on action recognition, while minimizing the inclusion of papers that merely mention the words “action” and “recognition” separately. This method yielded a corpus of 1,878 papers.

Publications that did not use datasets containing video data of humans were excluded from the analysis, resulting in 1,320 papers for analysis. Please refer to Table 3.1 for the breakdown of the paper corpus by conference.

3.2 Analysis

The analysis of the corpus was done using inductive thematic analysis. Thematic analysis is “a method for identifying, analysing and reporting patterns (themes) within data” [5]. It is an iterative and flexible process

Table 3.1: Paper Corpus Breakdown by Conference.

Conference	Number of Papers
CVPR	424
ICCV	334
ECCV	562
Total	**1320**

that generates rich descriptions of the data. Inductive thematic analysis is the process of coding the data "*without* trying to fit it into a pre-existing coding frame, or the researcher's analytic preconceptions" [5] (emphasis original). In practice, this means that I did not begin with a predetermined rubric for identifying patterns in the papers. Instead, the rubric developed organically as I read the papers and became more familiar with recurring themes.

Because this process is iterative, it allowed the rubric to be continuously refined as the analysis progressed: new items were added, overlapping categories were consolidated, and earlier phases of analysis were revisited to identify patterns that may have been missed during the first pass. Throughout this process, I held regular meetings with my major professor, Dr. Schlesinger, and committee member, Dr. Sun, to discuss progress and preliminary findings. Based on these discussions, we frequently adjusted the rubric to better capture patterns within the data. The final version of the rubric is presented in Table 3.2.

The analysis began with publications extracted from CVPR. I systematically reviewed these papers using the Excel spreadsheet where they were stored, identifying and cataloging datasets that contained videos of humans. Initially, I focused on noting the presence of such datasets, recording their names, and tracking how many datasets were referenced in each paper. As the analysis progressed, I expanded the tracked information to include additional details, such as whether authors actively used the datasets or merely mentioned them. I also collected excerpts of dataset descriptions from each paper and tracked the language authors used when introducing or discussing the datasets—all of which were documented within the same Excel spreadsheet.

Throughout this process, I held regular meetings with Dr. Schlesinger and Dr. Sun to review and refine the tracked features. New features were added as they emerged and were deemed relevant to the analysis. By the time I reviewed the final conference, ECCV, these evolving features had developed into the structured rubric presented in Table 3.2.

3.3 Repository Construction

Parallel to the analysis, I compiled the identified datasets into a repository in a separate Excel spreadsheet, available at the following link: `https://doi.org/10.5281/zenodo.15866883`. All included datasets contain videos of humans, ranging from videos of human hands to individual humans to groups of humans. Synthetic representations of humans, such as video data of human characters from the video

Table 3.2: Rubric for Data Analysis

This rubric was used to systematically analyze papers using categories like dataset usage and dataset descriptions.

Rubric Category	**Description**
Focus	What the focus of paper is.
Focus On Action Recognition?	Whether or not the paper focuses on action recognition
Video Data Present?	Whether a dataset containing video data is used in the paper
# Found	Total number of datasets found
Introduces New Dataset?	Whether or not the paper introduces a new dataset
Dataset Introduced	The name of the new dataset
Datasets Used	The datasets used
# Used	The total number of datasets used
Datasets Mentioned	The datasets mentioned
# Mentioned	The total number of datasets mentioned
Dataset Naming Issues?	Any issues found in the paper regarding the names of the datasets listed in either the introduced, named, or mentioned columns
# Unnamed	The number of datasets that were unnamed
Dataset Descriptions	Excerpts from the papers of where the datasets were described or discussed
Citation Issues?	Any citation issues noticed in relation to a used or mentioned dataset
Notes	Notes on what aspects of the datasets are mentioned/discussed in the paper
Purpose/Use of Datasets	What the paper states to be the purpose of the dataset and/or how the paper uses the dataset

game *Grand Theft Auto V*, were also included, as they serve as a form of human representation and may reflect assumptions about human movement. Why these criteria were selected will be further expanded upon in the discussion section.

The repository was designed using Excel formulas and cross-linked spreadsheets to track dataset usage across the three conferences. It records how frequently each dataset appeared and highlights the five most frequently used datasets. A dataset is referred to as a *used dataset* if a publication uses it but does not introduce it as a new dataset. If a publication does introduce a new dataset for the first time, it is referred to as an *introduced dataset*. A single dataset can be labeled as both *introduced* and *used* if it is introduced by one publication and subsequently used by others within the analyzed corpus. For each dataset, the repository includes detailed information such as usage across conferences, introduction history, related datasets, collection methods, content, intended purpose, and links to relevant websites and introducing papers.

3.4 Limitations

This methodology is limited by its scale, as well as time and resource constraints. Inductive thematic analysis is inherently time-consuming due to its iterative nature, which posed a challenge given the size of the corpus: 1,320 papers analyzed by a single researcher—myself. Despite these limitations, the rigor and quality of the analysis was ensured through frequent meetings with my advisors and the adoption of reflexive practices, in which I regularly evaluated my approach and progress. Additionally, due to the scale of the repository and limited time, comprehensive information was systematically maintained only for introduced datasets and the five most frequently used datasets.

Chapter 4

RESULTS

This chapter will begin with the ways a small subset of papers define "action" in computer vision. Next, I will provide an overview of the repository using the introduced datasets. I will then examine how these publications justify the creation of new datasets, describe the dataset, and discuss the inclusion of human subjects. Finally, I will detail how publications reference the video datasets they use, including what the five most commonly used datasets across the conferences are.

4.1 How is "Action" Defined in Computer Vision?

Out of the 1,320 papers examined, only two—Chen et al. (CVPR 2014) [12] and Luo et al. (ICCV 2015) [50]—attempt to define what "action" means within the context of computer vision. To do so, they draw from philosopher Donald Davidson's definition of "action" in his 1963 paper *Actions, Reasons, and Causes* [18], which Luo et al. summarizes as "intentional biological movement". Both papers recognize that this definition is not entirely compatible with action recognition due to the inclusion of "intent" in the definition. Chen et al. attempt to adapt Davidson's definition of "action" so that it is suitable for computer vision, because computer vision "cares more about what visual patterns an action may present than the philosophy of action." Luo et al. take a different approach, where they adhere strictly to Philosophy of Action's definition of "action", and assert that actions done unintentionally, like falling, should not be included in datasets like HMDB51 [39], unless they were done intentionally. To assess whether a given movement qualifies as intentional, both papers introduce the concept of "actionness" but their definitions differ. According to Chen et al., "actionness" is "intentional bodily movement of biological agents", while Luo et al. defines it as low-level attributes characterized by four cues that makes salient actions stand out.

Chen et al. base their definition of "actionness" in the four aspects of "action" according to Davidson. As described by Chen et al., they are: "first, action is what an **agent** can do; second, action requires an **intention**; third, action requires a **bodily movement** guided by an agent or agents; and fourth, action leads to **side-effects**" (emphasis original). Chen et al. consider the words they highlighted to be the key terms of these aspects, where "agent" is usually a person or an animal and "side-effects" are the results of the action. They use these terms to define "actionness" as "intentional bodily movement of biological agents," stating that it is "a subclass of general motion and a direct presentation of action." The inclusion of "intention" in their definition is justified by the explanation that "a non-biological agent, such as a bicycle, can not have intention, and hence the agents we care about are the people and animals.

Luo et al. disagree with Chen et al.'s definition of "actionness", claiming that they define "actionness entirely in high-level terms and assumes that the notion of action can be defined implicitly by annotated examples in a dataset." Instead, they propose an alternative approach, stating, "we explicitly posit that actionness can be defined in terms of low-level operations" and the strength of their definition is substantiated by the improved results they see in action detection. Drawing from the psychology paper *Chasing vs. stalking: Interrupting the perception of animacy* by Gao and Scholl [25], these "low-level operations" consist of the following 4 cues that are not mutually exclusive:

1. Sudden changes in direction and speed—indicate impulsive movements like striking or kicking.
2. Repetitive motions over time—characteristic of sustained movements such as walking or running.
3. Temporal synchrony—well-coordinated movements that are, either within a person (e.g., diving), between people (e.g., shaking hands), or between a person and an object (e.g., drinking). These indicate intent.
4. Association with salient regions—actions typically involve agents and objects rather than background elements.

Furthermore, in the second and third cues, Luo et al. give preference to trajectories that show significant direction changes because it indicates self-propelled motion, or movement with intent, rather than externally influenced movement.

Although the two publications offer differing definitions of "actionness", both generally agree that "action" is intentional movement. Additionally, neither paper is dedicated to defining "action". In Chen et al., the concepts of "action" and "actionness" are discussed briefly within a dedicated section. In Luo et al., the discussion appears mainly in the introduction, with a brief mention in the related works section.

4.2 Repository Overview

The datasets examined in this research are collections of video recordings of human subjects, designed for training and evaluating action recognition models. This section provides an overview of the types of datasets included in the repository, supported by specific examples to illustrate their nature and scope. Subsequent sections will focus on how these datasets are represented in the extracted publications, rather than analyzing the dataset contents directly.

There are 967 unique datasets organized into the repository. This section draws only from a subset referred to as *introduced datasets*. To reiterate, these are datasets first presented or made publicly available in the publications examined from CVPR, ICCV, and ECCV. A dataset is considered "introduced" if the authors explicitly describe it as a novel contribution or indicate that it has been made publicly available for the first time in their paper. In total, 294 such datasets were identified. A breakdown by conference can be found in Table 4.1.

Table 4.1: Introduced Datasets by Conference

Conference	**Number of Introduced Datasets**
CVPR	130
ICCV	74
ECCV	90
Total	**294**

To help illustrate the nature and range of these datasets, Table 4.2 presents a sample of introduced datasets. The table includes how the datasets' introducing papers describe the dataset, as well as video stills of the data. These examples are intended to give a clearer sense of what kinds of human activities and recording settings are commonly represented in action recognition research.

To organize the dataset characteristics in a clear and descriptive manner, I developed a taxonomic framework comprised of four main categories: (1) the point of view (POV) in which the video is recorded, (2) the sources of the videos, (3) the content of the videos, and (4) the metadata of the datasets. Figure 4.1 illustrates how these categories further divide into subcategories. It is important to note that no category holds precedence over the others, and the subcategories are not mutually exclusive—multiple subcategories may apply to a single dataset. Each of the categories and subcategories will be described in depth in this section.

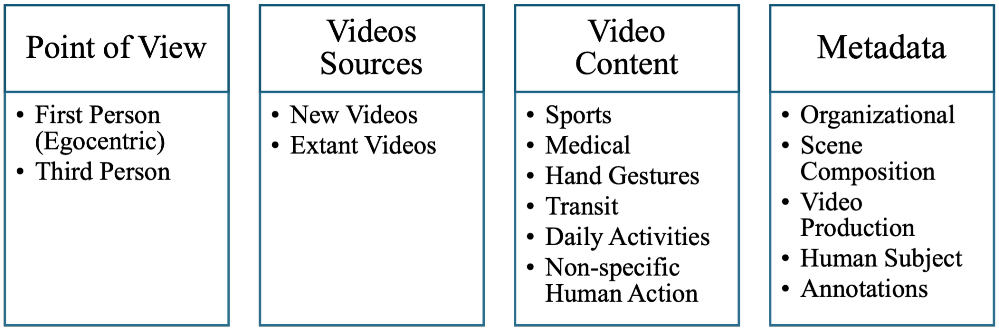


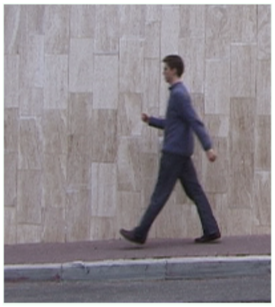
Figure 4.1: Dataset Taxonomy

4.2.1 Point of View

The first category, point of view (POV), refers to the perspective from which the video data is recorded in relation to the human subject present in the video. There are two video POVs in these datasets, third person and first person. The most common POV is third person, where an external camera captures

[1] In the context of this dataset, "abnormal events" can range from someone riding a bicycle or pushing a stroller to chasing or brawling. A "normal event" would be someone walking.

Table 4.2: Examples of Introduced Datasets

Dataset	**Description**	**Video Example**
Weizmann [26]	A database of 90 video sequences featuring nine individuals, each performing 10 "natural actions" such as walking and running, for use in action classification and clustering experiments.	
HMDB51 [39]	A dataset of non-specific human action with 51 action categories, each containing at least 101 clips, totaling 6,766 video clips sourced from diverse origins.	
MultiSports [45]	A sports dataset featuring 66 action categories across four sports—American football, basketball, volleyball, and aerobic gymnastics—comprising YouTube videos from 247 competitions.	
ShanghaiTech Campus [49]	A dataset of 3 scenes with varied lighting and camera angles, includes 130 *abnormal events*[1], over 270,000 training frames, and has pixel-level annotations for abnormalities.	

the subject's actions. The Weizmann dataset [26] is an example of a dataset that uses this POV is used in all of its videos. Another common POV is first person, or egocentric, in which a camera is mounted on a participant to record actions from their perspective. A notable example of this is the Something-Something V1 dataset [27].

4.2.2 Videos Sources

The second category, video source, refers to the origins of the video data in the datasets and can be divided into two subcategories: new videos and extant videos. New videos refers to videos created specifically for a dataset and can be either researcher-recorded or participant-recorded. In researcher-recorded datasets, researchers set up cameras and capture participant actions, as in the Weizmann dataset [26]. Meanwhile, participant-recorded datasets rely on individuals to record their own actions, often resulting in egocentric datasets. For example, the GTEA Gaze [21] and GTEA Gaze+ [21] datasets use smart glasses worn by participants to capture their perspective. However there are exceptions, like the Charades dataset [73] where participants recorded themselves using a stationary camera rather than a wearable device.

The second subcategory, extant videos, refers to datasets composed of previously existing video content from external sources. This category can be further divided into public extant videos and private extant videos. Datasets with public extant videos include content from publicly accessible platforms, such as YouTube (e.g., UCF YouTube [48]), movies (e.g., UCF Films [64]), broadcast television (e.g., UCF Sports [64], which includes footage from ESPN and BBC), and video games (e.g., GTA Combat [86], which features combat footage from the game Grand Theft Auto V). Some datasets, like HMDB51 [39], incorporate public extant videos from multiple sources. Others are created by modifying or re-purposing existing datasets, like J-HMDB [33], which is derived from HMDB51. On the other hand, datasets with private extant videos consist of footage from non-public sources, such as surveillance cameras. An example of this is the ShanghaiTech Campus dataset [49], which includes video data of humans collected from surveillance footage.

4.2.3 Video Content

The third category in the taxonomy is the content of the video data. The most common are sports datasets, like UCF Sports [64], which contains videos of various athletic actions, including diving, golf swinging, and kicking. Some sports datasets specialize in a single sport, such as Diving48 [44], which exclusively includes footage of diving performances. Similarly, other datasets focus on hand gestures, which are often created for human-computer interfaces. The NVIDIA Dynamic Hand Gestures dataset [58], for instance, includes recordings of hand gestures performed within an indoor car driving simulation. A subset of hand gesture datasets includes sign language datasets, such as BSL-1K [2], which contains videos of British Sign Language signing from publicly broadcasted TV programs.

Beyond action recognition, some datasets are designed for medical research and clinical applications. For example, the Alzheimer's Disease Patients (ADP) dataset [81] captures the facial actions—such as smiling, talking, singing—of individuals diagnosed with Alzheimer's disease. [81] then evaluates the performance of existing methods for assessing facial dynamics using their dataset. Similarly, the i-Walk dataset [10] includes action videos of both healthy individuals and patients with various physical and/or cognitive impairments, supporting research on action recognition for mobility assessment and rehabilitation.

Some datasets compile unexpected or erroneous actions, ranging from crimes to someone tripping and falling. Papers introducing such datasets often refer to this class of actions as anomalies, or anomalous actions [49, 76], thus they will be referred to here as anomaly datasets. An example is the Oops! dataset [20], which consists of fail compilation videos from YouTube, capturing mistakes such as falls or failed attempts at action. A subset of anomalous action datasets includes violence detection datasets, such as XD-Violence [85], which contains both violent and non-violent actions collected from movies and YouTube. Another example is UCF Crime [76], which contains surveillance footage of "13 real-world anomalies, including Abuse, Arrest, Arson, Assault, Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism".

Crowd datasets, which typically feature videos of pedestrians and traffic, are often related to anomaly datasets and frequently overlap with violence detection datasets. An example is the ShanghaiTech Campus dataset [49], which captures pedestrian activity, but focuses recognizing abnormal events, such as chasing and brawling. Another crowd dataset, WWW Crowd [71], includes violent actions like fighting but primarily aims to describe the crowd scenario, like what is happening, who is present, and where the events are taking place.

Daily activity datasets contain recordings of individuals performing routine, everyday tasks. An example is the Toyota SmartHome dataset [17], which captures senior citizens as they go about their daily routines, performing actions like drinking, sitting down, standing up, eating and cooking. While the Toyota SmartHome dataset is recorded in a third-person point of view, most of these datasets, like the EPIC-Kitchens dataset [16], are egocentric and often focus specifically on kitchen activities, such as cooking, food preparation, and washing dishes.

Finally, the most generic datasets are those described by researchers as containing human actions, where the specific actions included are determined at the discretion of the researchers with no further reasoning as to why. For example, in the construction of the HMDB51 dataset [39], researchers tasked students with collecting videos from various internet sources that depicted human actions. While guidelines were provided — such as requiring at least one discernible action per clip and ensuring each clip was at least one second long — the selection of specific actions was largely left up to the students.

4.2.4 Metadata

The final category of the taxonomy is the metadata of datasets. There are five types: organizational, scene composition, video production, human subject, and annotation metadata. Please refer to Figure 4.2 for the breakdown.

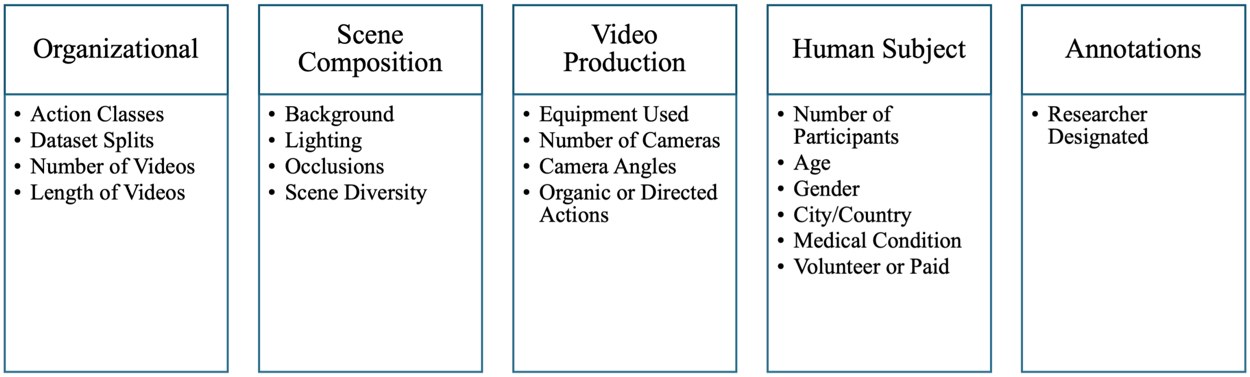


Figure 4.2: Types of Metadata

Organizational Metadata

Organizational metadata refers to how a dataset is structured. The most common attribute of this metadata are action classes, which categorize videos based on the specific actions they depict. These classes indicate both what kinds of actions are included in the dataset and how many there are. For instance, the paper introducing the Concurrent Actions dataset [82] reports that it contains 12 action classes, including: "drink, make a call, turn on monitor, type on keyboard, fetch water, pour water, press button, pick up trash, throw trash, bend down, sit, and stand." Though the term "action classes" is used by most publications, some use synonymous phrases like "action categories" [13].

Aside from sorting data by action class, datasets can also be organized into splits for training, testing, and sometimes validation. Some datasets, like the Cross-Modal Infrared-Visible dataset [80], use a two-way split where 75% of the paired videos for training and the remaining 25% for testing. Others introduce a third validation split, such as the Holistic Video Understanding (HVU) dataset [19], which includes 481,000 videos for training, 31,000 for validation, and 65,000 for testing.

The HVU dataset also presents a common metric that falls under organizational metadata, which is the number of videos in a dataset, or in this case the number of videos by split. It also common for datasets to provide the number of videos per class, like the BIT-Interaction dataset [38], or simply stating the total number of videos, as seen in the Composable Activities dataset [46]. Another commonly reported metric is the length of the videos in the dataset. Similar to how the number of videos in a dataset is reported, the length can be presented in a cumulative figure, or as the video length per class or split, or as the average duration for a video. For example, the BBDB dataset's [72] reports to having 4,254 hours of baseball game footage.

Scene Composition Metadata

Scene composition metadata describes the visual elements within the videos. This typically consists of attributes like background, occlusions, and lighting. Publications describe datasets that exhibit variations in these attributes across different videos as featuring diverse scenes, or "scene diversity" [20, 49, 71].

The background in a video is usually described in two ways. The first is by filming location, often categorized broadly as indoor or outdoor settings [23, 73, 90]. Some datasets provide more specific information—for example the i-Walk dataset [10] specifies that participants were recorded in a rehabilitation center. Similarly, the EPIC-Kitchens dataset [16] features videos recorded in the participants' own kitchens. The second method of describing background is by identifying the visual elements that could impact model performance, like clutter or movement [47].

Like background clutter or movements, occlusions impact model performance by making it difficult to recognize the performed action. This is because occlusions are anything that at least partially obscures the human subjects in the video. However, most publications do not describe the elements that create occlusion, instead reporting that occlusions are present and that they make the dataset difficult. The authors of the MultiSports dataset [45], for example, cite the "occlusion of actions in sports" as one of the reasons why their dataset is difficult.

Another frequently reported attribute of scene composition metadata is lighting. Specifically, authors commonly note if there are changes in illumination [38, 61, 81], because they can make it challenging for models to recognize actions in videos [81]. For similar reasons, publications also report if videos have complex light conditions like [49]. Some publications provide further information on how the lighting affects the scene composition, such as the LSE_eSaude_UVIGO dataset [79], where the videos were recorded in a studio environment that has uniform illumination and no shadow effects.

Video Production Metadata

Video production metadata pertains to both the filming process and the equipment used to record the videos. When discussing the filming process, the focus is typically on how actions were recorded. Many researchers instruct participants to perform specific actions, like the $ACT4^2$ [13] and Weizmann [26] datasets, where every participant performed a set number of actions a certain number of times. On the other hand, other datasets capture natural, unscripted behavior [42, 49]. Some datasets, such as Charades [73], have a hybrid approach, where participants are given scripts to follow but are allowed creative freedom in how they perform the actions.

Meanwhile, equipment details primarily concern the cameras used during video collection. The vast majority of the datasets in the repository are recorded using color cameras. Notable exceptions are synthetic datasets, such as the GTA Combat dataset [86], which is not captured with physical cameras but instead features color recordings of in-game actions from Grand Theft Auto V. Though some publications like [47] only mention the use of color camera, most include information on the camera model [8, 13, 21, 22, 61, 70, 84]. Microsoft Kinects are some of the most commonly used cameras [13, 70, 84]. For egocentric datasets, GoPro cameras and wearable devices like the Tobii eye-tracking glasses are frequently employed

[16, 21, 22]. Publications also often specify the number of cameras used [13, 61, 84], the number of camera angles [13, 61, 70], and the resolution of the collected footage [26, 45, 47, 61, 84].

Human Subject Metadata

Human subject metadata refers to information about the individuals featured in the video data. Among the five types of metadata, it is the least frequently reported. Datasets typically include two kinds of human subjects: ambiguous subjects and participants. Ambiguous subjects are individuals whose consent is neither explicitly documented nor can it be reasonably inferred. Their likenesses are typically sourced from public extant footage—either public videos available online [2, 3, 20, 39, 44, 64, 71, 85] or private recordings like surveillance footage [49, 76]. Participants, in contrast, are usually volunteers or paid actors who have explicitly consented to the use of their likeness in the dataset [10, 17, 27, 63, 73]. While many publications do not detail the exact nature of participation, it is often reasonable to assume consent when individuals are actively involved in the recording process [13, 21, 22, 26, 84].

Publications that include participants often report how the number of individuals featured in the recordings [8, 13, 17, 21, 22, 84]. Some also provide additional demographic attributes, such as participants' geographic locations [10, 16, 73], gender [10, 81] and age range [17, 70]. Gender is typically reported using a binary male/female classification [10, 81]. Only a few datasets mention diversity in body shape and ethnicity [8, 70], though none provide explicit statistical breakdowns.

Annotation Metadata

The final category of metadata is annotations, which refers to any information that the authors explicitly classify as annotations. Many datasets do not include annotations and therefore lack annotation metadata. Among those that do, the level of detail varies significantly — ranging from a simple mention of their existence to in-depth descriptions of the annotation content and process. For example, the ShanghaiTech Campus dataset [49] states that "pixel-level ground truth of abnormal events is also annotated in our dataset." Through accompanying figures and their descriptions, it is deduced that these annotations are represented by green dots marking regions where abnormal events occur. At the other end of the spectrum, the MultiSports dataset [45] offers a detailed explanation of its two-stage annotation process. In the first stage, professional athletes create annotations like bounding boxes, while in the second stage, a group of crowd-sourced annotators refines those annotations for accuracy. This publication, along with the authors of the COIN dataset [77], are some of the few that describe the annotation interface they use and provide images of what the interface looks like.

Other than bounding boxes, videos can also be annotated for the start and end frame of each action, which is also present in the MultiSports dataset. Another frequent form is joint annotation, where body joints are labeled. The UAV-GESTURE dataset [63], includes 13 annotated joints across 37,151 frames, covering the ankles, knees, hips, wrists, elbows, shoulders, and head. Other datasets, like WebVid-2M [3], have textual description annotations for every video in the dataset. Similarly, the WWW Crowd dataset

[71] includes semantic annotations for each video that address the following questions: “Where is the crowd?”, “Who is in the crowd?”, and “Why is crowd here?”.

4.3 Introduction of New Datasets

When introducing new datasets, authors often justify their creation by claiming that existing datasets lack the features that they deem necessary for the advancement of the field of action recognition. The ACT4^2 dataset [13], for instance, is introduced to “break the bottleneck of existing action recognition approaches” by providing “an infrastructure for investigating both color and depth information in human action analysis and handling action variations over viewpoints”. Similarly, the Charades dataset [73] was compiled in response to the perceived absence of “boring videos of daily activities” that could support practical applications like helping people find lost items. In contrast, some researchers introduce new datasets without explaining their necessity, often stating — explicitly or implicitly — that the dataset was created to evaluate their proposed methodology. This is frequently done without clarifying why existing datasets are inadequate for the researchers’ purposes. For example, Lin, Jiang, and Davis [47] evaluate their action recognition method using the Weizmann [26] and KTH [69] datasets, in addition to their newly introduced Military Gesture dataset. However, they do not explain why this new dataset was created, despite presenting it as one of the four contributions of their work and using two existing datasets for evaluation. Instead, they describe their Military Gesture dataset as challenging, which is a frequent descriptor of introduced datasets.

The term “challenging” is often treated as a desirable quality of datasets because it is associated with better applicability to the real world and is often used as a synonym for real-life data or real-life scenarios. Although its definition varies across publications, the four most common metadata attributes that authors argue contribute to the “challenging” nature of the dataset are: lighting [81], background [47], occlusions [45, 81], and scene diversity [49, 73]. All of these attributes have a shared trait of introducing noise that can hinder a model’s ability to recognize action. For example, the authors of the Military Gestures dataset [47] consider it to be challenging because its test videos have dynamic backgrounds, where a person “(regarded as ‘noise’) moved continuously behind the actor, making recognition more challenging”. Similarly, the creators of the ADP dataset [81] deliberately included videos with lighting variations and partial occlusions as they consider these factors to be typical challenges in facial behavior analysis from naturalistic videos.

In addition to “challenging”, introduced datasets are frequently marketed as large and in some way the first of its kind to distinguish them from the existing abundance of datasets. The UCF Crime dataset [76], for example, claims to be “the first of its kind” and that it is “by far the largest dataset with more than 25 times [as many] videos than [the] existing largest anomaly dataset”. Similarly, the EPIC-Kitchens dataset [16] claims to be “the largest first-person dataset to date”, while the ShanghaiTech Campus dataset [49] claims to be “even larger than the summation of all existing dataset for anomaly detection in terms of both the volume of data and the diversity of scenes”. This is often coupled with describing the contents of the dataset in terms of its metadata. Usually referred to as “Dataset Statistics”, authors often provide some combination of the 5 kinds of metadata, though they most often report organizational metadata [45, 77]

Another feature of dataset introductions is the attention given to how datasets are constructed [45, 73]. The authors of the MultiSports dataset explain their rationale for selecting certain sports, which in turn contributed to the choice of actions to include. They also describe the annotation process, noting that both athletes and crowd-sourced annotators were employed to ensure high quality annotations. However, some publications offer minimal information about on their datasets. A notable example is the ShanghaiTech Campus dataset [49], which is described in three sentences:

> "Our new proposed dataset has 13 scenes with complex light conditions and camera angles. It contains 130 abnormal events and over 270, 000 training frames. Moreover, pixel level ground truth of abnormal events is also annotated in our dataset."—**ShanghaiTech Campus**

One aspect that is frequently under-reported is information about the human subjects in videos. Most datasets that are composed of extant videos do not address the individuals featured in the footage. If they do, it is often in assurance of the quality of their dataset. An example of this is the MultiSports dataset [45], where they claim that their dataset is less biased and well balanced for realistic sports analysis because they selected videos from competitions across various performance levels, countries, and genders. Most datasets with new videos do not explain how participants are recruited [13, 26], instead opting to report how many people participated in the recording of the dataset. There are a few exceptions to this, like Something-Something V1 [27] and Charades [73], that explain participants were paid actors recruited through crowd-sourcing platforms like Amazon Mechanical Turks. If additional participant information is provided, it is usually in terms of human subject metadata as discussed in Section 4.2.4, which is most commonly found in medical datasets like ADP [81] and i-Walk [10]. The HuMMan dataset [8] is unique in that it is one of the few to engage with the appearance of participants, stating that the dataset "consists of 1000 subjects with a wide coverage of genders, ages, body shapes (heights, weights), and ethnicity". However, it does not elaborate further, instead referring readers to the participant statistics provided in the supplementary material.

Overall, dataset introductions tend to emphasize attributes such as being "challenging," large-scale, or novel, while details about dataset construction and human subjects vary considerably, with some datasets providing extensive information and others offering only minimal description.

4.4 Description of Used Datasets

The vast majority of the extracted publications, including those that introduce their own datasets, rely on existing action recognition datasets containing videos of humans. A few exceptions use only the datasets they introduce [34, 43, 80, 90]. Throughout this paper, I refer to datasets that are used but not introduced by the authors as *used datasets*. As stated previously, and introduced dataset can also be a used dataset if it is introduced by one publication and used by others within the corpus.

Most publications provide brief descriptions of these used datasets, typically focusing on metadata attributes. For example, Zhao & Wildes [89] describe the three datasets they use—Breakfast [40], 50 Salads

[75], and EPIC-Kitchens [16]—in terms of the number of videos, actors, and action classes. Like many other papers [60, 88], they also briefly explains how the datasets were processed for their experiments. Some publications offer more detailed descriptions, though these still tend to focus on metadata. For example, [88] describes the Weizmann dataset [26] in terms of its background settings and action classes, listing each class to specify the types of actions captured, as well as detailing how many actions are performed in each video and by how many individuals. Conversely, some publications only name the datasets they use and proceed directly to describing how they were used, without providing any description of the dataset [31, 56, 83, 87]. For example, [56] includes a subsection meant for describing the three datasets it uses—KTH [69], UCF Sports [64], and Hollywood2 [53]—but it instead briefly explains how each was used in the study.

Among the publications that describe the used datasets, few engage with the human subjects within the datasets. When they do, it is in terms of human subject metadata like the number of subjects, age, and gender within a male/female binary [7, 14, 36, 37, 51, 89]. Publications like [14] are among the rare few that discuss diversity among subjects. [14] in particular notes that the DISFA dataset [54] "relative ethnic diversity", but—like the publication that introduces the HuMMan dataset [28]—this is not elaborated upon.

Table 4.3: Most Used Datasets by Conference.

	Conferences			
Dataset	**CVPR**	**ICCV**	**ECCV**	**Total**
UCF101	162	80	82	324
HMDB51	123	59	59	241
Kinetics 400	116	49	50	215
NTU RGB+D 60	46	30	33	109
KTH	52	22	31	105

While tracking the datasets identified through the extracted publications, five were found to be the most commonly used across CVPR, ICCV, and ECCV. Ranked by the total number of times were used in all three conferences, these datasets are: UCF101 [74], HMDB51 [39], Kinetics 400 [35], NTU RGB+D 60 [70], and KTH [69]. Table 4.3 provides a breakdown for how frequently each dataset was used in CVPR, ICCV, and ECCV. Two of these datasets—HMDB51 and NTU RGB+D 60—are also introduced datasets. HMDB51 is from ICCV 2011 and NTU RGB+D 60 is from CVPR 2016. The oldest of the 5 is the KTH dataset, which was published in 2004. The newest is Kinetics 400, which is was published in 2017. Finally, the most used dataset, UCF101, was published in 2012. Most of these datasets have a wide range of human actions. For example, UCF101 actions like "apply eye makeup", "playing dhol", and "baby crawling". Three of the datasets—HMDB51, Kinetics 400, and NTU RGB+D 60—include unintentional actions like falling, staggering, and sneezing.

Chapter 5

DISCUSSION

5.1 What is "Action" in "Action Recognition"?

5.1.1 Attempts to Explicitly Define "Action"

Chen et al. [12] and Luo et al. [50] are the only publications identified in this research that attempt to explicitly define "action". Both adhere to the definition "action" as proposed by philosopher Davidson [18], which can be summarized as intentional bodily movement. However, the inclusion of "intent" in this definition limits its applicability to action recognition and computer vision as a whole. By requiring actions to be intentional, these definitions exclude a range of involuntary movements that action recognition systems are often designed to detect, such as seizures, tics, or falls. For instance, tripping and falling are unintentional movements, yet detecting them is crucial in safety-critical contexts like pedestrian recognition for autonomous vehicles. In a scenario where someone trips and falls in front of a self-driving car, the system's ability to recognize and respond appropriately could be the difference between life and death. Similarly, in clinical environments, recognizing involuntary actions such as seizures is essential for ensuring patient safety. A system incapable of identifying such events risks failing to provide timely alerts, which can lead to severe health consequences, including death. Clearly, a definition of "action" that excludes unintentional or involuntary movements is too narrow to meet the practical demands of computer vision. The requirement of intent as a defining feature is overly restrictive. While philosophical contributions from the philosophy of action offer valuable insights, they alone do not provide a sufficiently comprehensive framework for the full range of actions that computer vision systems must be capable of recognizing.

Additionally, for both papers, defining "action" and "actionness" was one of their purposes, not the main goal of the paper. Though they treat these definitions as some of their contributions, as seen in their abstracts, they use the definitions as a means to set the foundations for their novel methodologies. This likely explains why both papers were selective in the sources they use to support their definitions of "action" and "actionness". Chen et al. and Luo et al. base their definition of "action" on Donald Davidson's 1963 philosophical work *Actions, Reasons, and Causes* [18], where he argues that providing a reason for an action gives it a causal explanation. To define "actionness", Luo et al. additionally draw from Gao and Scholl's 2011 psychology paper *Chasing vs. Stalking: Interrupting the Perception of Animacy* [25], which examines how participants distinguish between a target that is actively chasing their character and other distractors moving in similar ways. While these sources offer an initial foundation for defining "action" in the context of action recognition, a more robust and comprehensive definition would draw from multiple and interdisciplinary sources from relevant fields—like philosophy of action and psychology—while remaining practical for application in computer vision. Ideally, such a definition would be inherently compatible with action recognition without requiring an auxiliary concept like "actionness" to make it applicable.

5.1.2 The Need for a Standard Definition of Action

As I analyzed the corpus during the course of this research, I noticed that action recognition researchers often operate on implicit and subjective definitions of "action". The absence of a clear, shared definition has led to inconsistencies in how actions are represented across datasets. For example, the authors of the Weizmann dataset [26] include actions like "run" and "walk". Meanwhile the MultiSports dataset has the action "basketball dribble" [45]. Though MultiSports treats this as one action, dribbling in basketball consists of multiple actions, like running or walking while simultaneously bouncing a ball. In this case, "basketball dribble" is considered an action of its own, while by other definitions, like Weizmann's, it would be broken down into the multiple actions that compose it.

These inconsistencies also arise within datasets themselves. For instance, in the HMDB51 dataset, the authors aimed to create a dataset that consists of everyday human actions by asking students to collect any single action videos clips from the internet and movies. Consequently, the dataset consists of everyday actions like "hand-waving", "chewing", "talking", and "drinking"—but also includes actions like "sword fighting", "draw sword", "shoot bow", and "shoot gun". What began as a dataset intended to represent "everyday actions" became one that also includes actions that are typically far removed from daily life. This is because "action", specifically "everyday actions", was not explicitly defined, and there was a reliance on an implicit definition that was subjective to each participating student.

These examples, along with the treatment of "action" in other publications, suggest that "action" is generally assumed to refer to any movement or behavior that researchers expect a system to recognize and interpret. Establishing a standardized definition of "action" within the field of action recognition—and in computer vision more broadly—would promote consistency in how actions are represented in datasets. It would also encourage researchers to be more explicit about their criteria for including specific actions.

The absence of a standardized definition of "action" has impacted this research as well. When constructing the dataset repository, I included any video dataset featuring humans, since there is no universally accepted definition of what constitutes an "action." As a result, the criteria for what qualifies as action recognition—and, more specifically, what counts as an action recognition dataset—remain largely subjective. Nonetheless, I felt justified in including these datasets because they were identified in papers that were retrieved from CVPR, ICCV, and ECCV using the search term "action recognition". Furthermore, of the 754 papers extracted from CVPR, "action recognition" appeared in approximately 25% of titles and 47% of abstracts. At ICCV, the term appeared in about 24% of titles and 46% of abstracts among the 369 papers extracted. For the 752 ECCV papers, "action recognition" was found in around 12% of titles; however, abstract data was unavailable due to limitations in the extracted information. It is also important to note that not all extracted papers were analyzed, only the ones that used datasets containing video data of human. Referring to Table 3.1, the number of papers analyzed per conference are as follows: 424 from CVPR, 334 from ICCV, and 562 from ECCV.

5.1.3 What is "Action" According to Action Recognition Research?

Chen et al. and Luo et al. emphasize the inclusion of "intent" in the definition of "action", as it distinguishes action from movement. This raises critical questions for computer vision: Does it matter to an algorithm whether a person in a video is moving intentionally? Is the use of "intent" a human-centric way of thinking, or is it a concept machines must also grasp to effectively recognize actions?

As Alex Taylor explores in *Machine Intelligence*, humans and machines excel at different tasks and exhibit "intelligence" in different ways [78]. Applying a definition of action rooted in the philosophy of action—which seeks to understand what constitutes an action from a human perspective—may not translate directly to machine perception. While such philosophical frameworks prioritize internal states like intention, machine learning models typically rely on external, observable patterns in data.

For instance, humans often predict others' actions by drawing on both observed behavior and understanding their intent. In contrast, algorithms tasked with action prediction rely solely on data from past actions. One might argue that machines can approximate intent by identifying statistical regularities or contextual cues in the data. Some might also say that this is fundamentally different from the human understanding of intention, as the machine does not understand intent in a cognitive or phenomenological sense—it merely models patterns that correlate with it. Still, some may say that this process mirrors the internal one in humans to understand "intent", so can it really be said that machines are not capable of understanding intent? At this point, it is worth pausing to consider if this conversation is even necessary for the purposes of action recognition.

As previously discussed, defining "action" in terms of intent risks excluding some of the field's most high stakes applications, such as pedestrian detection in self-driving cars or seizure recognition in medical monitoring systems. In these scenarios, the focus lies on observable outcomes rather than internal motivations. Based on my research and analysis, the implicit definitions used by action recognition researchers, and the real-world applications of action recognition, I propose the following definition of "action" for this field: ***observable human movement that conveys meaningful behavior in a given context***. *Observable* is a necessary part of the definition because the behavior needs to be visible for computers to recognize it. Meditating, for instance, can be considered to be an action to most people, but visually it is indistinguishable from the act of "sitting". Therefore, within the context of action recognition for computer vision, an action must be observable. The action must also be *meaningful* within its context, because the significance of the action can change depending on the situation. For example, raising a hand might mean asking a question in a classroom or signaling to stop traffic.

Since action recognition in computer vision operates on visual data, emphasizing observable and contextually meaningful behavior offers a more practical and effective framework. The definition I propose aligns with the field's applications because it is intentionally designed for machine processes, focusing on what algorithms can reliably detect and interpret.

5.2 How Datasets Are Valued

5.2.1 Model Work Over Data Work

Although models are highly dependent on data—performing only as well as the data they are trained on—data work is frequently undervalued [65, 66]. This is evident in how the Kinetics 400 dataset is introduced and described by its authors. The dataset was initially introduced in a 2017 arXiv preprint [35], and later that year formally presented in the CVPR paper *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset* [9]. In the CVPR paper, the dataset is described in one section consisting of three paragraphs through organizational metadata—a practice more typical of papers that only use datasets rather than introduce them. Other than describing the data, which was collected from YouTube, as challenging in the abstract, *Quo Vadis* refrains from marketing its dataset the way introducing papers typically do, as is described in Section 4.3. The authors refer readers to the preprint if they want more information on the dataset. This aligns with observations by Scheuerman et al 2021 [66], who note that datasets are "rarely unaccompanied by a methodological innovation" despite authors considering them as crucial for scientific progress.

The calls for papers at top computer vision conferences—CVPR, ICCV, and ECCV—only recently began explicitly recognizing "Datasets and Evaluation" as a formal area of interest. CVPR introduced this category in 2019[1], followed by ECCV in 2020[2], and ICCV in 2021[3]. Prior to these changes, none of these conferences listed datasets as a standalone topic, making it unlikely for papers focused solely on data work to be accepted. This context likely explains the manner in which Kinetics 400 was introduced in 2017: the arXiv preprint focused exclusively on the dataset, while the subsequent CVPR paper offered a secondary introduction that both broadened its visibility and conferred the prestige of being presented at a leading conference.

Though top computer vision conferences now encourage contributions centered on data work, established norms around how datasets are introduced and discussed continue to shape current practices. The way new datasets are presented today often reflects these earlier precedents. For example, *LAEO-Net: Revisiting People Looking at Each Other in Videos* by Marín-Jiménez et al. [52], was published in CVPR 2019 and introduces two new datasets, UCO-LAEO and AVA-LAEO. In a dedicated section titled "Datasets", they are briefly described by their content, some metadata, and what the data was annotated for, alongside the other two datasets that the authors use. The rest of the paper focuses on their novel method. Simi-

[1] `https://cvpr2019.thecvf.com/files/CFP_CVPR2019.pdf`

[2] `https://eccv2020.eu/submission/`

[3] `https://iccv2021.thecvf.com/node/4#call-for-papers`

larly, the CVPR 2022 paper titled *Programmatic Concept Learning for Human Motion Description and Synthesis* introduces the MotiCon dataset [41] within a single-paragraph subsection by its content and metadata, while the remainder of the paper is dedicated to their novel methodology. Despite conferences adding "Datasets and Evaluation" as an area of interest, publications that have been released so far display little to no change in how they introduce and describe datasets. Model work continues to be included, with papers often focusing on the model work at the expense of data work. This is likely due to concerns that papers will not be published without inclusion of model work, as historically "the vast majority of these datasets do not get published unless they report some kind of algorithmic improvement" [66].

Furthermore, much of the language utilized when describing datasets remains the same, with authors relying on metadata to describe the datasets they introduce this. Metadata offers a way to quickly classify the aspects of datasets so they are easy to describe. However, for publications introducing new datasets , this approach overlooks the contextuality of the data. As defined by Scheuerman et al. [66], contextuality is how "circumstances such as time, location, or use shape the world and thus the data in a dataset". Providing this context would involve, for example, addressing why a new dataset was created despite existing similar ones (other than furthering the field of action recognition), and articulating how the data was thoughtfully and intentionally curated. This includes explaining how human subjects were selected and considering how their demographics—like age, gender, ethnicity, ableness, and geographic origin—might influence how actions are represented. For example, expressions of emotion, particularly facial expressions, are not universal and often significantly varies by culture [29]. This extends to how actions are performed, with people from different regions and different cultures performing actions, like counting [32]. When authors discuss what makes their datasets "challenging", they often focus on the metadata attributes that make it so (lighting, background, occlusions, and scene diversity). They could also examine how these attributes affect the representation of the human subjects within the data, and how those effects may vary based on factors like their skin tone and physical impairments.

Additionally, publications that describe the datasets they use often rely heavily on metadata to provide brief summaries, possibly due to word count or page limitations. However, this practice can obscure important contextual information. Researchers should consider engaging more deeply with their chosen datasets—for example, by explaining why a particular dataset was selected, what populations it represents, and how these factors might influence the performance of their methods in real-world settings. Such engagement would enhance transparency around the dataset's life-cycle and help mitigate the risk of data cascades [65].

5.2.2 The Marketing of New Datasets

Publications introducing new datasets often frame their contributions with broad, promotional language, frequently using terms like "challenging," "largest," or "first-of-its-kind". Authors appear to value these descriptors for two reasons: 1) there is a tendency in the field of action recognition, and more broadly computer vision, to prefer masculine discourse, and 2) they help distinguish their novel dataset from the existing abundance of datasets while indicating its applicability to the real-world.

Masculine Discourse

The emphasis on novelty, scale, and difficulty in dataset introductions reflects a broader tendency in computer science for masculine discourse, which prefers "hard, technical, work over work done for the sake of helping others" [30]. As discussed by Hermans & Schlesinger [30], masculine discourse privilege masculine-coded values such as technical rigor, formalism, and quantitative analysis, while undervaluing feminine-coded practices like user studies, qualitative methods, and work oriented toward social benefit. To be clear, emphasizing technical rigor is not inherently problematic, but it becomes limiting when it is consistently elevated at the expense of other perspectives. This dynamic is evident in how authors justify the creation of new datasets: they often highlight technical limitations in existing datasets or claim that their contributions advance the field of action recognition. Describing a dataset as "challenging," "large-scale," or "first-of-its-kind" positions it as a meaningful technical accomplishment, reinforcing dominant values around innovation and difficulty rather than accessibility or utility. It is worth expanding these narratives to also include how datasets might benefit people, support social good, or address real-world needs—complementing technical goals rather than displacing them.

What often gets left out when only technical values are foregrounded are the contextual factors that shape how datasets are constructed. Elements such as time, place, intended application, and the people represented in the data are often omitted or downplayed. As a result, datasets are frequently presented as neutral, universal resources—even though they are products of social and political processes [59, 66]. Additionally, the context of the researcher—such as their disciplinary background, institutional goals, or assumptions about what counts as an "action"—shapes how data is collected, labeled, and structured. Meanwhile, the context of the subject—their lived experience, cultural background, or personal intentions—is frequently ignored or flattened. When classification schemes are imposed without reflexivity, they can trap subjects within categories that do not reflect the complexity of their actions or identities. Failing to explain seemingly "arbitrary" decisions—such as which human subjects or actions to include, or how labels are defined—gives the impression that they required little thought, reinforcing an illusion of objectivity. Integrating feminine-coded values like reflexivity into dataset creation encourages researchers to examine the assumptions underlying their work, leading to more thoughtful and accountable research practices.

Real World Applicability

Given the abundance of available datasets, researchers often frame new releases as valuable contributions by emphasizing their distinctiveness. Common descriptors include terms like "challenging," "largest," and "first-of-its-kind," which serve to convey significance and justify a dataset's relevance. In defining what makes a dataset "challenging," researchers frequently point to metadata attributes like lighting, backgrounds, occlusions, and scene diversity. However, these characteristics are likely not unique to action recognition; Scheuerman et al. [66] have identified similar language patterns across computer vision research more broadly.

The emphasis on "challenging" data is rooted in a longer history of the field's evolving relationship with visual complexity. In the 1950's and 1960's, artificial intelligence researchers believed vision would be relatively easy to solve—an optimism encapsulated by the 1966 *The Summer Vision Project* which tasked undergraduates with building a complete visual recognition system over a summer [62]. Nearly six decades later, the persistence of vision as an unsolved problem illustrates how deceptively difficult this task has proven to be. Reiterating Alex Taylor [78], humans and machines excel at different tasks; recognizing a dog in a photo may be effortless for a human, but for a machine, even minor variations in lighting, background, or occlusion can disrupt recognition [57]. These challenges have led to a push for large-scale datasets with diverse instances of objects—and, for action recognition, diverse instances of actions—to help models learn to generalize better. As Crawford notes, insufficient variety in training data—such as only using red apples—can lead to brittle systems that fail on slightly different inputs, like green apples [15].

Within this context, "challenging" datasets are positioned as those that attempt to replicate the unpredictability and variation of real-world environments, under the assumption that exposure to such data improves a model's performance outside controlled settings [66]. These datasets are often presented as necessary for advancing not only action recognition but computer vision more broadly. Similarly, large-scale datasets are assumed to offer greater value by better capturing the complexity of real-world phenomena, reinforcing the association between scale, challenge, and realism. Claims of novelty are often tied to dominant conceptions of progress in the field, with "first-of-its-kind" positioning used to signal innovation and academic contribution.

Chapter 6

Future Work

As stated earlier, the repository currently contains comprehensive information for the 294 introduced datasets and the five most frequently used datasets. Future work will involve expanding this information to the remaining 673 datasets by documenting their publishing details, related websites, and content. In addition, because the repository tracks the video content for all datasets, further research could analyze which types of video content are most common in action recognition (e.g., sports), which types are most frequently used, and what implications this concentration may have for the field.

A related direction for future research involves a deeper investigation of the five most frequently used datasets. This could include analyzing their video sources, content, and metadata, as well as examining how they represent human subjects and why they have become preferred resources within the action recognition community.

Additionally, this research focuses specifically on action recognition within the top computer vision conferences. Future work could expand the scope to include broader artificial intelligence conferences such as ICLR and NeurIPS. This could provide insight into how the patterns and findings from this study applies to broader AI research community.

Finally, a standardized definition of "action" should be proposed and adopted for the field of action recognition. As discussed earlier, no widely accepted definition currently exists in the field, and this lack of clarity limits the ability to consistently evaluate and design datasets. While this thesis offers a working definition, there is value in pursuing a dedicated, interdisciplinary effort to define “action”—one that integrates insights from fields such as philosophy, cognitive science, and disability studies, while remaining grounded in the practical needs of action recognition systems. Establishing a definition grounded in a theoretically robust framework would not only bring conceptual clarity but also support the intentional and inclusive representation of diversely abled individuals in datasets. A clear, inclusive definition could guide researchers in recognizing a broader range of human activities and embodiments, ultimately leading to more equitable models and applications.

Chapter 7

Conclusion

This research has examined how datasets are introduced and discussed within the field of human action recognition, revealing patterns that reflect broader cultural tendencies in computer vision. A critical finding is the absence of a standardized definition of "action" within the field. This lack of clarity leads to inconsistent representations of action both across and within datasets. It has even affected the construction of the repository for this research. Based on observed explicit and implicit definitions of "action" for action within the analyzed paper corpus and the applications of action recognition, I propose defining "action" as observable human movement that conveys meaningful behavior in a given context.

This research also finds an over-reliance on metadata when describing datasets. For example, if human subjects are addressed, they are often abstracted through metadata and described with minimal attention to their demographics, environments, and how factors in data collection affect their representation. Removing them of their context and positioning them as a universal representation of people reinforces inequities in how technologies perform across different populations and contexts.

Additionally, despite the foundational role of data in developing action recognition systems, data work remains undervalued relative to model development. Datasets are often introduced with an emphasis on scale, novelty, and technical challenge, while overlooking contextual and human-centered details such as human subject selection, cultural variation, and the conditions under which data is collected. Addressing this imbalance requires a cultural shift within the computer vision community—one that elevates the value of data work, integrates reflexive practices in data curation, and embraces interdisciplinary and intercultural perspectives on human behavior.

As the field continues to evolve, reexamining assumptions about datasets will be essential to ensuring that technological progress aligns with ethical and social responsibility.

Bibliography

[1] URL: https://portal.core.edu.au/conf-ranks/?search=computer+vision&by=all&source=CORE2023&sort=atitle&page=1 (visited on 03/07/2025).

[2] Samuel Albanie et al. "BSL-1K: Scaling Up Co-articulated Sign Language Recognition Using Mouthing Cues". en. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 35–53. ISBN: 978-3-030-58621-8. DOI: 10.1007/978-3-030-58621-8_3.

[3] Max Bain et al. "Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. ISSN: 2380-7504. Oct. 2021, pp. 1708–1718. DOI: 10.1109/ICCV48922.2021.00175. URL: https://ieeexplore.ieee.org/document/9711165 (visited on 05/06/2025).

[4] Abeba Birhane et al. "The Values Encoded in Machine Learning Research". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. New York, NY, USA: Association for Computing Machinery, June 2022, pp. 173–184. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533083. URL: https://dl.acm.org/doi/10.1145/3531146.3533083 (visited on 06/04/2025).

[5] Virginia Braun and Victoria Clarke. "Using thematic analysis in psychology". In: *Qualitative Research in Psychology* 3.2 (Jan. 2006). Publisher: Routledge, pp. 77–101. ISSN: 1478-0887. DOI: 10.1191/1478088706qp063oa. URL: https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa (visited on 03/07/2025).

[6] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, Feb. 2018, pp. 77–91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html.

[7] Wonmin Byeon et al. "ContextVP: Fully Context-Aware Video Prediction". en. In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Cham: Springer International Publishing, 2018, pp. 781–797. ISBN: 978-3-030-01270-0. DOI: 10.1007/978-3-030-01270-0_46.

[8] Zhongang Cai et al. "HuMMan: Multi-modal 4D Human Dataset for Versatile Sensing and Modeling". en. In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan et al. Cham: Springer Nature Switzerland, 2022, pp. 557–577. ISBN: 978-3-031-20071-7. DOI: 10.1007/978-3-031-20071-7_33.

[9] João Carreira and Andrew Zisserman. "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919. July 2017, pp. 4724–4733. DOI: 10.1109/CVPR.2017.502. URL: https://ieeexplore.ieee.org/document/8099985 (visited on 05/20/2025).

[10] Georgia Chalvatzaki et al. "i-Walk Intelligent Assessment System: Activity, Mobility, Intention, Communication". en. In: *Computer Vision – ECCV 2020 Workshops*. Ed. by Adrien Bartoli and Andrea Fusiello. Cham: Springer International Publishing, 2020, pp. 500–517. ISBN: 978-3-030-66823-5. DOI: 10.1007/978-3-030-66823-5_30.

[11] Cheng Chen and S. Shyam Sundar. "Is this AI trained on Credible Data? The Effects of Labeling Quality and Performance Bias on User Trust". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. event-place: Hamburg, Germany. New York, NY, USA: Association for Computing Machinery, 2023. ISBN: 978-1-4503-9421-5. DOI: 10.1145/3544548.3580805. URL: https://doi.org/10.1145/3544548.3580805.

[12] Wei Chen et al. "Actionness Ranking with Lattice Conditional Ordinal Random Fields". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. ISSN: 1063-6919. June 2014, pp. 748–755. DOI: 10.1109/CVPR.2014.101. URL: https://ieeexplore.ieee.org/document/6909496 (visited on 03/07/2025).

[13] Zhongwei Cheng et al. "Human Daily Action Analysis with Multi-view and Color-Depth Data". en. In: *Computer Vision – ECCV 2012. Workshops and Demonstrations*. Ed. by Andrea Fusiello, Vittorio Murino, and Rita Cucchiara. Berlin, Heidelberg: Springer, 2012, pp. 52–61. ISBN: 978-3-642-33868-7. DOI: 10.1007/978-3-642-33868-7_6.

[14] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. "Deep Structure Inference Network for Facial Action Unit Recognition". en. In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Cham: Springer International Publishing, 2018, pp. 309–324. ISBN: 978-3-030-01258-8. DOI: 10.1007/978-3-030-01258-8_19.

[15] Kate Crawford. *The Atlas of AI : Power, Politics, and the Planetary Costs of Artificial Intelligence*. eng. New Haven: Yale University Press, Jan. 2021. ISBN: 978-0-300-20957-0. URL: https://research.ebsco.com/linkprocessor/plink?id=5bc8c3f7-ec08-3525-8cc4-f10e4a814e3f.

[16] Dima Damen et al. "Scaling Egocentric Vision: The Dataset". en. In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Cham: Springer International Publishing, 2018, pp. 753–771. ISBN: 978-3-030-01225-0. DOI: 10.1007/978-3-030-01225-0_44.

[17] Srijan Das et al. “Toyota Smarthome: Real-World Activities of Daily Living”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. ISSN: 2380-7504. Oct. 2019, pp. 833–842. DOI: 10.1109/ICCV.2019.00092. URL: https://ieeexplore.ieee.org/document/9008135 (visited on 04/06/2025).

[18] Donald Davidson. “Actions, Reasons, and Causes”. en. In: *The Journal of Philosophy* 60.23 (Nov. 1963), pp. 685–700. DOI: 10.2307/2023177. URL: https://www.pdcnet.org/pdc/bvdb.nsf/purchase?openform&fp=jphil&id=jphil_1963_0060_0023_0685_0700 (visited on 03/07/2025).

[19] Ali Diba et al. “Large Scale Holistic Video Understanding”. en. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 593–610. ISBN: 978-3-030-58558-7. DOI: 10.1007/978-3-030-58558-7_35.

[20] Dave Epstein, Boyuan Chen, and Carl Vondrick. “Oops! Predicting Unintentional Action in Video”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 2575-7075. June 2020, pp. 916–926. DOI: 10.1109/CVPR42600.2020.00100. URL: https://ieeexplore.ieee.org/document/9156404 (visited on 03/07/2025).

[21] Alireza Fathi, Yin Li, and James M. Rehg. “Learning to Recognize Daily Actions Using Gaze”. en. In: *Computer Vision – ECCV 2012*. Ed. by Andrew Fitzgibbon et al. Berlin, Heidelberg: Springer, 2012, pp. 314–327. ISBN: 978-3-642-33718-5. DOI: 10.1007/978-3-642-33718-5_23.

[22] Alireza Fathi, Xiaofeng Ren, and James M. Rehg. “Learning to recognize objects in egocentric activities”. In: *CVPR 2011*. ISSN: 1063-6919. June 2011, pp. 3281–3288. DOI: 10.1109/CVPR.2011.5995444. URL: https://ieeexplore.ieee.org/document/5995444 (visited on 05/05/2025).

[23] David F. Fouhey et al. “People Watching: Human Actions as a Cue for Single View Geometry”. en. In: *Computer Vision – ECCV 2012*. Ed. by Andrew Fitzgibbon et al. Berlin, Heidelberg: Springer, 2012, pp. 732–745. ISBN: 978-3-642-33715-4. DOI: 10.1007/978-3-642-33715-4_53.

[24] Batya Friedman and Helen Nissenbaum. “Bias in computer systems”. In: *ACM Trans. Inf. Syst.* 14.3 (July 1996), pp. 330–347. ISSN: 1046-8188. DOI: 10.1145/230538.230561. URL: https://dl.acm.org/doi/10.1145/230538.230561 (visited on 06/04/2025).

[25] Tao Gao and Brian J. Scholl. “Chasing vs. stalking: Interrupting the perception of animacy.” eng. In: *Journal of Experimental Psychology: Human Perception and Performance* 37.3 (June 2011). Publisher: American Psychological Association, pp. 669–684. ISSN: 0096-1523. DOI: 10.1037/a0020735. URL: https://research.ebsco.com/linkprocessor/plink?id=cf85a9b6-5fab-3bbc-beee-b65121736f2f (visited on 05/27/2025).

[26] Lena Gorelick et al. "Actions as Space-Time Shapes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.12 (Dec. 2007). Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 2247–2253. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2007.70711. URL: https://ieeexplore.ieee.org/document/4359333 (visited on 03/07/2025).

[27] Raghav Goyal et al. "The "Something Something" Video Database for Learning and Evaluating Visual Common Sense". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. ISSN: 2380-7504. Oct. 2017, pp. 5843–5851. DOI: 10.1109/ICCV.2017.622. URL: https://ieeexplore.ieee.org/document/8237884 (visited on 03/07/2025).

[28] Yun He et al. "Human Action Recognition Without Human". en. In: *Computer Vision – ECCV 2016 Workshops*. Ed. by Gang Hua and Hervé Jégou. Cham: Springer International Publishing, 2016, pp. 11–17. ISBN: 978-3-319-49409-8. DOI: 10.1007/978-3-319-49409-8_2.

[29] Douglas Heaven. "Why faces don't always tell the truth about feelings". en. In: *Nature* 578.7796 (Feb. 2020). Bandiera_abtest: a Cg_type: News Feature Publisher: Nature Publishing Group Subject_term: Psychology, Society, Computer science, pp. 502–504. DOI: 10.1038/d41586-020-00507-5. URL: https://www.nature.com/articles/d41586-020-00507-5 (visited on 06/02/2025).

[30] Felienne Hermans and Ari Schlesinger. "A Case for Feminism in Programming Language Design". In: *Proceedings of the 2024 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*. Onward! '24. New York, NY, USA: Association for Computing Machinery, Oct. 2024, pp. 205–222. ISBN: 9798400712159. DOI: 10.1145/3689492.3689809. URL: https://dl.acm.org/doi/10.1145/3689492.3689809 (visited on 05/27/2025).

[31] Yi Huang, Shang-Hong Lai, and Shao-Heng Tai. "Human Action Recognition Based on Temporal Pose CNN and Multi-dimensional Fusion". en. In: *Computer Vision – ECCV 2018 Workshops*. Ed. by Laura Leal-Taixé and Stefan Roth. Cham: Springer International Publishing, 2018, pp. 426–440. ISBN: 978-3-030-11012-3. DOI: 10.1007/978-3-030-11012-3_33.

[32] Anand Jagatia. *How the way you count reveals more than you think*. en-GB. Sept. 2021. URL: https://www.bbc.com/future/article/20210902-how-finger-counting-gives-away-your-nationality (visited on 06/02/2025).

[33] Hueihan Jhuang et al. "Towards Understanding Action Recognition". In: *2013 IEEE International Conference on Computer Vision*. ISSN: 2380-7504. Dec. 2013, pp. 3192–3199. DOI: 10.1109/ICCV.2013.396. URL: https://ieeexplore.ieee.org/document/6751508 (visited on 03/07/2025).

[34] Baoxiong Jia et al. “LEMMA: A Multi-view Dataset for L Earning Multi-agent Multi-task Activities”. en. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 767–786. ISBN: 978-3-030-58574-7. DOI: 10.1007/978-3-030-58574-7_46.

[35] Will Kay et al. *The Kinetics Human Action Video Dataset*. arXiv:1705.06950 [cs]. May 2017. DOI: 10.48550/arXiv.1705.06950. URL: http://arxiv.org/abs/1705.06950 (visited on 03/07/2025).

[36] Seong Tae Kim and Yong Man Ro. “Facial Dynamics Interpreter Network: What Are the Important Relations Between Local Dynamics for Facial Trait Estimation?” en. In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Cham: Springer International Publishing, 2018, pp. 475–491. ISBN: 978-3-030-01258-8. DOI: 10.1007/978-3-030-01258-8_29.

[37] Takumi Kobayashi and Nobuyuki Otsu. “Efficient Optimization for Low-Rank Integrated Bilinear Classifiers”. en. In: *Computer Vision – ECCV 2012*. Ed. by Andrew Fitzgibbon et al. Berlin, Heidelberg: Springer, 2012, pp. 474–487. ISBN: 978-3-642-33709-3. DOI: 10.1007/978-3-642-33709-3_34.

[38] Yu Kong, Yunde Jia, and Yun Fu. “Learning Human Interaction by Interactive Phrases”. en. In: *Computer Vision – ECCV 2012*. Ed. by Andrew Fitzgibbon et al. Berlin, Heidelberg: Springer, 2012, pp. 300–313. ISBN: 978-3-642-33718-5. DOI: 10.1007/978-3-642-33718-5_22.

[39] H. Kuehne et al. “HMDB: A large video database for human motion recognition”. In: *2011 International Conference on Computer Vision*. ISSN: 2380-7504. Nov. 2011, pp. 2556–2563. DOI: 10.1109/ICCV.2011.6126543. URL: https://ieeexplore.ieee.org/document/6126543 (visited on 03/07/2025).

[40] Hilde Kuehne, Ali Arslan, and Thomas Serre. “The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. ISSN: 1063-6919. June 2014, pp. 780–787. DOI: 10.1109/CVPR.2014.105. URL: https://ieeexplore.ieee.org/document/6909500 (visited on 03/07/2025).

[41] Sumith Kulal et al. “Programmatic Concept Learning for Human Motion Description and Synthesis”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 2575-7075. June 2022, pp. 13833–13842. DOI: 10.1109/CVPR52688.2022.01347. URL: https://ieeexplore.ieee.org/document/9879574 (visited on 06/02/2025).

[42] Tian Lan et al. “Learning Action Primitives for Multi-level Video Event Understanding”. en. In: *Computer Vision - ECCV 2014 Workshops*. Ed. by Lourdes Agapito, Michael M. Bronstein, and Carsten Rother. Cham: Springer International Publishing, 2014, pp. 95–110. ISBN: 978-3-319-16199-0. DOI: 10.1007/978-3-319-16199-0_7.

[43] Yin Li, Miao Liu, and James M. Rehg. "In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video". en. In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Cham: Springer International Publishing, 2018, pp. 639–655. ISBN: 978-3-030-01228-1. DOI: 10.1007/978-3-030-01228-1_38.

[44] Yingwei Li, Yi Li, and Nuno Vasconcelos. "RESOUND: Towards Action Recognition Without Representation Bias". en. In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Cham: Springer International Publishing, 2018, pp. 520–535. ISBN: 978-3-030-01231-1. DOI: 10.1007/978-3-030-01231-1_32.

[45] Yixuan Li et al. "MultiSports: A Multi-Person Video Dataset of Spatio-Temporally Localized Sports Actions". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. ISSN: 2380-7504. Oct. 2021, pp. 13516–13525. DOI: 10.1109/ICCV48922.2021.01328. URL: https://ieeexplore.ieee.org/document/9711267 (visited on 04/06/2025).

[46] Ivan Lillo, Alvaro Soto, and Juan Carlos Niebles. "Discriminative Hierarchical Modeling of Spatio-temporally Composable Human Activities". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. ISSN: 1063-6919. June 2014, pp. 812–819. DOI: 10.1109/CVPR.2014.109. URL: https://ieeexplore.ieee.org/document/6909504 (visited on 04/04/2025).

[47] Zhe Lin, Zhuolin Jiang, and Larry S. Davis. "Recognizing actions by shape-motion prototype trees". In: *2009 IEEE 12th International Conference on Computer Vision*. ISSN: 2380-7504. Sept. 2009, pp. 444–451. DOI: 10.1109/ICCV.2009.5459184. URL: https://ieeexplore.ieee.org/document/5459184 (visited on 04/30/2025).

[48] Jingen Liu, Jiebo Luo, and Mubarak Shah. "Recognizing realistic actions from videos "in the wild"". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. ISSN: 1063-6919. June 2009, pp. 1996–2003. DOI: 10.1109/CVPR.2009.5206744. URL: https://ieeexplore.ieee.org/document/5206744 (visited on 03/07/2025).

[49] Weixin Luo, Wen Liu, and Shenghua Gao. "A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. ISSN: 2380-7504. Oct. 2017, pp. 341–349. DOI: 10.1109/ICCV.2017.45. URL: https://ieeexplore.ieee.org/document/8237307 (visited on 03/07/2025).

[50] Ye Luo, Loong-Fah Cheong, and An Tran. "Actionness-Assisted Recognition of Actions". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. ISSN: 2380-7504. Dec. 2015, pp. 3244–3252. DOI: 10.1109/ICCV.2015.371. URL: https://ieeexplore.ieee.org/document/7410728 (visited on 03/07/2025).

[51] Alessandro Manzi, Filippo Cavallo, and Paolo Dario. "A 3D Human Posture Approach for Activity Recognition Based on Depth Camera". en. In: *Computer Vision – ECCV 2016 Workshops*. Ed. by Gang Hua and Hervé Jégou. Cham: Springer International Publishing, 2016, pp. 432–447. ISBN: 978-3-319-48881-3. DOI: 10.1007/978-3-319-48881-3_30.

[52] Manuel J. Marín-Jiménez et al. "LAEO-Net: Revisiting People Looking at Each Other in Videos". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 2575-7075. June 2019, pp. 3472–3480. DOI: 10.1109/CVPR.2019.00359. URL: https://ieeexplore.ieee.org/document/8954303 (visited on 06/02/2025).

[53] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. "Actions in context". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. ISSN: 1063-6919. June 2009, pp. 2929–2936. DOI: 10.1109/CVPR.2009.5206557. URL: https://ieeexplore.ieee.org/document/5206557 (visited on 03/07/2025).

[54] S. Mohammad Mavadati et al. "DISFA: A Spontaneous Facial Action Intensity Database". In: *IEEE Transactions on Affective Computing* 4.2 (Apr. 2013), pp. 151–160. ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2013.4. URL: https://ieeexplore.ieee.org/document/6475933 (visited on 05/13/2025).

[55] Milagros Miceli et al. "Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 161–172. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445880. URL: https://dl.acm.org/doi/10.1145/3442188.3445880 (visited on 06/04/2025).

[56] Yusuke Mitarai and Masakazu Matsugu. "Visual Code-Sentences: A New Video Representation Based on Image Descriptor Sequences". en. In: *Computer Vision – ECCV 2012. Workshops and Demonstrations*. Ed. by Andrea Fusiello, Vittorio Murino, and Rita Cucchiara. Berlin, Heidelberg: Springer, 2012, pp. 321–331. ISBN: 978-3-642-33863-2. DOI: 10.1007/978-3-642-33863-2_32.

[57] Melanie Mitchell. *Artificial Intelligence: A Guide to Thinking Humans*. en-US. New York, NY, USA: Farrar, Straus and Giroux, 2019.

[58] Pavlo Molchanov et al. "Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919. June 2016, pp. 4207–4215. DOI: 10.1109/CVPR.2016.456. URL: https://ieeexplore.ieee.org/document/7780825 (visited on 03/07/2025).

[59] C.Thi Nguyen. *The Limits of Data*. en-US. 2024. URL: https://issues.org/limits-of-data-nguyen/.

[60] Bingbing Ni, Pierre Moulin, and Shuicheng Yan. "Order-Preserving Sparse Coding for Sequence Classification". en. In: *Computer Vision – ECCV 2012*. Ed. by Andrew Fitzgibbon et al. Berlin, Heidelberg: Springer, 2012, pp. 173–187. ISBN: 978-3-642-33709-3. DOI: 10.1007/978-3-642-33709-3_13.

[61] Juan Diego Ortega et al. "DMD: A Large-Scale Multi-modal Driver Monitoring Dataset for Attention and Alertness Analysis". en. In: *Computer Vision – ECCV 2020 Workshops*. Ed. by Adrien Bartoli and Andrea Fusiello. Cham: Springer International Publishing, 2020, pp. 387–405. ISBN: 978-3-030-66823-5. DOI: 10.1007/978-3-030-66823-5_23.

[62] Seymour Papert. *The Summer Vision Project*. Accepted: 2004-10-04T14:40:06Z. July 1996. URL: https://dspace.mit.edu/handle/1721.1/6125 (visited on 07/02/2025).

[63] Asanka G. Perera, Yee Wei Law, and Javaan Chahl. "UAV-GESTURE: A Dataset for UAV Control and Gesture Recognition". en. In: *Computer Vision – ECCV 2018 Workshops*. Ed. by Laura Leal-Taixé and Stefan Roth. Cham: Springer International Publishing, 2018, pp. 117–128. ISBN: 978-3-030-11012-3. DOI: 10.1007/978-3-030-11012-3_9.

[64] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah. "Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. ISSN: 1063-6919. June 2008, pp. 1–8. DOI: 10.1109/CVPR.2008.4587727. URL: https://ieeexplore.ieee.org/document/4587727 (visited on 03/07/2025).

[65] Nithya Sambasivan et al. ""Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–15. ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445518. URL: https://dl.acm.org/doi/10.1145/3411764.3445518 (visited on 05/31/2025).

[66] Morgan Klaus Scheuerman, Alex Hanna, and Remi Denton. "Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW2 (Oct. 2021), 317:1–317:37. DOI: 10.1145/3476058. URL: https://dl.acm.org/doi/10.1145/3476058 (visited on 05/27/2025).

[67] Morgan Klaus Scheuerman et al. "How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis". In: *Proc. ACM Hum.-Comput. Interact.* 4.CSCW1 (May 2020). Place: New York, NY, USA Publisher: Association for Computing Machinery. DOI: 10.1145/3392866. URL: https://doi.org/10.1145/3392866.

[68] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. "Intersectional HCI: Engaging Identity through Gender, Race, and Class". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. New York, NY, USA: Association for Computing Machinery, May 2017, pp. 5412–5427. ISBN: 978-1-4503-4655-9. DOI: 10.1145/3025453.3025766. URL: https://dl.acm.org/doi/10.1145/3025453.3025766 (visited on 06/04/2025).

[69] C. Schuldt, I. Laptev, and B. Caputo. "Recognizing human actions: a local SVM approach". In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* Vol. 3. ISSN: 1051-4651. Aug. 2004, 32–36 Vol.3. DOI: 10.1109/ICPR.2004.1334462. URL: https://ieeexplore.ieee.org/document/1334462 (visited on 03/07/2025).

[70] Amir Shahroudy et al. "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919. June 2016, pp. 1010–1019. DOI: 10.1109/CVPR.2016.115. URL: https://ieeexplore.ieee.org/document/7780484 (visited on 03/07/2025).

[71] Jing Shao et al. "Deeply learned attributes for crowded scene understanding". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919. June 2015, pp. 4657–4666. DOI: 10.1109/CVPR.2015.7299097. URL: https://ieeexplore.ieee.org/document/7299097 (visited on 04/06/2025).

[72] Minho Shim et al. "Teaching Machines to Understand Baseball Games: Large-Scale Baseball Video Database for Multiple Video Understanding Tasks". en. In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Cham: Springer International Publishing, 2018, pp. 420–437. ISBN: 978-3-030-01267-0. DOI: 10.1007/978-3-030-01267-0_25.

[73] Gunnar A. Sigurdsson et al. "Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding". en. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, 2016, pp. 510–526. ISBN: 978-3-319-46448-0. DOI: 10.1007/978-3-319-46448-0_31.

[74] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. *UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild*. arXiv:1212.0402 [cs]. Dec. 2012. DOI: 10.48550/arXiv.1212.0402. URL: http://arxiv.org/abs/1212.0402 (visited on 03/07/2025).

[75] Sebastian Stein and Stephen J. McKenna. "Combining embedded accelerometers with computer vision for recognizing food preparation activities". In: *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. UbiComp '13. New York, NY, USA: Association for Computing Machinery, Sept. 2013, pp. 729–738. ISBN: 978-1-4503-1770-2. DOI: 10.1145/2493432.2493482. URL: https://dl.acm.org/doi/10.1145/2493432.2493482 (visited on 03/06/2025).

[76] Waqas Sultani, Chen Chen, and Mubarak Shah. "Real-World Anomaly Detection in Surveillance Videos". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. ISSN: 2575-7075. June 2018, pp. 6479–6488. DOI: 10.1109/CVPR.2018.00678. URL: https://ieeexplore.ieee.org/document/8578776 (visited on 04/30/2025).

[77] Yansong Tang et al. "COIN: A Large-Scale Dataset for Comprehensive Instructional Video Analysis". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 2575-7075. June 2019, pp. 1207–1216. DOI: 10.1109/CVPR.2019.00130. URL: https://ieeexplore.ieee.org/document/8953268 (visited on 05/11/2025).

[78] Alex S. Taylor. "Machine intelligence". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. New York, NY, USA: Association for Computing Machinery, Apr. 2009, pp. 2109–2118. ISBN: 978-1-60558-246-7. DOI: 10.1145/1518701.1519022. URL: https://dl.acm.org/doi/10.1145/1518701.1519022 (visited on 06/06/2025).

[79] Manuel Vázquez Enríquez et al. “ECCV 2022 Sign Spotting Challenge: Dataset, Design and Results”. en. In: *Computer Vision – ECCV 2022 Workshops*. Ed. by Leonid Karlinsky, Tomer Michaeli, and Ko Nishino. Cham: Springer Nature Switzerland, 2022, pp. 225–242. ISBN: 978-3-031-25085-9. DOI: 10.1007/978-3-031-25085-9_13.

[80] Lan Wang et al. “PM-GANs: Discriminative Representation Learning for Action Recognition Using Partial-Modalities”. en. In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Cham: Springer International Publishing, 2018, pp. 389–406. ISBN: 978-3-030-01231-1. DOI: 10.1007/978-3-030-01231-1_24.

[81] Yaohui Wang et al. “Comparing Methods for Assessment of Facial Dynamics in Patients with Major Neurocognitive Disorders”. en. In: *Computer Vision – ECCV 2018 Workshops*. Ed. by Laura Leal-Taixé and Stefan Roth. Cham: Springer International Publishing, 2018, pp. 144–157. ISBN: 978-3-030-11024-6. DOI: 10.1007/978-3-030-11024-6_10.

[82] Ping Wei et al. “Concurrent Action Detection with Structural Prediction”. In: *2013 IEEE International Conference on Computer Vision*. ISSN: 2380-7504. Dec. 2013, pp. 3136–3143. DOI: 10.1109/ICCV.2013.389. URL: https://ieeexplore.ieee.org/document/6751501/citations#citations%20 (visited on 04/04/2025).

[83] Junwu Weng, Chaoqun Weng, and Junsong Yuan. “Spatio-Temporal Naive-Bayes Nearest-Neighbor (ST-NBNN) for Skeleton-Based Action Recognition”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919. July 2017, pp. 445–454. DOI: 10.1109/CVPR.2017.55. URL: https://ieeexplore.ieee.org/document/8099538/ (visited on 05/13/2025).

[84] Chenxia Wu et al. “Watch-n-patch: Unsupervised understanding of actions and relations”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919. June 2015, pp. 4362–4370. DOI: 10.1109/CVPR.2015.7299065. URL: https://ieeexplore.ieee.org/document/7299065 (visited on 05/05/2025).

[85] Peng Wu et al. “Not only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision”. en. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 322–339. ISBN: 978-3-030-58577-8. DOI: 10.1007/978-3-030-58577-8_20.

[86] Liang Xu et al. “ActFormer: A GAN-based Transformer towards General Action-Conditioned 3D Human Motion Generation”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. ISSN: 2380-7504. Oct. 2023, pp. 2228–2238. DOI: 10.1109/ICCV51070.2023.00212. URL: https://ieeexplore.ieee.org/document/10377513 (visited on 03/07/2025).

[87] Chunfeng Yuan et al. “3D R Transform on Spatio-temporal Interest Points for Action Recognition”. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. ISSN: 1063-6919. June 2013, pp. 724–730. DOI: 10.1109/CVPR.2013.99. URL: https://ieeexplore.ieee.org/document/6618943/ (visited on 05/13/2025).

[88] Fei Yuan, Véronique Prinet, and Junsong Yuan. “Middle-Level Representation for Human Activities Recognition: The Role of Spatio-Temporal Relationships”. en. In: *Trends and Topics in Computer Vision*. Ed. by Kiriakos N. Kutulakos. Berlin, Heidelberg: Springer, 2012, pp. 168–180. ISBN: 978-3-642-35749-7. DOI: 10.1007/978-3-642-35749-7_13.

[89] He Zhao and Richard P. Wildes. “On Diverse Asynchronous Activity Anticipation”. en. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 781–799. ISBN: 978-3-030-58526-6. DOI: 10.1007/978-3-030-58526-6_46.

[90] Kang Zheng et al. “Video-Based Action Detection Using Multiple Wearable Cameras”. en. In: *Computer Vision - ECCV 2014 Workshops*. Ed. by Lourdes Agapito, Michael M. Bronstein, and Carsten Rother. Cham: Springer International Publishing, 2014, pp. 727–741. ISBN: 978-3-319-16178-5. DOI: 10.1007/978-3-319-16178-5_51.