1 Introduction

In video understanding tasks, action recognition and detection are prominent and meaningful due to their practical applications in daily life. Some notable applications include Surveillance and Security, Human-Computer Interaction, Sports Analysis, Entertainment and Gaming, among others. Although deep learning models designed to solve these problems often require significant computational resources, with the advancement of computer hardware, the deployment in real-world scenarios while meeting real-time processing speed has become more feasible over time.

Besides the requirement for significant computational resources, they also demand a large and sufficiently complex dataset. In addition to serving as training data, datasets also provide a portion of data specifically for evaluating models, thereby establishing a common benchmark for comparing different models. Over the years, new datasets have emerged, either as additions to existing datasets or as entirely new ones based on different construction perspectives. This has increased both the diversity and quantity of available data, but also inadvertently posed challenges in selecting an appropriate dataset. Evaluating whether a dataset is suitable for a given research problem is not merely a matter of its scale. Other characteristics must also be considered, such as the dataset creator's perspective, data collection methods, sample size, number of classes, level of annotation detail (spatial, temporal, sound, etc.), popularity within the research community, the baseline for comparison, and various other factors. Therefore, it is necessary to carefully examine datasets relevant to the task, gather information, evaluate, and then compare them to ultimately select the desired dataset for research purposes. This process typically consumes a significant amount of time and effort. To address this issue, in this paper, we aim to compile notable datasets in the fields of action detection and action recognition, listing them chronologically while providing concise necessary information regarding:

- *Context and construction perspective of the dataset*: Since the datasets are presented chronologically, this section clarifies the information regarding the background and the authors' perspectives on the shortcomings or the necessary additions to older datasets.
- *Dataset distribution*: Information about the dataset, such as the number of data samples, the number of classes, the train-validation splits, and any other available details.
- *Annotations*: Explanation of the annotations provided in the dataset.
- *Data collection methods*: We summarize the data collection process employed by the respective author groups on that dataset. This allows for a more objective assessment of the dataset's reliability and quality based on the researcher's perspective.

In section II, we will provide a brief overview of the history and context of the field of artificial intelligence research from its inception to the emergence of CNN models and their dominance from image task to video task. Having a general understanding of the history and context will help readers understand why datasets have their limitations and continue to evolve over the years.

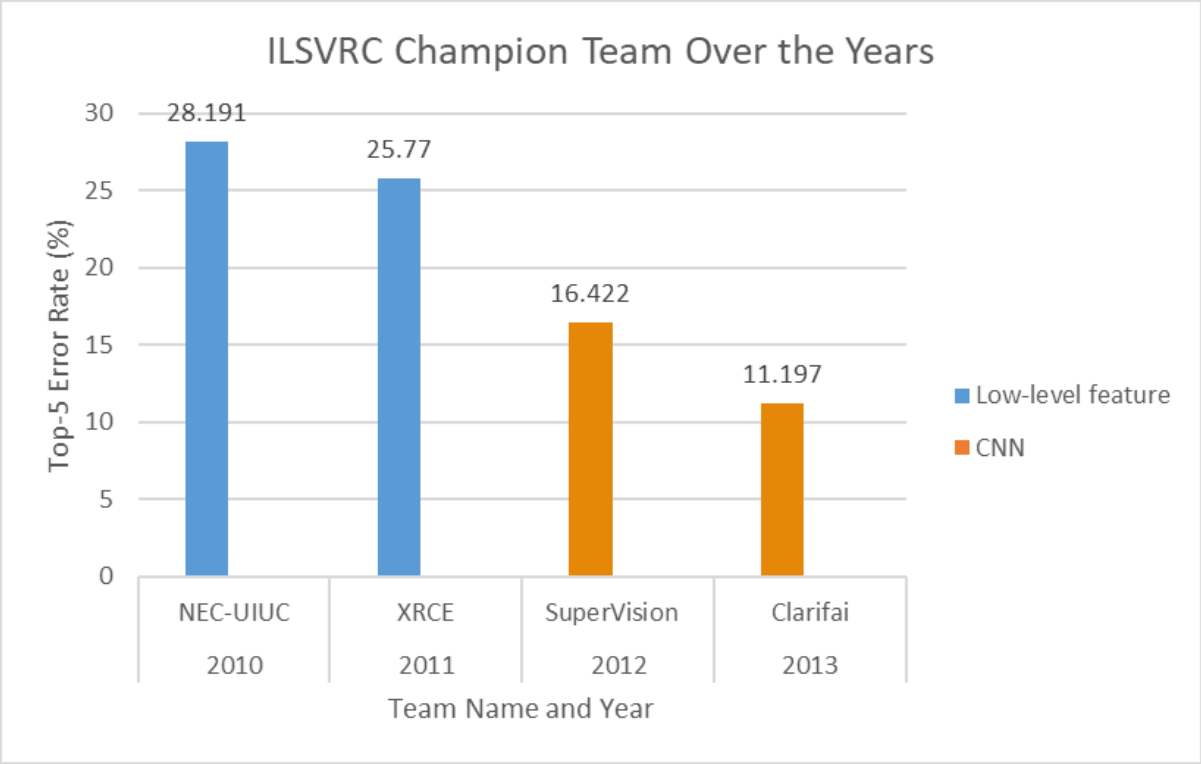
In section III, we list the datasets in the order of their publication time (measured from the time the accompanying paper is published). Each dataset includes four sections presented in the following order: "Context and Construction Perspective of the Dataset," "Annotations," "Dataset Distribution," and "Data Collection Methods." If some information is not provided by the authors in the original paper, it will be left blank or omitted. Additionally, if the authors provide any additional information included in the dataset, we will allocate a separate section below to describe it. The list of datasets, along with a brief overview of their publication dates and the mentioned data quantities, can be found in Fig1..

2 Overview History

Although the field of artificial intelligence emerged in the mid-1990s, it took several decades for significant progress to be made, thanks to the remarkable advancements in computer hardware - greater computing power and easier accessibility. As a result, the research community's interest in AI has significantly increased. AI competitions began to be organized, particularly in computer vision, attracting numerous research groups. In 2010, the ImageNet Large Scale Visual Recognition Challenge

(ILSVRC) [?] was initiated, aiming to build upon the success of the PASCAL VOC challenge [?] by evaluating model performance in image recognition tasks.

ILSVRC, upon its initial launch, garnered significant attention and credibility within the research community due to its unprecedented scale of data: 1.2 million training images and 1,000 object classes. It attracted participation from top researchers in the field and further solidified its reputation.



Hình 1: ILSVRC champion team over years on classification task

Go back to 1998 when LeNet [?], the first CNN model, was introduced. At that time, CNN was just one of many research directions and had not received much attention. It wasn't until 2012 when the SuperVison team, led by researchers at the University of Toronto, proposed a CNN model called AlexNet and convincingly won the ILSVRC2012 in image classification with a top-5 error rate of only 16.422% (Fig 1). They completely outperformed other competitors at that time, paving the way for the era of CNN and the dawn of Deep Learning. Since then, winning solutions in subsequent years of ILSVRC have consistently utilized CNN. Over time, there has been an increasing number of research studies applying CNN models to various tasks. Through experimentation, CNN has proven to be effective not only in image classification but also in localization, segmentation, and even beyond computer vision, extending to other fields such as speech processing and natural language processing. CNN has shown great potential in developing solutions for previously challenging problems that were not adequately addressed. One such problem is action recognition, which is a highly significant task with practical applications. CNN has opened up possibilities for developing solutions to previously unresolved problems, and action recognition is just one of them, receiving considerable attention and practical implications.

The problem of Action Recognition existed before the rise of CNN. Solutions for action recognition during this period often involved feature extraction using various methods to obtain a feature vector from the data, followed by a classifier, typically a Support Vector Machine (SVM). This approach is called "hand-crafted feature" and it continued to dominate other methods, including CNNs, until 2015 because CNN models were still relatively new and not extensively explored. Over time, the research community gradually replaced these hand-crafted feature methods with CNNs. Continuous advancements and proposals of CNN models for action recognition have been made, such as Two-stream networks, Segment-based methods, Multi-stream networks, 3D CNNs, and so on. Alongside these developments, there has been an increasing demand for computational power and a significant growth in the amount of data. The datasets used in this period were also limited, as shown in Table 3, which lists prominent datasets from before 2012. It can be observed that in terms of scale (number of classes, number of video clips),

the datasets were still quite limited.

Bảng 1: Action recognition dataset

Name	Year	NumClass	Clip/Class	Ref
KTH	2004	6	100	[?]
Weizmann	2005	9	9	[?]
IXMAS	2007	11	33	[?]
Hollywood	2008	8	30-129	[?]
UCF Sports	2008	9	14-35	[?]
Hollywood2	2009	12	61-278	[?]
UCF YouTube	2009	11	100	[?]
Olympic	2010	16	50	[?]

Building a dataset typically goes through four steps: (1) Defining a set of predefined actions, (2) Collecting videos from data sources, (3) Annotating the data (either automatically, semi-automatically, or manually), (4) Cleaning and filtering the data to remove duplicates and noise. Each step presents its own challenges. In step (1), defining an action is not a simple task as humans often perform complex combinations of gestures. So, what constitutes an "atomic action"? In step (2), do the video sources comply with copyright rules? Privacy regulations? Is the dataset stable (not prone to loss or replacement)? In step (3), the workload scales with the dataset size, and there can be vague boundaries in determining the start and end of an action. In step (4), what criteria are used to evaluate whether a video meets the standards for usability? There are many related questions, such as the availability of human and financial resources, required to meet the demands of building a comprehensive research dataset. Furthermore, considering the context before CNNs gained significant prominence, investing in developing a large-scale dataset was highly risky.

CNN models possess immense power that scales with their complexity., is prone to overfitting, especially when dealing with small amounts of data. During the explosion of CNN, research groups faced many challenges due to data scarcity. Methods like data augmentation were effective solutions, but it was still necessary to supplement larger and more complex datasets to meet the growing demand for data in CNN models. Another reason is that the remarkable success of CNNs in image processing tasks has been greatly contributed by large-scale datasets like ImageNet. However, in the video domain, there is currently no comparable dataset to ImageNet. Realizing this need, research groups from all over the AI research community have continuously improved and published increasingly refined datasets. These datasets play a crucial role as a common benchmark for comparing different models.

3 Overview Dataset

Nói về context và sơ lược về thông tin, quy mô của bộ dữ liệu.

3.1 HMDB51

In 2011, HMDB51 [?] was introduced to highlight differences among action categories based on motion rather than static poses, which were commonly used in datasets like KTH [?] and Weizmann [?]. With the growing demand for datasets that present more substantial challenges and enhance the capabilities of action recognition systems, HMDB51 aims to enrich the contextual background and increase the number of action categories, thereby improving the utility of recognition systems in real-life applications.

It includes 6,766 video clips covering 51 different action categories sourced from movies and Internet, with each category having at least 101 clips. Each video contains a single action described in its name, with a quality of 240px and length of more than 1 second.

3.2 UCF101

Introduced in 2012 as an expansion of UCF50 by adding 51 classes, UCF101 [**?**] aimed to refine and expand the range of actions captured in the previous datasets, contributing to a more comprehensive understanding of human activities. Additionally, the dataset also blurs the difference between artificial and real-life videos, therefore accurately reflects the diversity of human actions represented in real-world scenarios. Among the four versions of the UCF family: "UCF Sports [**?**], UCF11 [**?**], UCF50, and UCF101", the dataset being referred to is the largest and most widely used.

The dataset contains 13,320 videos demonstrating 101 actions. Each clip has a resolution of 320x240 pixels at 25 FPS. Additionally, this dataset also includes audio data for 50 action classes. Each clip has its augmented version, which expands into various variants of itself, such as increasing or reducing the length, changing the segment, or adding audio.

3.3 Sport-1M

In 2014, Sport-1M [**?**] was developed to apply CNN methods to action recognition categories, marking a departure from datasets such as HMDB51 [**?**] and UCF101 [**?**], which were designed to accommodate traditional machine learning methods, especially SVM, thus became non-equivalent. Therefore, with the rise of CNNs, there has been encouragement to develop equivalent resources that can enhance their capabilities.

The dataset contains 1,133,158 YouTube videos, covering a diverse range of content. There are 487 unique classes in the dataset, with each class ranging from 1000 to 3000 videos. Notably, each video may be assigned multiple labels, indicating that a single video could have more than one annotation, accounting for approximately 5% of the dataset.

3.4 ActivityNet

First introduced in 2015, ActivityNet [**?**] offer a solution with a large-scale dataset that provides a high level of specificity for human daily life in a hierarchical structure. At that time, UCF101 [**?**] and THUMOS-14 [**?**] already have their own category distributions, but they lack detailed organization and depth of levels, which limits the amount of information they provide. This hierarchical organization allows for a detailed understanding of the diverse range of behaviors captured in this dataset, especially in every day activities.

In the version 1.3 released in Mar 2016, ActivityNet contains 849 hours from 27,801 videos of indoor actions with a total of 68.8 hours that appear 200 human-centric activities. Most of the videos have lengths ranging from 5 to 10 minutes at 30 FPS have HD resolution quality (1280x720).

3.5 Youtube-8M

In 2016, YouTube-8M [**?**] was introduced as the largest multi-label video classification dataset, which utilized the content-based annotation method. Unlike previous datasets such as Sports-1M [**?**] and ActivityNet [**?**], which only assigned a limited number of action categories, YouTube-8M employed Knowledge Graph entities to filter topics presented on YouTube and therefore broadening the scope to cover a wide range of activities. The purpose of this dataset is to understand the main actions in the video and summarize them into key topics.

In the newest update in May 2018, YouTube-8M underwent a cleanup process to ensure quality for both video resolution and annotation vocabulary. It removed the private, unfamous and sensitive contents for safety purpose. Currently, the dataset comprises over 6.1 million video IDs from 3,862 entities, grouped into 24 high-level topics. Each video is required to be between 120 and 500 seconds long and must contain at least one target vocabulary. The dataset allows for multi-label assignments, with each video typically having an average of 3.0 assigned classes.

3.6 Charades

Introduced in 2016, Charades [?] is described as "Hollywood in Homes" when using a man-made dataset instead of downloading videos from YouTube. Using a similar method to the Something [?] dataset, a group of AMT workers was hired to employ the Hollywood filming method to create clips from diverse environments. Due to the noisy labels and background context from datasets sourced from the internet, such as HMDB51 [?] and UCF101 [?], Charades aims to create a high-quality and realistic dataset, particularly focusing on daily activities.

The dataset comprises of 9,848 videos with an average length of 30 seconds, which demonstrates 157 action classes and 46 object classes. Overall, there are 27,847 video descriptions and 66,500 temporally localized action contributed by 267 people.

3.7 Something Something

Introduced its first version in 2017, Something V1 [?] emphasizes detailed interactions between human actions and objects, aiming to provide fine-grained videos that reflect real-world aspects. The YouTube-sourced datasets, such as Sport-1M [?] or YouTube-8M [?], although notably large in size, still involve combining features extracted from frames, thus becoming a "set of images" classification task.

When an action is combined with various objects, it can potentially mislead the model since it diverges from its learned associations due to the lack of contextual understanding of how different actions correlate with each other of datasets. As an illustration, consider the action of "pointing" which can result in two scenarios: "Pointing a finger" (Harmless) or "Pointing a knife" (Dangerous). The main objective of Something Something dataset is to address this problem.

In the newer V2 version released in 2018, the number of videos has increased to 220,847 clips, which is twice the number in V1, while retaining the same set of labels totaling 174. Additionally, each clip has been upgraded to a quality of 240px. Each clip has an average length of 2-6 seconds performing a single action at 12 FPS. Overall, there are 318,572 annotations, which involve 30,408 unique objects.

3.8 Kinetics

In 2017, one of the most famous action classification datasets, Kinetics, was introduced. Its method combines elements from the previous HMDB51 [?] and UCF101 [?]datasets and expands the number of action classes to 400. By collecting videos from YouTube, the dataset can capture various camera motions, angles, lighting conditions, etc., and therefore covering a broad range of human actions.

Later in 2018, an updated version, Kinetics-600, was introduced. Kinetics-600 is a superset of Kinetics-400, retaining the original 368 classes and splitting 32 classes to provide clearer explanations. Additionally, a new filtering method was used to gather videos correlated to the action classes. Due to some validation sets from Kinetics-400 becoming part of the Kinetics-600 test set, it is recommended not to evaluate Kinetics-600 with a pre-trained Kinetics-400 model.

In 2020, Kinetics-700-2020 was introduced, expanding by 30% compared to Kinetics-600. Additional actions were sourced partly from EPIC-Kitchens and AVA datasets. Rare actions were gathered from more videos, and duplicated videos were removed. These changes resulted in a more balanced dataset.

The Kinetics dataset contains 10 seconds short clips demonstrating the mentioned action. The table below shows the size of the three Kinetics version:

Dataset	Total videos	Action classes	Average clips per class
Kinetics-400	306,245	400	683
Kinetics-600	495,547	600	762
Kinetics-700-2020	647,907	700	926

3.9 AVA

Introduced in 2018, AVA's [?] main goal is to overcome weaknesses of previous datasets like Sports-1M [?], YouTube-8M [?], Something Something [?], and Moments in Time [?], which focus on large-scale datasets and are often annotated automatically, leading to noisy annotations. Other datasets such as ActivityNet [?], THUMOS [?], and Charades utilize a large number of videos containing multiple actions but only provide temporal annotations. Therefore, AVA provides realistic fine-grained recognition in a complex environment where actors perform a set of combined actions, aiming to enhance spatio-temporal action localization.

Currently, there are four different versions of AVA. The newest version, v2.2, consists of a total of 430 videos covering 80 classes extracted from movies. Each video contributes 15 minutes of footage sampled at a rate of 1Hz, which translates to one frame per second, resulting in 897 segments per 15 minutes.

3.10 EPIC-KITCHENS

Since its first introduction in 2018, EPIC-KITCHENS (2020) now extends to provide a fully version of a large-scale egocentric dataset. Recently, ATM workers are being frequently utilized to collect desired video footage scripted scenarios, resulting in great contributions to projects like Something Something [?] and Hollywood in Home [?]. However, this practice also leads to a lack of natural actions in real life. Given that situation, EPIC-KITCHENS captures random multitasking actions performed by real individuals without any scripts. By recording daily kitchen activities from the first-person perspective of 32 participants from 10 different countries, it aims to present a challenging real-life scenario.

The number of records in the dataset amounts to 55 hours in length. Within it, 39,596 action segments and 454,158 object bounding boxes are extracted. Recorded with GoPro devices, the clips are captured in Full HD resolution at 60 FPS, resulting in 11.5 million frames. The average length of each clip is 1.7 hours, starting from the moment the actor goes to their kitchen and ending when they finish their work, describing both the preparation and cooking process. After that, both objects and actions were annotated manually.

3.11 Moments in Time

In 2019, Moments in Time [?] was introduced and became one of the largest datasets comprising hundreds of verbs depicting moments lasting a few seconds. Over the years, the rapid growth of datasets has expanded the usability of human action understanding. Large-scale video datasets such as Kinetics and YouTube-8M [?] play significant roles in studying open-world vocabulary from the internet. Other datasets, such as ActivityNet [?] and AVA [?], explore recognizing and localizing fine-grained actions by linking correlations. To enhance these characteristics, Moments in Time aims to ensure a high-quality and balanced dataset, capturing both inter- and intra-class variations across different levels of abstraction for video understanding.

The dataset contains more than 1,000,000 labeled 3-second videos, which include 339 action classes. The actors performing actions are not just limited to humans but also include animals or cartoon characters. Therefore, this dataset proposes a new challenge in recognizing events across various actors. Moreover, sound-dependent classes are added to expand the capability of understanding auditory cues.

3.12 HACS

HACS [?] emerged in response to the increasing need for extensive datasets, facilitating the development of more sophisticated models in the realm of action recognition. Inspired by the notable expansions witnessed in large-scale action recognition datasets like Sport-1M, Kinetics, and Moments in Time, HACS enhances both its scale and quality to offer a more encompassing resource. Moreover, it builds upon the strengths of past action localization datasets such as THUMOS [?], AVA [?], Charades [?], and especially ActivityNet [?].

The dataset provides 504K videos sourced from YouTube, categorized into 200 action classes. These videos are trimmed into shorter segments, resulting in a total of 1.5M clips, each lasting 2 seconds,

for more accurate labeling which is called HACS Segments. Then, it is annotated into positive (has action) and negative (doesn't has action) samples.

3.13 HVU

Introduced in 2020 as a multi-label and multi-task fully annotated dataset, HVU [**?**] provides a multi-label and multi-task large-scale video benchmark with a comprehensive list of tasks and annotations for video analysis and understanding. CNNs model has envolved to be stronger and faster in recent years, but the datasets just allow them to recognize single label per task, which hinders the learning of ConvNets.

HVU comprises a total of 572,000 videos and 3,143 labels. It consists of trimmed video clips with varying durations, capped at a maximum length of 10 seconds. Additionally, HVU does not solely rely on a single action class but instead includes multiple tags which is organized into six main categories: scene, object, action, event, attribute, and concept.

3.14 AViD

Introduced in 2020, AViD [**?**] aims to provide an Anonymized Videos from Diverse Countries dataset. In the past, datasets such as Kinetics, HACS [**?**], and HVU [**?**], although containing numerous labeled video clips, were predominantly limited to the USA and other English-speaking countries. Moreover, those datasets were mainly sourced from YouTube links, which may not be available in some countries. AViD solves that problem by saving it as a static dataset, which can be found at the relevant link provided by the authors. When collecting videos, the authors blurred all the actors' faces to prevent machines from recognizing people in the videos but still reliably recognized actions, which is also a unique characteristic of this dataset.

After the filtering process, the dataset has a total of more than 800K videos from over the world, demonstrating 887 classes. The labels follow hierarchy structure from general to particular action for studying various aspects of action performance.

3.15 TinyVIRAT

3.16 ToyotaSmarthome

3.17 FineAction

4 Analysis

Các thông tin ở từng mục sẽ được nói kĩ nhưng không đi sâu vào từng bộ dữ liệu mà mang tính tổng quát, sau đó thống kê vào bảng. Thông tin trên bảng sẽ được điền dạng kí hiệu hoặc yes/no, sau đó chú thích bên dưới.

4.1 Characteristic

Nói về đặc điểm hiện có của các bộ dữ liệu theo các yếu tố trong bảng bên dưới.

Bảng 2: Action recognition dataset

tên	phân loại(như trong sheet)	temporal	spatial	classification-only	Focus on (scale/diverse class ...)

4.2 Data collection method

Thảo luận 3 mục :

- Build Action class list : Dùng các công trình nghiên cứu về ngôn ngữ / Tự tiến hành nghiên cứu / Sử dụng nguồn từ dataset trước đó có/không bổ sung thêm .
- Collect video : Từ Internet / Tự quay.
- Annotation : Tự động / Thủ công / Bán tự động.

Bảng dự kiến :

Bảng 3: Action recognition dataset

tên	pp tạo act list	nguồn vid	pp anno	các tool anno được sử dụng

4.3 Data statistic

Phần này chưa biết phải viết gì nhiều, có thể giải thích tại sao một số data không công bố test anno.

Bảng dự kiến :

Bảng 4: Action recognition dataset

tên	numclass	train	val	test	duration/sample

4.4 Benchmark and metric

Phần này cũng chưa biết viết gì.

Bảng dự kiến :

Bảng 5: Action recognition dataset

tên	benchmark	metric	eval protocol

4.5 state of the art method result

Chưa biết viết gì

4.6 Discussion

Chỉ vừa nghĩ ra được một mục

4.6.1 Limitations of current datasets

Thuyết minh + đưa ra dẫn chứng cụ thể.

- Những data cũ :
  - Data : Quy mô nhỏ, đơn giản.
  - Staturation : sota method đạt ngưỡng rất cao.
- Những data mới :
  - Annotation : vẫn còn nhiều nhiễu/thiếu chính xác.
  - Data : Chất lượng phân giải / thiếu ổn định (bị gỡ bỏ, không thể tiếp cận).

Bảng 6: Action recognition dataset

tên	SOTA method	result	metric	eval Protocol

5 Conclusion

6 References