Video Understanding : A review of action detection-recognition dataset

First Author
firstauthor@i1.org

Second Author
secondauthor@i2.org

Ngày 26 tháng 2 năm 2024

Mục lục

Tóm tắt nội dung

In this article, we provide a summary and an overview of the datasets used in the task of action detection/recognition. The datasets will be presented in the order of their publication time. For each dataset, we sequentially present four aspects: the context of its creation, data distribution, explanations of annotations, and data collection methods.

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.

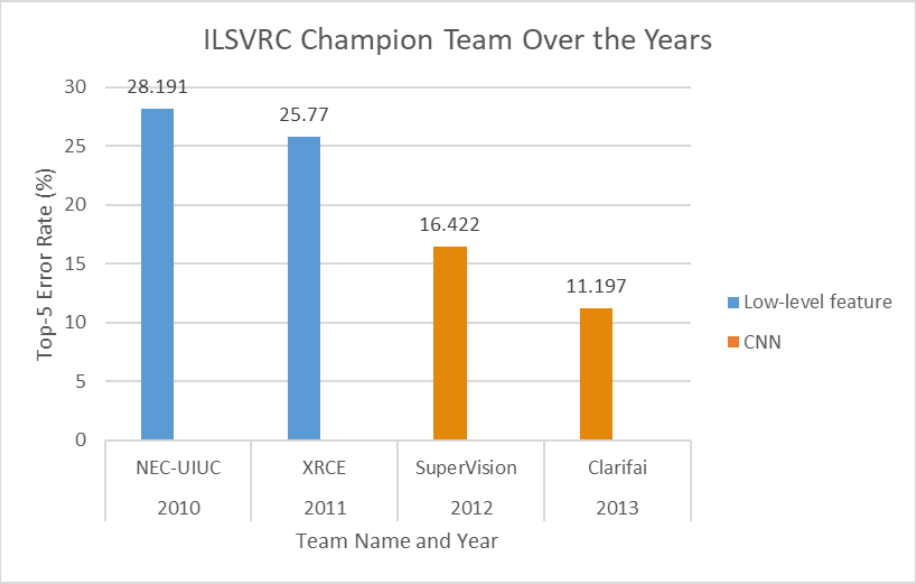
In section II, we will provide a brief overview of the history and context of the field of artificial intelligence research from its inception to the emergence of CNN models and their dominance from image task to video task. Having a general understanding of the history and context will help readers 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 of 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) [1] was initiated, aiming to build upon the success of the PASCAL VOC challenge [2] 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 [3], 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. Solu-

tions 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 1, 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.

Bång 1: Action recognition dataset

Name	Year	NumClass	Clip/Class	Ref
KTH	2004	6	100	[4]
Weizmann	2005	9	9	[5]
IXMAS	2007	11	33	[6]
Hollywood	2008	8	30-129	[7]
UCF Sports	2008	9	14-35	[8]
Hollywood2	2009	12	61-278	[9]
UCF YouTube	2009	11	100	[10]
Olympic	2010	16	50	[11]

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 Review

3.1 HMDB51

- Year : 2011
- Paper : HMDB: A Large Video Database for Human Motion Recognition [12].

3.1.1 Context and construction perspective

KTH [4] and Weizmann [5] have long been regarded as pioneering datasets in the early stages of action recognition. However, over time, their limitations have become evident, particularly in terms of a restricted number of action categories and simplistic background contexts. With the advent of recent models, achieving accuracy rates exceeding 90% became commonplace. As a result, there is a growing demand for alternative datasets that present more substantial challenges, aiming to push the boundaries and enhance the capabilities of action recognition systems. The dataset is carefully designed to highlight differences among action categories based on motion rather than static poses. With the valuable contribution, the dataset has a potential of significantly enhance the evaluation and future utilization of recognition systems in real life.

3.1.2 Data collection methods

A group of student was asked to collect and annotate any segment which represents a single non-ambiguous human action from videos. The videos are sourced from digitized movies, public databases like the Prelinger archive, additional online videos, as well as content from YouTube, Google videos and other videos from internet. Students also were asked to consider some minimum quality standard : only one action per clip, minimum height for main actor should be 60 pixels, minimum contrast level, minimum of clip length is 1 second, acceptable compression artifacts, etc.

3.1.3 Data distribution

It includes 6,766 video clips covering 51 different action categories, with each category having at least 101 clips.

- Train : for each action, 70 random clips are used.
- Validation : no validation set.
- Test : for each action, 30 random clips are used.

The authors generated three distinct training/testing split follow above rule, ensure that no test clip is also not in train clip.

3.1.4 Annotations

Hình 2:

For each class, there is 3 split files as mentioned above. Each file includes the names of all the videos in the same class along with an index belonging to $\{0, 1, 2\}$ as shown in Figure 2. The file used for training is indexed as 1, for testing it is 2, and 0 indicates that it is not used in this split.

Bảng 2: Action recognition dataset

PROPERTY	CATEGORIES
Visible body parts	head(h), upper body(u), full body (f), lower body(l)
Camera motion	motion (cm), static (nm)
Number of people involved in the action	single (np1), two (np2), three (np3)
Camera viewpoint	front (fr), back (ba), left(le), right(ri)
Video quality	good (goo), medium (med), ok (bad)

The name of each video also carries a meaning, with a specific structure as follows:
vid-name_class-name_vible-body-part_cam-motion_num-of-people_cam-viewpoint_vid-quality
The abbreviations used are listed in Table 2.

For example, the name IPL_Awards_Ceremony_shake_hands_f_cm_np2_ba_med_0.avi represents a video titled IPL Awards Ceremony, belonging to the shake hands class, with 'f' indicating full body, 'cm' indicating motion, 'np2' indicating two people, 'ba' representing back viewpoint, and 'med' indicating medium quality.

3.2 UCF-101

- Year : 2012
- Paper : UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild Actions [13]

3.2.1 Context and construction perspective

In the past, Action Recognition datasets are made artificially and often involving actors performing various actions in controlled environments. Consequently, traditional action recognition datasets often suffer from the lack of realistic context, which poses challenges for effectively addressing real-world problems. Therefore, the UCF-50 was initially introduced with the primary aim of addressing this issue through the incorporation of real-life videos which reflect the diversity of human actions represented in real-world scenarios.

Serving as an extension of the UCF50 dataset, UCF101 retained the original 50 action classes from its predecessor while introducing an additional 51 new classes. The aim was to refine and expand the range of actions captured in the dataset, contributing to a more comprehensive understanding of human activities.

Among the UCF family of action datasets, there are four version: UCFSports [8], UCF11 [8], UCF50, and UCF101. Of these, the dataset being referred to is the largest and most iconic.

3.2.2 Data collection methods

The videos in the dataset are primarily sourced from YouTube. Researchers begin by conducting searches related to the desired action categories. After that, they undergo the preprocess and then annotation procedures on the identified videos. Beside, these videos are selected to keep up various aspects such as camera angles, motion, spectator viewpoint, external conditions, object interactions, contextual backgrounds, and more.

3.2.3 Data distribution

The dataset contains of 13,320 video demonstrating of 101 actions. Each clip has a resolution of 320x240 pixels at 25 FPS.

Beside, this dataset is also include audio data for 50 action classes, providing valuable supplementary information to the videos. In each clip, it is separated into various variants of itself, such as reducing or increasing the length or adding

audio called augmented group. These variations aid the model in learning both similarities and differences between classes.

The train/test split uses the same method as mentioned above for HMDB51 which divided into three independent sets. Each split is carefully chosen to ensure the minimal correlation between them.

- Train : for each action, 70 random clips are used.
- Validation : no validation set.
- Test : for each action, 30 random clips are used.

3.2.4 Annotation

In the related paper, actions are split into action groups depending on their purpose: *Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments, Sports*. The portion of each groups are shown the figure below:

In the provided Figure *b*, the Sport group significantly dominates the entire dataset. As the author mentions, the distinctive motions which related to sports tend to achieve higher accuracy than others action groups when applying the SVM classisfication technique.

Within each set, there exist both training and testing components. Each file contains a list of video names utilized for the proposed purpose. These video names following to a specific rule in order of: *action class*, *augmented group*, and *clip number*.

For example, *v_ PlayingGuitar_ g25_ c07* indicates the clip number **07** shows up the action **Playing Guitar** in the augmented group **25**.

3.3 Sport-1M

- Year : 2014
- Paper : Large-scale Video Classification with Convolutional Neural Networks Actions [14]

3.3.1 Context and construction perspective

In recent years, the rapid growth of Artificial Intelligence has led to the evolution of image-related datasets. However, they were designed to satisfy old traditional Machine Learning methods, such as SVM [12] [13]. This method is now prone to being left behind by the newborn CNNs method, whose effectiveness was proven by AlexNet [15] in 2012. The problem arises concerning the equivalent metrics for those datasets. Therefore, a gigantic sports action dataset was developed to address this issue which is called Sports-1M.

3.3.2 Data collection methods

The videos are automatically collected from YouTube based on related sports tags. Subsequently, their associated links are stored in a text file for later download. Each clip ensures that the link is accessible before being saved.

3.3.3 Data distribution

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. The dataset is divided into training, validation, and testing sets in a 7:1:2 ratio, respectively.

- Train and Validation : 914,491 clips
- Test: 218,667 clips

3.3.4 Annotation

The annotation process was performed automatically by a special machine algorithm. It follows the rule of analyzing the text metadata around the dataset and finding correlations to its labels. First, it identifies associated videos with pre-defined tags and then uses weakly supervised learning to structure the categories into hierarchical levels. For example, in the node "Ball Sports," it can include "freestyle football" and "ball hockey."

Despite the automatic process of acquiring and annotating videos, the portion of duplicated frames are tiny (1755 out of more than 1 million clips). This is primarily attributed to the variability of frames within individual videos. Additionally, variations in actions may happened due to external factors such as camera angles or viewing distances.

Each label describes in detail the associated sports. Therefore, numerous unique and rarely seen sports are also found here. For example, a line containing a link and labels is found in the training set:
`https://www.youtube.com/watch?v=BWYPOToJu24` 436,431
The number "436" and "431" indicate its labels in order of "akido" and "grappling"

3.4 ActivityNet

- Year : 2015
- Paper : ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding Actions [16]

3.4.1 Context and construction perspective

Although various action recognition datasets have been released in recent times, only a small portion of them focus on daily activities. For instance, datasets like UCF101 [13] and THUMOS 14 [17] have their own category distributions, but they lack detailed organization and depth of levels, which limits the amount of information they provide. Therefore, ActivityNet is introduced to offer a solution with a large-scale dataset that provides a high level of specificity for human daily life in a hierarchical structure.

3.4.2 Data collection methods

According to the related paper, the data collection method follows three steps:

- **Internet's data gathering:** Due to the abundance of videos available on the Internet, text based queries are used to filter out the relevant videos.
- **Confirming the labels of untrimmed videos:** AMT workers verified those clips to determine whether they actually contain the action with the related label. As a result, untrimmed videos would be associated with at least one ground truth label.
- **Annotating videos in the hierarchical structure**: The AMT workers then proceed to trim each video to include only one activity. Afterwards, the trimmed videos are rechecked by others to ensure the required quality

3.4.3 Data distribution

It contains 849 hours from 27,801 videos of indoor actions with a total of 68.8 hours that appear 203 human-centric activities. Most of the videos have lengths ranging from 5 to 10 minutes at 30 FPS. Additionally, more than half of them have HD resolution quality (1280x720). The training, validation, and testing sets follow a split of 5:2.5:2.5 proportions.

- Train : 10,024 videos
- Validation : 4,926 videos
- Test: 5,044 videos

3.4.4 Annotation

ActivityNet activities follows a hierarchical structure which referred as "roots." These roots different areas of human behavior including: *Personal Care*, *Eating and Drinking*, *Household*, *Caring and Helping*, *Working*, *Socializing and Leisure*, and *Sports and Exercises*. Within each root category, there are more specific nodes that provide additional granularity. Finally, at the leaves, the dataset provides ground truth labels for individual activities. This hierarchical organization allows for a detailed understanding of the diverse range of behaviors captured in this dataset.

The annotation process divided into two different parts: *label part* and *link part*.

The *label part* provides the following: .1 Taxonomy. .2 Parent name. .2 Node name. .2 Node ID. .2 Parent ID.

Node ID and Parent ID have a close relationship to define where the node is in the hierarchy

The *link part* provides the following: .1 Database. .2 Youtube ID. .3 Duration. .3 Subset. .3 Resolution. .3 URL. .3 Annotations. .4 Segment. .4 Label.

3.5 Something Something V2

- Year : 2016
- Paper : The "something something" video database for learning and evaluating visual common sense Actions [**?**]

3.5.1 Context and construction perspective

The recent expansion of the Action Recognition dataset has led to a rapid increase in the number of available videos, particularly sourced from platforms like YouTube such as Sport-1M [14] and Youtube-8M [18]. However, the model's work associated with these datasets still involves combining features extracted from frames, and therefore becoming a "set of image" classification task. That's the reason even these datasets empower models to infer numerous actions depicted in videos, they often lack the contextual understanding of how different actions correlate with each other. Consequently, when an action is combined with various objects, it can potentially mislead the model since it diverges from its learned associations.

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 particular problem.

In the related paper [**?**], Something V1 emphasizes detailed interactions between human actions and objects, aiming to provide fine-grained videos that reflect real-world aspects. One year after, the V2 version was released in [**?**] and significantly increasing the number of videos and improving label quality.

3.5.2 Data collection methods

A group of workers from the crowdsourcing service Amazon Mechanical Turk (AMT) were asked to record and label clips. They recorded videos that demonstrated specific actions with the given labels. After the video is uploaded, it is divided into categories depends on its action, label, and object information. Afterward, they submitted the videos to an online platform, which underwent careful automatic quality checks. Each submission was then rechecked by other workers to make sure there is no mistaken.

3.5.3 Data distribution

In the newer V2 version, 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, compared to the previous 100px. Each clip has an average length of 2-6 seconds, demonstrating only a single action mentioned in the its label. Overall, there are 318,572 annotations, which involve 30,408 unique objects. The data is split into training, validation, and test sets with a ratio of 8:1:1. Additionally, measures are taken to ensure that videos created by a single actor are exclusively allocated to one of these three splits.

- Train : 168,913 clips
- Validation : 24,777 clips
- Test: 27,157 clips

3.5.4 Annotation

An interesting aspect of this dataset is its provision of "sentence-like" information alongside the focus on human actions. This augmentation allows for descriptions of how actors interact with objects. For instance, a video featuring the phrasal verb's action "moving away" might be presented as: "Moving a bag of popcorn away from the camera". By integrating object names with actions, the dataset offers detailed action descriptions.

Due to the unique labeling approach using "natural language," conventional one-hot encoding methods are not applicable. Consequently, it is recommended that model should be first pretrain on ImageNet to enhance the capture of distinctive object characteristics.

For videos in the training set marked by a unique ID, there is an annotated JSON file records detailed descriptions which includes Video IDs (containing only the training video's ID), completed labels (providing precise descriptions of the actor's actions), template categories (outlining the action skeleton), and placeholders (identifying objects used in the video).

For example, a line in JSON file follows:

- **id:** 217769
- **label:** moving a bag of popcorn away from the camera
- **template**: Moving [something] away from the camera
- **placeholders**: a bag of popcorn

In the testing set, each video is assigned its own true label, which corresponds to the template in the JSON file. For example, the video show that a person is moving a calculator away, it will have the true label "Moving [something] away from the camera".

3.6 AVA

- Year : 2018
- Paper : AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions [19]

3.6.1 Context and construction perspective

Popular action classification datasets such as KTH [4], Weizmann [5], HMDB51 [12], and UCF101 [13] only consist of short clips manually trimmed to capture a single action. Newer datasets like Sports-1M [14], YouTube-8M [18], Something-something [**?**], and Moments in Time [20] focus on larger-scale data and are often annotated automatically, leading to noisy annotations. Some recent studies, such as ActivityNet [16], THUMOS [17], and Charades [21], utilize a large number of videos containing multiple actions but only provide temporal annotations. Recognizing the limitations from their perspective, the authors of this paper proposed the AVA dataset.

3.6.2 Data collection methods

The process of constructing the AVA dataset involves five steps : action vocabulary generation, movie and segment selection, person bounding box annotation, person linking and action annotation.

- **Action vocabulary generation** : They generate action vocabulary base on three principle :

  Generality : Generic action in daily-life scenes, opposite of specific actions in a specific situation, for example, playing football on a football field, would be general or nonspecific actions in a general situation.

  Atomicity : Independent of interacted objects (e.g., hold without specifying what object to hold).

  Exhaustivity : The authors initialized list of actions using knowledge from previous datasets. They iterated through this list multiple times until it covered approximately 99% of the actions in the AVA dataset.

- **Movie and segment selection** : Raw video content is sourced from YouTube. The authors initially compile a list of top actors from various nationalities. Each actor is then searched using a YouTube query, retrieving up to 2000 results. The selected videos must fall under the categories of "film" or "television," have a duration of over 30 minutes, be uploaded at least one year prior, and have a minimum of 1000 views. Additionally, black and white videos, low-resolution videos, animated and cartoon content, gaming videos, as well as videos containing adult content, are excluded. Each video is partition into a 15-minute long segment and divided into 897 movie segments. The result is 430 videos.

- **Person bounding box annotation** : First, the bounding boxes are automatically detected using the Faster-RCNN [22] person detector, which significantly reduces the manual annotation time. Then, annotators re-annotate the missed boxes to ensure complete coverage. In the final step, incorrectly labeled boxes are marked and removed.
- **Person link annotation** : The bounding boxes are linked using person embeddings [23], and then the optimal matching with the Hungarian algorithm [24] is applied to match the boxes together. In order to increase accuracy, annotators remove false positive boxes in the next step.
- **Action annotation** : Recognizing the practicality that annotators may make labeling mistakes is unavoidable when dealing with up to 80 classes, the authors divided Action annotation into two stages: action proposal and verification. In the proposal stage, annotators are required to suggest action classes, and these proposals are verified by annotators in the verification stage.

3.6.3 Data distribution

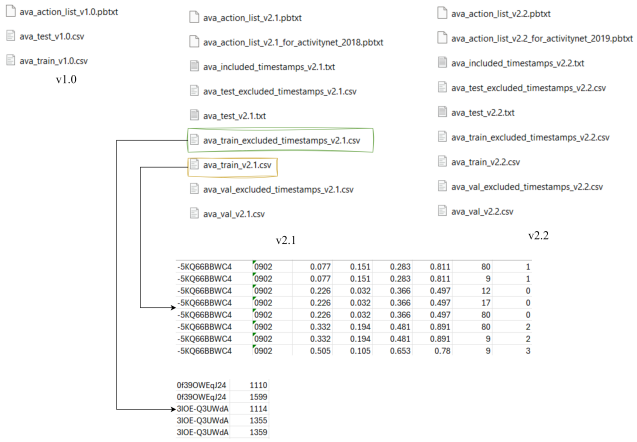
In total, there are 430 videos covering 80 classes, with each video contributing 15 minutes at a sample rate of 1Hz (meaning one frame per second, 897 segments per 15 minutes). The train/val/test ratio is divided as 55:15:30, respectively.

- Train : 235 videos, 211k segments.
- Val : 64 videos, 57k segments.
- Test : 131 videos, 118k segments.

3.6.4 Annotation

The AVA action dataset has four versions : v1.0, v2.0, v2.1 and v2.2. There are two key differences between AVA v2.2 and v2.1. Firstly, an additional round of human rating was carried out to include missing labels, resulting in a 2.5% increase in the total number of annotations. Secondly, box locations were corrected for a few videos that had aspect ratios significantly larger than 16:9. Regarding AVA v2.1, the only modification compared to v2.0 was the removal of a few duplicate movies. The class list and label map remained unchanged from v1.0. The differences between v1.0 and v2.0 is not mentioned by authors.

Figure 3 lists the annotation files provided across different versions, v2.0 is not mentioned because it has been replaced by v2.1. Unlike v1.0, v2.1, and v2.2 include additional action list files (60 classes as a subset of the 80 classes in AVA), which serve the purpose of the ActivityNet Challenge [16]. Additionally, there are included files and excluded files. Raters typically provided annotations at timestamps ranging from 902 to 1798, inclusive, in seconds, with a 1-second interval. Performance evaluation includes all these "included" timestamps, even those where raters indicated the absence of any action. However, for certain



Hình 3:

videos, specific timestamps were excluded from annotation due to raters flagging the corresponding video clips as inappropriate. Evaluation of performance does not take into account the "excluded" timestamps.

The format of a row is the following: video_id, middle_frame_timestamp, person_box, action_id, person_id , as described on the official website :

- video_id: YouTube identifier
- middle_frame_timestamp: the timestamp in seconds from the start of the YouTube video
- person_box: the bounding box coordinates of the person, given as the top-left (x_1, y_1) and bottom-right (x_2, y_2) points normalized with respect to the frame size. The coordinate range is from $(0.0, 0.0)$ at the top left to $(1.0, 1.0)$ at the bottom right.
- action_id: the identifier of an action class
- person_id: a unique integer that allows linking this box to other boxes depicting the same person in adjacent frames of the video.

3.7 Moments in Time

- Year : 2019
- Paper : Moments in Time Dataset: one million videos for event understanding [20]

3.7.1 Context and construction perspective

The authors want to create a high-coverage, highdensity, balanced dataset of hundreds of verbs depicting moments of a few seconds. To ensure top-notch data quality, datasets should encompass a wide range of topics, exhibit substantial sample diversity and density, and possess the capability to scale effectively.

3.7.2 Data collection methods

The construction of the Moments in Time dataset involves three main steps : Building a Vocabulary of Active Moments, Collection and Annotation.

- **Building a Vocabulary of Active Moments** : Retrieve the 4,500 most commonly used verbs from VerbNet, then perform clustering and select the most frequently used words from each cluster. Once a word is chosen, it is removed from all clusters to which it belongs. This process yields a result of 339 frequently used and semantically diverse verbs.
- **Collection data** : Authors conduct an Internet search by analyzing video metadata and crawling search engines from variety of different sources (Youtube, Flickr, Vine, Metacafe, Peeks, Vimeo, VideoBlocks, Bing, Giphy, The Weather Channel, and Getty-Images) to build a list of candidate videos for each class in the vocabulary. Each video will be randomly cut a 3-second section corresponding verb.
- **Annotation** : The AMT workers are shown a pair of videos and verbs and are asked to confirm whether the action is completed in the video or not. The confirmed samples are then annotated. Each HIT (a worker's annotation request) consists of 64 three-second videos related to a single verb, with 10 ground truth videos used for control purposes. Each HIT includes four initial questions that workers must answer correctly in order to proceed. Only the results from HITs that achieve a 90% or higher accuracy on the control videos are included in the dataset. Each video in the training set must be reviewed at least three times and must receive at least 75% confirmation from the AMT workers to be considered a positive sample. For the test and validation sets, the authors increase the minimum review frequency to four times and the confirmation rate to 85%.

3.7.3 Data distribution

3.7.4 Annotation

3.8 HACS

- Year : 2019
- Paper : HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization [25].

3.8.1 Context and construction perspective

Recent advancements in computer vision have been propelled by the increasing size of datasets. For image categorization, datasets have expanded significantly in just a few years. Caltech101, with only 9.1K examples, was quickly surpassed by the ImageNet dataset, which now boasts over 1.2M examples. Object detection has seen a similar trend, with Pascal VOC starting at 1.6K examples and the COCO dataset currently containing 200K images and 500K object-instance annotations. In the video domain, action recognition datasets have also experienced substantial growth. Older benchmarks like HMDB51, UCF101, and Hollywood2 consisted of only a few thousand examples, but newer datasets like Sports1M, Kinetics, and Moments-in-Time contain significantly more videos. However, the growth in action localization datasets has been limited. THUMOS, ActivityNet, AVA, and Charades, while valuable, do not possess comparable dataset sizes. To address this gap, the authors introduce a new video benchmark called Human Action Clips and Segments (HACS), motivated by the need for large-scale action datasets. HACS aims to provide a comprehensive dataset for action localization, facilitating the exploration of more advanced models in this domain.

3.8.2 Data collection methods

The authors employed 200 action labels (which is identical to that of the ActivityNet-v1.3 dataset) to search the YouTube video search engine, resulting in the retrieval of 890K potentially relevant videos. The number of videos per action class varied from 1100 to 6600. To ensure dataset quality, two types of de-duplication were performed. Firstly, duplicate videos within the HACS dataset were eliminated. Secondly, to ensure fair evaluation on other benchmarks, videos that overlapped with samples in the validation or test sets of datasets such as Kinetics, ActivityNet, UCF-101, and HMDB-51 were also removed. The authors observed that manually annotating the start and end of action segments in untrimmed videos is time-consuming. Therefore, they propose sampling short clips from the videos. They experimented with three different sampling methods and selected the best-performing one. Subsequently, the selected samples are annotated.

3.8.3 Data distribution

HACS dataset utilizes a taxonomy consisting of 200 action classes, which aligns with the ActivityNet-v1.3 dataset. It comprises a total of 504K videos obtained from YouTube, with each video strictly shorter than 4 minutes, and an average length of 2.6 minutes. The dataset includes 1.5M clips, each lasting 2 seconds, which are sparsely sampled using a combination of uniform randomness and consensus/disagreement of image classifiers. Among these clips, 0.6M are annotated as positive samples, while 0.9M are labeled as negative samples.

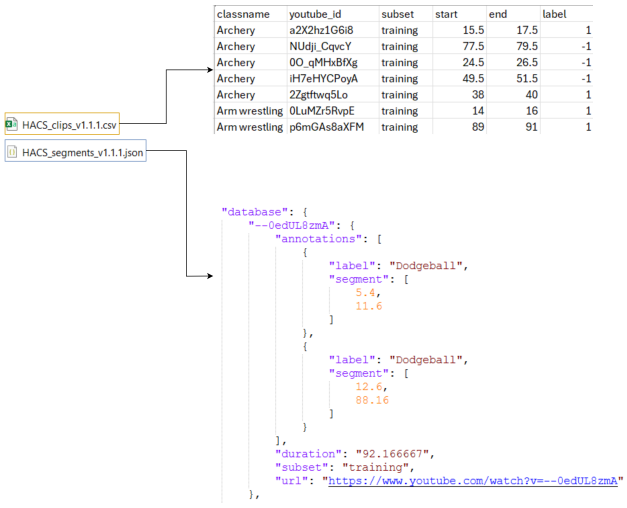
- **HACS Clips** : The dataset is divided into training, validation, and testing sets. The training set consists of 1.4M clips sampled from 492K videos,

while the validation and testing sets contain 50K clips each, sampled from 6K videos .

- **HACS Segments** : For a subset of 50K videos (38K for training, 6K for validation, and 6K for testing), manual boundaries are collected to define the start, end, and action label of each action segment within the video. It is ensured that all videos in this subset contain at least one action segment.

3.8.4 Annotation

There are two annotation files available for HACS Clips and HACS Segments. The file structure is explained in Figure 4, where the label section has values of 1 and -1 corresponding to positive and negative samples, respectively.



Hình 4:

3.9 HVU

- Year : 2020.
- Paper : Large Scale Holistic Video Understanding [26].

3.9.1 Context and construction perspective

The authors argue that training ConvNets to understand videos with a single label is insufficient to describe the content and hinders the learning of ConvNets. Furthermore, existing datasets also lack in this aspect. Recognizing this, the authors propose HVU - a dataset aiming to provide a multi-label and multi-task large-scale video benchmark with a comprehensive list of tasks and annotations for video analysis and understanding.

3.9.2 Data collection methods

- **Data collection** : They utilize existing video sources such as YouTube-8M, Kinetics-600, and HACS. By using these datasets as sources, they can avoid copyright issues and ensure that none of the test videos from existing datasets are included in the training set of HVU.

- **Anotation** : The authors employ a two-stage framework for the annotation process in HVU. In the first stage, they utilize the Google Vision API and Sensifai Video Tagging API to obtain rough annotations for the videos. These APIs predict around 30 tags per video, with a relatively low probability threshold (around 30%) to avoid false rejects of tags. The selected tags are chosen from a dictionary of nearly 8,000 words, resulting in approximately 18 million tags for the entire dataset. In the second stage, human verification is applied to remove any potentially mislabeled noisy tags and add any missing tags that were not captured by the APIs. This involves reviewing and correcting the annotations based on human judgment. As a result, the human annotation step produces approximately 9 million tags for the entire dataset, encompassing around 3,500 different tags. This two-stage annotation process helps improve the accuracy and quality of the annotations in the HVU dataset.

3.9.3 Data distribution

The HVU dataset comprises a total of 572k videos. These videos are divided into different sets, with 481k video clips in the training set, 31k clips in the validation set, and 65k clips in the test set. The dataset consists of trimmed video clips, where the duration of the videos varies, with a maximum length of 10 seconds. HVU does not solely rely on a single action class but instead includes multiple tags. It is organized into six main categories, which are scene, object, action, event, attribute, and concept. The dataset consists of a total of 3,143 tags.

3.9.4 Annotation

HVU provides three annotation files. One file contains tags along with their respective categories. The other two files, namely train and val, have the same structure (see fig 5), including columns for Tags, youtube_id, time_start, and time_end. To access the test video and missing videos, it is necessary to fill out a form available on the author's official GitHub page.

3.10 AViD

- Year : 2020

- Paper : AViD Dataset: Anonymized Videos from Diverse Countries [27]

Hình 5:

3.10.1 Context and construction perspective

The authors argues that current datasets have three shortcomings:

- **Geographical bias and limited geographic coverage**: Videos are often sourced in a geographically biased manner, resulting in limited geographic diversity in the coverage of the datasets.
- **Lack of temporal consistency in large datasets**: Due to the data collection process primarily relying on YouTube as the main source, the videos may be unstable over time. For example, in the first year of its release, Kinetics-400 dataset had 10% of its videos removed by YouTube.
- **Difficult accessibility**: Downloading tools often encounter request errors, and furthermore, some videos may not be available in certain countries.

3.10.2 Data collection methods

- **Action Definition**: Sử dụng lại action vocab từ Charades, Moments in Time và Kinetics. Loại bỏ các action có liên quan đến khuôn mặt do AViD làm mờ khuôn mặt của các actor. Thêm nữa, quá trình annotation cũng đề xuất ra thêm các class action. Kết quả của quá trình này gồm 887 action class. Ngoài ra nhóm tác giả cũng đề xuất hệ thống phân cấp cho các action class.
- **Data Collection**: The action classes were translated into 22 different languages and searched across various platforms. The group of authors ensured that all these videos comply with the Creative Commons license. The outcome of this process is a collection of 800K videos.
- **Data Annotation**: The combination of manual (Amazon Mechanical Turk) and automated (I3D model) methods was employed.
- **Data Cleaning and Filtering**:

3.10.3 Data distribution

- **Train clips** : 410k.
- **Validation clips** : no mentioned.
- **Test clips** : 40k.

3.10.4 Annotation

Classification-only annotation. Structure : [video_name] : [class_name] .


```
{"btlieszn.mp4": "rope pushdown", "znnbyryk.mp4": "fireworks", "qqkftvms.mp4": "tasting wine", "cbfpndym.mp4": "playing field hockey", "avgicwxb.mp4": "kickboxing", "ftqmjqim.mp4": "shaking head",
"earkjgvj.mp4": "busking", "linwlalp.mp4": "none", "lstzthpb.mp4": "milking goat", "roujhucf.mp4": "playing cymbals", "asesywto.mp4": "construction", "msoswyza.mp4": "playing marbles", "sgsfkbay.mp4": "combat",
"smuwbzqv.mp4": "ice climbing", "jrjwggpw.mp4": "sailing", "huzmamzu.mp4": "nightclub", "gvltuhxh.mp4": "backflip (human)", "ztvkmrpw.mp4": "sign language interpreting", "mfvzeydn.mp4": "belly dancing",
"yxvvsqex.mp4": "archaeological excavation", "aejxqnrh.mp4": "curling (sport)", "aycyxjcp.mp4": "washing dishes", "aytwtfym.mp4": "none", "owibjlrx.mp4": "driving tractor", "myqkhcxw.mp4": "none",
"dbhwlmml.mp4": "bouncing on trampoline", "cbcxrgua.mp4": "none", "rjratdzd.mp4": "none", "hjphknhr.mp4": "playing guitar", "pczpzlcn.mp4": "bodyweight exercise", "jlchqofo.mp4": "card throwing",
"irlfnigl.mp4": "none", "fkeohbsr.mp4": "none", "dadaduhj.mp4": "tai chi", "mgwzfhtm.mp4": "dancing ballet", "vuglxxvw.mp4": "making jewelry", "qmuwjwdr.mp4": "none", "urypibdm.mp4": "playing harmonica",
"wulqlwfu.mp4": "shooting off fireworks", "mkxsfsdx.mp4": "rock climbing", "qdlnpvxo.mp4": "protesting", "yxhfvvek.mp4": "none", "vmzyxjjq.mp4": "clay pottery making", "nkuljpiy.mp4": "news anchoring",
"fmacirqw.mp4": "speed skating", "swzlvpfg.mp4": "heavy equipment", "vnnjgisw.mp4": "steering car", "ofmxphxu.mp4": "none", "vqdzjekc.mp4": "playing clarinet", "hmzcvfds.mp4": "carving ice", "pjtfceue.mp4":
"abseiling", "oculeick.mp4": "rock climbing", "kzbayoqt.mp4": "getting a piercing", "bgensupo.mp4": "squeezing orange", "saqniyyu.mp4": "making bed", "tycoybym.mp4": "peeling banana", "fynwswrx.mp4": "building
lego", "mkplvhlw.mp4": "golf putting", "ktmozszz.mp4": "base jumping", "avvkmlmf.mp4": "robot dancing", "elqvyhpl.mp4": "playing cricket", "dnqpyixk.mp4": "none", "bkeptqgq.mp4": "sled dog racing",
"ofupwavw.mp4": "grooming cat", "kvjpwohu.mp4": "land sailing", "yncusucw.mp4": "shaking head", "cptcfsxy.mp4": "playing laser tag", "zararrqs.mp4": "waterfall", "twwwmpct.mp4": "feeding fish", "xpflnake.mp4":
"leatherworking", "ivkamamf.mp4": "polishing metal", "jzxkfjbm.mp4": "changing gear in car", "iticzlys.mp4": "casting fishing line", "kjtgjter.mp4": "playing piano", "rroypwzt.mp4": "wakeboarding",
"jrkoaqpp.mp4": "home improvement", "ogjmurjq.mp4": "eating ice cream", "mquzdiht.mp4": "none", "nkebcphe.mp4": "unloading truck", "llqwmhfk.mp4": "shaving head", "preqriqu.mp4": "none", "uimtgrhp.mp4":
"canoeing or kayaking", "mfiysnfg.mp4": "casting fishing line", "qoswtllx.mp4": "cooking chicken", "pxbfzeeo.mp4": "sanding floor", "ninenvix.mp4": "cleaning windows", "ynrvngkn.mp4": "archaeological
excavation", "xqiwtgwi.mp4": "dumpster diving", "sgcybrjh.mp4": "fly tying", "jtunflkc.mp4": "playing monopoly", "qugxjvmd.mp4": "none", "wzeobknu.mp4": "installing carpet", "khiykswf.mp4": "surfing water",
"evwlcdyq.mp4": "rafting", "ysfedjhc.mp4": "bodyweight exercise", "exfocrnr.mp4": "playing cello", "tgxdtozb.mp4": "deadlifting", "sumgckbs.mp4": "needle felting", "frbkuvir.mp4": "stretching leg",
"ffexzibb.mp4": "scuba diving", "tcusdmxp.mp4": "eating cake", "yxlrbyqz.mp4": "laying bricks", "jmuetszi.mp4": "gargling", "ttbyensl.mp4": "splashing", "pvpopgtp.mp4": "crawling baby", "vmseihdh.mp4": "none",
```


Hình 6:

3.11 THUMOS

- Year : 2015
- Paper : The THUMOS challenge on action recognition for videos “in the wild”

3.11.1 Context and construction perspective

THUMOS Challenge is a competition held from 2013 to 2015, serving as a benchmark for action recognition.

3.11.2 Data collection methods

THUMOS base on UCF101 with extra data.

- **Action Definition**: Reuse action list of UCF101.
- **Data Collection**: Background : search on Internet.
- **Data Annotation**: Manually annotate with predefined-criteria.
- **Data Cleaning and Filtering**:

3.11.3 Data distribution

- **Train** : 13k.
- **Validation** : 21k.
- **Test set** : 5.6l
- **Background set** : aka negative set, 3k .

3.11.4 Annotation

Base on UCF101.

3.12 YouTube-8M

- Year : 2016
- Paper : YouTube-8M: A Large-Scale Video Classification Benchmark [18]

3.12.1 Context and construction perspective

3.12.2 Data collection methods

- **Action Definition**:
- **Data Collection**:
- **Data Annotation**:
- **Data Cleaning and Filtering**:

3.12.3 Data distribution

3.12.4 Annotation

3.13 Charades

- Year : 2016
- Paper : Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding [21]

3.13.1 Context and construction perspective

The group of authors argues that everyday human activities are often mundane, whereas filming scenes or creating videos on social media platforms like YouTube always require engaging content to attract viewers. As a result, everyday life scenes and actions have not been fully utilized and there are no specific methods to collect them. The authors explain their video collection process and provide the research community with their achievement, the Charades dataset.

3.13.2 Data collection methods

- **Action Definition**: Use triplet verb-noun-scene to generate a script. This resulting 157 action class.
- **Data Collection**: AMT worker capture video base on generated script.
- **Data Annotation**:
- **Data Cleaning and Filtering**:

3.13.3 Data distribution

- **Train** : 7,985.
- **Validation** : not mentioned.
- **Test set** : 1,863.

3.13.4 Annotation

3.14 NTU RGB+D

- Year : 2016
- Paper : NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis [28]

3.14.1 Context and construction perspective

3.14.2 Data collection methods

- **Action Definition**:
- **Data Collection**:
- **Data Annotation**:
- **Data Cleaning and Filtering**:

3.14.3 Data distribution

3.14.4 Annotation

4

- Year :
- Paper :

4.0.1 Context and construction perspective

4.0.2 Data collection methods

- **Action Definition**:
- **Data Collection**:
- **Data Annotation**:
- **Data Cleaning and Filtering**:

4.0.3 Data distribution

4.0.4 Annotation

5

- Year :
- Paper :

5.0.1 Context and construction perspective

5.0.2 Data collection methods

- **Action Definition**:
- **Data Collection**:
- **Data Annotation**:
- **Data Cleaning and Filtering**:

5.0.3 Data distribution

5.0.4 Annotation

Tài liệu

[1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[2] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 06 2010.

[3] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[4] 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, pp. 32–36 Vol.3, 2004.

[5] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2, pp. 1395–1402 Vol. 2, 2005.

[6] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3d exemplars," in *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–7, 2007.

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[8] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[9] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2929–2936, 2009.

[10] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1996–2003, 2009.

[11] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Computer Vision – ECCV 2010* (K. Daniilidis, P. Maragos, and N. Paragios, eds.), (Berlin, Heidelberg), pp. 392–405, Springer Berlin Heidelberg, 2010.

[12] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *2011 International Conference on Computer Vision*, pp. 2556–2563, 2011.

[13] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," 2012.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, p. 84–90, may 2017.

[16] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, 2015.

[17] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah, "The thumos challenge on action recognition for videos "in the wild"," *Computer Vision and Image Understanding*, vol. 155, p. 1–23, Feb. 2017.

[18] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," 2016.

[19] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "Ava: A video dataset of spatio-temporally localized atomic visual actions," 2018.

[20] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfruend, C. Vondrick, and A. Oliva, "Moments in time dataset: one million videos for event understanding," 2019.

[21] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," 2016.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016.

[23] L. Wu, C. Shen, and A. van den Hengel, "Personnet: Person re-identification with deep convolutional neural networks," 2016.

[24] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics (NRL)*, vol. 52, 1955.

[25] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "Hacs: Human action clips and segments dataset for recognition and temporal localization," 2019.

[26] A. Diba, M. Fayyaz, V. Sharma, M. Paluri, J. Gall, R. Stiefelhagen, and L. V. Gool, "Large scale holistic video understanding," 2020.

[27] A. Piergiovanni and M. S. Ryoo, "Avid dataset: Anonymized videos from diverse countries," 2020.

[28] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," 2016.