A Survey of Video Datasets for Grounded Event Understanding

Kate Sanders, Benjamin Van Durme

Johns Hopkins University

Q: How do we develop and evaluate systems that can answer complex video understanding requests like this?

```
How are these videos related to the recent European election,
and what are the most recent developments? Who are the main
political figures of interest, and what do they do in the video?
```

A: We need vision-language agents that can understand multimodal events the way humans do. But what resources do we have?

WHEN REVIEWING EVENT-CENTRIC VIDEO BENCHMARKS, WHAT SHOULD BE CONSIDERED?

SEMANTIC CONTENT

- What is the domain of events depicted?
- How many events are depicted?
- Are there multiple videos for each event?
- Were the events natural or staged?

DATA PRESENTATION

- What is the purpose of the videos?
- What modalities are recorded?
- Are the videos compilations?
- Are text or graphics added in post?

EVENT MODELING

Learn more in the following panels . . .

We produced a spreadsheet of 105 video datasets with relevant event-centric info. See QR code.

Dataset Name	Website Link	Paper Link	Year	Task	Event Types	Event/Video Cov.	Instertiation	Genre	Text/Graphics	Modalities	Annotations	Motivation/Contribution (excerpt from paper)
MMSRV	https://github.com/https://arxiv.org/a		2022	Video Classificat	Short social med	50k short videos	Both	Social media vid	F	Video	Class labels, emotion labels, audi	We present 3MASRV, a multilingual, multimodal and multi-aspect, ex
ActivityNet	http://activity-net.https://www.cv-fo		2015	Action Recognitio	Daily human acti	200 classes, 10C	Unknown	Internet videos	F	Video	Class labels, class temporal boun	current computer vision algorithms are still severely limited in terms o
ActivityNet-Capti	https://cs.stanfor https://arxiv.org/a		2017	Captioning	Daily human acti	20K videos, 100	Unknown	Internet videos	F	Video	Video captions, temporal intervals	We introduce the task of dense-captioning events, which involves bot
ActivityNet-Entit	https://github.com https://openaccess		2019	Captioning	Daily human acti	15K videos, 52K	Unknown	Internet videos	F	Video	ActivityNet-Captions annotations -	Video description is one of the most challenging problems in vision a
ADGA	https://cs.stanfor https://arxiv.org/a		2021	QA	daily indoors act	160W unbalance	Scripted	Dataset content	F	Video	Spatio-temporal scene graphs, Qa	We present Action Genome Question Answering (AGQA), a new ben
ADLA	https://catalog.id _		2023	Scene Parsing	Current events	291 videos	Both	Internet news vi	T	Video, text article	Event-relevant labels - Event slots	To develop a multi-hypothesis semantic engine to generate explicit an
Animal Kingdom	https://sukritsy.gith https://openacce		2022	Action recognitio	Diverse multi-lab	30K video sequ	Unknown	Animal behavior	F	Video	Action and behavior labels, temp	existing animal behavior datasets have limitations in multiple aspects
Assembly101	https://assembly https://openacce		2022	Action recognitio	Fine- and coarse	4321 videos of p	Scripted	Dataset content	F	Video	Coarse- and fine-grained action la	Assembly101 is a new procedural activity dataset featuring 4321 vide
AVA Actions	https://research.r https://openacce		2018	Action recognitio	Atomic actions o	80 atomic visual	Scripted	Movies, TV	F	Video	Action labels, bounding boxes, en	This paper introduces a video dataset of spatio-temporally localized A
Breakfast	http://serre-lab.cl https://deepai.org/m		2014	Hierarchical acti	Cooking activitie	77 hours of video	Scripted	Multiview dataset	F	Video	Hierarchical class labels, tempora	we collected a large dataset of daily cooking activities: The dataset in
CCV	https://www.ee.c https://www.ee.c		2011	Classification	Daily life	9,317 YouTube v	Unknown	Internet videos	F	Video	Class labels	Existing corpora for video analysis lack scale and/or content diversity,
Charades	https://prior.allen.http://ai2-website		2016	Action Recognitio	daily indoors act	157 classes, 984	Scripted	Dataset content	F	Video	(Verb, proposition, noun) triplets,	computer vision methods need to be trained from real and diverse ex
Charades-Ego	https://prior.allen https://arxiv.org/a		2018	Action Recognitio	daily indoors act	7860 videos, 15	Scripted	First and third pe	F	Video	Charades action labels	Several theories in cognitive neuroscience suggest that when people
ChinaVisOpen	https://vuo.qmul.https://arxiv.org/a		2023	Captioning/Retri	Random internet	50k videos	Unknown	Internet videos	F	Video	Multilingual video titles, video tags	While the state-of-the-art multimodal learning networks have shown s
COIN	https://coin.nelnet https://arxiv.org/a		2019	Hierarchical Acti	nursing & caring	11,827 videos, 1	Scripted	Instructional vide	F	Video	Class labels, temporal intervals	However, most existing datasets for instructional video analysis have
CondensedMovi	https://www.robo https://arxiv.org/a		2020	Retrieval	General movie a	3K movies, 33,0	Scripted	Movies	F	Video	High level semantic scene descri	Our objective in this work is long range understanding of the narrativ
CrossTask	https://github.io/ https://arxiv.org/a		2019	Task ordering	Instruction steps	83 tasks, 4.7K vi	Scripted	Class labels, temporal intervals		Video	Class labels, temporal intervals	weakly supervised learning may be easier if a model shares compon
DiDeMo	https://github.io/ https://arxiv.org/a		2017	Action Localizati	Personal daily a	10K videos	Unknown	Internet videos	F	Video	Clip captions, temporal intervals	current video datasets do not include pairs of localized video segmen
DiscoVerb	http://www.eecs.q https://www.ecva.n		2019	Hierarchical Acti	Hierarchical alm	198k+ videos, 4M	Natural	Disney videos	F	Video	Hierarchical class labels	A procedure, RESOLVND, is proposed to quantify and minimize repet
DramaQA	https://dramaqa.i https://arxiv.org/a		2020	QA	Character-centri	17,983 QA pairs	Scripted	TV drama	F	Video	QA pairs, character bounding box	we propose a novel video question answering (Video QA) task, Dram

WHAT BENCHMARKS INVOLVE IMPLICIT VISUAL EVENT UNDERSTANDING?

BENCHMARK TASKS

ACTIVITY RECOGNITION

SCENE GRAPH GENERATION

CAPTIONING AND RETRIEVAL

QUESTION ANSWERING

UNIQUE DOMAINS

TV SHOWS AND MOVIES

NEWS CONTENT AND BROADCASTS

HOW ARE EVENTS IMPLICITLY MODELED IN VIDEO BENCHMARKS?

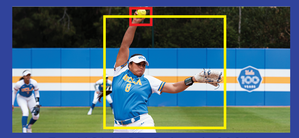
VISUAL SEMANTIC STRUCTURES

PROPOSITIONAL

E.g., Sports-1M (Karpathy et al., 2014)

$Throwing$

DAVIDSONIAN



E.g., HOMAGE (Rai et al., 2021)

$\exists e, x, y. [\ Throwing(pitcher(x), baseball(y), e)]$

NEO-DAVIDSONIAN

E.g., MovieGraphs (Vicol et al., 2018)

$\exists e, x, y, z. [\ Throwing(e) \wedge Agent(pitcher(x), e) \wedge Obj(baseball(y), e) \wedge Team(Tampa\ Bay\ Rays(z), e) \ldots$

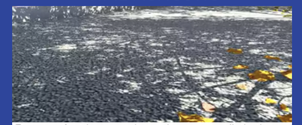
STRING-BASED

E.g., VaTeX (Wang et al., 2020)

The US pitcher throws the last ball of the third inning.

TEMPORAL STRUCTURES

TIME-AGNOSTIC

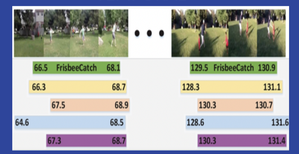


Runners feet in a sneakers close up. realistic three dimensional animation.

WEBVID-2M (Bain et al., 2021)

$\exists e.\ Event(e)$

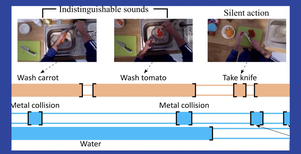
TEMPORAL



THUMOS (Long et al., 2020)

$\exists e, t. [\ Event(e) \wedge At(e, t_1, t_2)]$

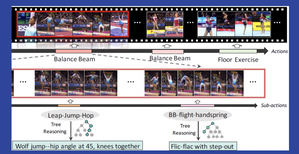
COMPOSITIONAL



EpicKitchens (Damen et al., 2018)

$\exists e, t. [\ Event(e_x) \wedge At(e_x, t_{1(x)}, t_{2(x)}) \wedge Event(e_y) \wedge At(e_y, t_{1(y)}, t_{2(y)})]$

HIERARCHICAL

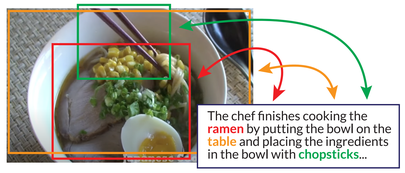


FineGym (Shao et al., 2020)

$\exists e_x, e_y, e_z. [\ Event(e_x) \wedge Event(e_y) \wedge Event(e_z) \wedge \mathbf{Comprises}(e_x, e_y, e_z)]$

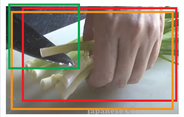
VIDEO EVENT EXTRACTION TASKS

MULTIMODAL EVENT EXTRACTION (CHEN ET AL., 2021)



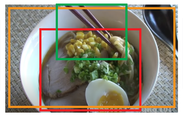
EVENT	COOKING
FOOD	RAMEN
INSTRUMENT	CHOPSTICKS
PLACE	TABLE
. . .	. . .

VISUAL SEMANTIC ROLE LABELING (SADHU ET AL., 2021)



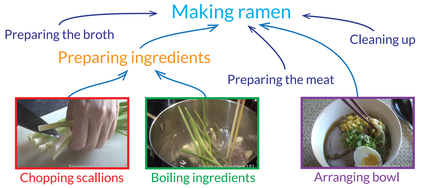
EVENT	COOKING
FOOD	SCALLIONS
INSTRUMENT	KNIFE
PLACE	TABLE

PREPARATION STEP FOR



EVENT	COOKING
FOOD	RAMEN
INSTRUMENT	CHOPSTICKS
PLACE	TABLE

MULTIMODAL EVENT HIERARCHY EXTRACTION (AYYUBI ET AL., 2022)



The chef begins cooking the ramen by preparing the ingredients. They first chop scallions and other vegetables, and then bring water to a boil and put the vegetables in the pot.

...

Finally, they arrange the ingredients in the bowl, and serve the dish.

KEY SURVEY TAKEAWAYS

- Natural language and formal structures
- Temporality and hierarchy modeling
- Uncertainty is a powerful modeling tool
- Multi-faceted diversity strengthens datasets and benchmarks