Generic Event Boundary Detection: A Benchmark for Event Segmentation

Mike Zheng Shou[1,2], Stan Weixian Lei[2], Weiyao Wang[1], Deepti Ghadiyaram[1], Matt Feiszli[1]

[1]Facebook AI [2]National University of Singapore

Abstract

This paper presents a novel task together with a new benchmark for detecting generic, taxonomy-free event boundaries that segment a whole video into chunks. Conventional work in temporal video segmentation and action detection focuses on localizing pre-defined action categories and thus does not scale to generic videos. Cognitive Science has known since last century that humans consistently segment videos into meaningful temporal chunks. This segmentation happens naturally, without pre-defined event categories and without being explicitly asked to do so. Here, we repeat these cognitive experiments on mainstream CV datasets; with our novel annotation guideline which addresses the complexities of taxonomy-free event boundary annotation, we introduce the task of ***Generic Event Boundary Detection (GEBD)*** *and the new benchmark* ***Kinetics-GEBD****. We view GEBD as an important stepping stone towards understanding the video as a whole, and believe it has been previously neglected due to a lack of proper task definition and annotations. Through experiment and human study we demonstrate the value of the annotations. Further, we benchmark supervised and un-supervised GEBD approaches on the TAPOS dataset and our Kinetics-GEBD. We release our annotations and baseline codes at* ***CVPR'21 LOVEU Challenge****:* `https://sites.google.com/view/loveucvpr21`.

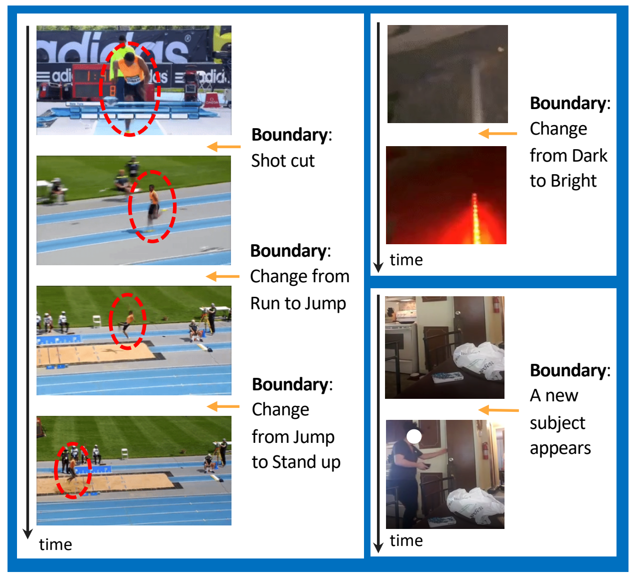

Figure 1: Examples of generic event boundaries: 1) A long jump is segmented at a shot cut, then between actions of Run, Jump and Stand up (dominant subject in red circle). 2) color / brightness change 3) new subject appears.

1. Introduction

Cognitive science tells us [49] that humans perceive video in terms of "events" (goal-directed sequences of actions, like "washing a car" or "cooking a meal"), and further, people segment events naturally and spontaneously while perceiving video, breaking down longer events into a series of shorter temporal units. However, mainstream SOTA video models [47, 48, 13, 7, 28, 12] still commonly process short clips (e.g. 1s long), followed by some kind of pooling operation to generate video-level predictions.

Recent years have seen significant progress in temporal action detection [8, 9, 14], segmentation [24, 3, 21, 11] and parsing [35, 38] in videos. Despite this, we have not seen major developments in modeling long-form video. The cognitive science suggests that one underlying deficit is event segmentation: unlike our SOTA models, humans naturally divide video into meaningful units and can reason about these units. In contrast to our current methods build upon limited sets of predefined action classes, humans perceive a broad and diverse set of segment boundaries **without any predefined target classes**.

To enable machines to develop such ability, we propose a new task called **Generic Event Boundary Detection (GEBD)** which aims at localizing the moments where humans naturally perceive event boundaries. As Fig. 1 shows, our event boundaries could happen at the moments where

	#videos	#segments	#boundaries	video domain	boundary cause	#boundary classes	#Annotations per video
THUMOS'14	2700	18K	36K	sports	action	20	1
ActivityNet v1.3	27801	23K	46K	in-the-wild	action	203	1
Charades	67000	10K	20K	household	action	157	1
HACS Segments	50000	139K	278K	in-the-wild	action	200	1
AVA	214	197K	394K	movie	action	80	1
EPIC-Kitchens	432	39K	79K	kitchen	action	2747, open-vocab	1
EPIC-Kitchens-100	700	89K	179K	kitchen	action	4053, open-vocab	1
TAPOS Instances	16294	48K	33K	sports	action	open-vocab	1
Kinetics-GEBD (raw)	55351	1771K	1498K	in-the-wild	generic	taxonomy-free	4.93
Kinetics-GEBD (clean)	54691	1561K	1290K	in-the-wild	generic	taxonomy-free	4.94

Table 1: Comparing our Kinetics-GEBD with other video boundary datasets. Our Kinetics-GEBD has the largest number of temporal boundaries (e.g. 32x of ActivityNet, 8x of EPIC-Kitchens-100), spans a broad spectrum of video domains in the wild in contrast to sports or kitchen centric, is open-vocabulary rather than building on a pre-defined taxonomy, contains boundaries caused by not only action change but also generic event change, and has almost 5 annotations per video to capture human perception differences and therefore ensure diversity. Note that for ActivityNet and TAPOS, since ground truths of test set are withheld, we do not include #segments and #boundaries of their test sets.

the action changes (e.g. Run to Jump), the subject changes (e.g. a new person appears), the environment changes (e.g. suddenly become bright), for example.

To annotate ground truths of such taxonomy-free event boundaries, the common strategies used by the existing temporal tasks with pre-defined taxonomy do not work:

1. Existing tasks require us to manually define each target class carefully i.e. its semantic differentiators compared to other classes. But it is impractical to enumerate and manually define all candidate generic event boundary classes.

2. The existing tasks typically focus on shot and action boundaries, neglecting other generic event boundaries as the examples shown in Fig. 1 like change of subject.

In this paper, we propose to follow cognitive experiments [49] in annotating event boundaries on computer vision datasets. We choose the popular Kinetics [20] dataset as our video source and construct a new event segmentation benchmark **Kinetics-GEBD**. The marked boundaries are relatively consistent across different annotators; the main challenge raising ambiguity is the level of detail. For example, one annotator might mark boundaries at the beginning and end of a dance sequence, while another might annotate every dance move. We develop several novel principles in design annotation guideline to ensure consistent level of detail across different annotators while explicitly capturing the human perception differences with a multi-review protocol.

Our new GEBD task and benchmark will be valuable in:

1. Immediately supporting applications like video editing, summarization, keyframe selection, highlight detection. Event boundaries divide a video into natural, meaningful units and can rule out unnatural cuts in the middle of a unit, for example.

2. Spurring progress in long-form video; GEBD is a first step towards segmenting video into meaningful units and enabling further reasoning based on these units.

In summary, our contributions are four-fold:

- A new task and benchmark, Kinetics-GEBD, for detecting generic event boundaries without the need of a predefined target event taxonomy.

- We propose novel annotation task design principles that are effective yet easy for annotators to follow. We disambiguate what shall be annotated as event boundaries while preserving diversity across individuals in the annotation.

- We benchmark a number of supervised and unsupervised methods on the TAPOS [38] dataset and our Kinetics-GEBD.

- We demonstrate the value of our event boundaries on downstream applications including video-level classification and video summarization.

2. Related Work

Temporal Action Detection or localization methods attempt to detect the start time and end time for action instances in untrimmed, long videos. Standard benchmarks include THUMOS [19], ActivityNet [1], HACS [53], etc. All of them target a list of specified action classes and manually define the criteria for determining the start point and end point of each action, preventing annotations at scale.

Numerous methods have been developed for temporal action detection [8, 9, 14, 31, 40, 42, 4, 30, 54, 51, 33]. Notably, many of them contain a temporal proposal module which solves a binary classification problem analogous to foreground-background segmentation. "Background" segments contain no pre-defined action classes. However, many other generic events could appear in background segments, and segmenting generic events is the main focus in this paper.

Temporal Action Segmentation [24, 3, 21, 11, 26, 2, 27, 37, 18] means labeling the action classes in every frames. Some popular benchmarks are 50Salads [45], GTEA [26], Breakfast [21, 22], MERL Shopping [43], etc. Another task called **Temporal Action Parsing** was recently proposed in [38]; parsing aims to detect the temporal boundaries for segmenting an action into sub-actions. This is more closely related to our current work. However, these annotations and methods are also developed for pre-defined action classes only, not generic boundaries.

Shot Boundary Detection is a classical task to detect shot transitions due to video editing such as scene cuts, fades/dissolves, and panning. Some recent works are [5, 15, 46, 39, 44]. These shot boundaries are well-defined and an overcomplete set is easy to detect since the changes between shots are often significant. In this paper, we also annotate and detect shot boundaries in our Kinetics-GEBD benchmark; however, the main novelty lies in event boundaries which are useful for breaking generic videos into semantically-coherent subparts.

3. Definition of the GEBD Task

3.1. Task Definition

GEBD localizes the moments where humans naturally perceive taxonomy-free event boundaries that break a longer event into shorter temporal segments. To obtain ground truth annotations, we begin with the cognitive experiments' guideline [49] which achieved consistent boundaries marked by different annotators. However, the cognitive experiments typically cover a limited number of scenarios in simple videos, e.g. a single actor, free of distractions from the event of interest. We target diverse, natural human activity videos like Kinetics [20] which contain multiple actors, background distractions, different levels of detail in both space and time, etc. Thus, there is more ambiguity about what are the event boundary positions.

3.2. Principles for Designing Annotation Guideline

To overcome these ambiguities in natural videos, we arrived at the following design principles throughout multiple iterations of improving annotation guidelines.

(a) Detail in space: Focus on the dominant subject. In order to avoid getting distracted by background events, annotators shall focus on the salient subject performing the event. The subject could be a person, a group, an object, or a collection of objects, depending on the video content.

(b) Detail in time: Find event boundaries at "1 level deeper" granularity compared to the video-level event. Given a video, it can be segmented at different temporal granularities. For example, the event boundaries of a long jump video could be 1) coarse: Long Jump starts / ends, or 2) intermediate: Long Jump is broken into running, jumping, and landing, or 3) fine: every foot step. All variants are legitimate segmentations. We embrace this ambiguity to a limited degree: we instructed annotators to mark boundaries "1 level deeper" than the video-level event, and provided some examples but no precise definition of "1 level". Sometimes there is no one single video-level event; yet the merit of this principle is to ensure the segmented subparts are at the same level of granularity. This technique can be recursively applied to the segmented subparts when finer granularity is desired. With this principle implemented, we find that humans can reliably agree on event boundaries without the need of a hand-crafted event boundary taxonomy.

(c) Diversity of perception: Use multi-review. Sometimes people have different interpretations of "1 level deeper" and go slightly deeper or coarser. For example, in a video of two consecutive Long Jump instances, some might segment two instances of long jump, while others would segment the running and jumping units. In practice, we consider both are correct and find that one video usually has at most 2-3 such possible variations due to the human perceiving differences rather than the ambiguity of task definition. Thus, to capture such diversity, we assign 5 annotators for each video based the rule of thumb in user experience research.

(d) Annotation format: Timestamps vs Time Ranges. The above principles clarify when to mark an event boundary. The remaining question is marking where. Following previous works, we can accommodate some ambiguity in "where" during evaluation by varying an error tolerance threshold; more details can be found in Sec. 3.3. We provide two options for marking an event boundary: 1) A single "Timestamp", typically used for instantaneous change (e.g. the moment when jumping begins in long jump). 2) A time "Range", typically used for short yet gradual change e.g. the interval between the end of landing and the start of standing up. More detailed can be found in Supp.

More details of our annotation guideline for Kinetics-GEBD (e.g. our own annotation interface, task rejection criteria, annotation format) can be found in Supp.

3.3. Evaluation Protocol

As described in Sec. 3.2, a boundary can be either a timestamp or a short range. If it is a range, we represent it by its middle timestamp during evaluation. Thus, our evaluation task is to measure the discrepancy between the detected timestamp and the ground truth timestamp, regardless of their types or semantic meanings. To measure the discrepancy between timestamps, we follow previous works such as temporal parsing of an action instance [38] and online detection of action start [41] and use the Relative Distance (Rel.Dis.) measurement. Inspired by the Intersection-over-Union measurement, Rel.Dis. is the error between the detected and ground truth timestamps, divided by the length of the corresponding whole action instance. Given a fixed threshold for Rel.Dis., we can determine whether a detection is correct (i.e. $\leq$ threshold) or incorrect (i.e. $>$ threshold), and then compute precision, recall, F1 score for the whole dataset. Note that duplicated detection for the same boundary is not allowed. Also, Each rater's annotation is used separately. A detection result is compared against each rater's annotation and the highest F1 score is treated as the final result. We have also explored other metrics. Detailed discussions can be found in Supp.

4. Benchmark Creation: Kinetics-GEBD

4.1. Video Sources

Our Kinetics-GEBD Train Set contains 20K videos randomly selected from Kinetics-400 Train Set [20]. Our Kinetics-GEBD Test Set contains another 20K videos randomly selected from Kinetics-400 Train Set. Our Val Set contains all 20K videos in Kinetics-400 Val Set.

4.2. Annotator Training

To ramp up a new annotator, we provide a training curriculum consisting of a cascade of 5 training batches. Each training batch contains 100 randomly sampled Kinetics videos with some reference annotations. We make it clear to the annotator that different people may segment the same video in different ways, thus our provided annotations are only for reference. Once a batch is done and before moving the annotator to the next batch, we will review its annotations for all 100 videos and provide specific feedback regarding errors made due to misunderstanding or misconduct of the guideline. Overall, we do observe steady improvements over training batches for each new annotator.

4.3. Quality Assurance

We present our detailed quality assurance mechanism in Supp. Briefly, annotators were trained on 5 cascaded batches of 100 videos, with a QA mechanism before they worked on real jobs. Typical issues early in training included misunderstanding of the tool or guidelines, as well as annotating too much or too little detail. Training videos were rated on a scale of 1 (good), 2 (minor errors like inaccurate timestamps), and 3 (bad, typically misunderstanding of guidelines). Raters progressed to real jobs when their average rating was deemed sufficient. In practice, the performance of an annotator is satisfying and acceptable if its average rating is below 1.3.

4.4. Common Characteristics of Boundary Causes

Cognitive studies [6] suggest that event boundaries can be characterized by several high-level causes. Throughout our pilot annotation tasks for refining guideline, we confirmed such finding and arrived at the following high-level causes of event boundaries: (1) **Change of Subject**: new subject appears or old subject disappears and such subject is dominant. (2) **Change of Object of Interaction**: the subject starts to interact with a new object or finishes with an old object. (3) **Change of Action**: an old action ends, or a new action starts. Note that this characteristic includes when the subject changes physical direction (e.g. a runner suddenly changes direction) and when the same action is being performed multiple times (e.g. several consecutive push-up instances). (4) **Change in Environment**: significant changes in color or brightness of the environment or the dominant subject (e.g. a light is turned on, illuminating a previously darker environment). Further, **Shot Change** boundaries are also common in Kinetics videos. Thus, we also annotate shot boundaries and the instructions can be found in Supp. In a video of multiple shots, the target granularity for event boundaries is 1 level deeper than the corresponding shot-level event. Sometimes an event boundary might be due to **Multiple** coupled causes or **Others**. As the distribution shown in Fig. 2: **Others** is negligible; **Change of Action** is the most common cause. Note that the actions leading to boundaries in our dataset are *much more generic and diverse* than the pre-defined taxonomies in the current CV action datasets.

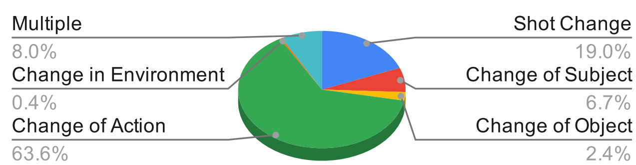

Figure 2: Distribution of boundary causes on Kinetics-GEBD Val.

4.5. Annotation Results Summary and Analysis

Annotation capacity. In total, around 40 qualified annotators were trained to annotate our Kinetics-GEBD. The average speed is around 5mins per video per annotator.

Statistics of #annotations received. Recall that each video

#Annotations	0	1	2	3	4	5
#videos	101	141	203	342	805	18166
Per. (%)	0.51	0.71	1.03	1.73	4.07	91.94

Table 2: For our Kinetics-GEBD Val set, #annotations received per video vs. #videos and its percentage .

is annotated by 5 annotators. Annotators can reject a video due to the reasons stated in Supp. Table 2 shows that most videos receive all 5 annotations without rejection.

The extent of consensus for GEBD annotation. Given the construction of the dataset, a natural question is "how consistent are the annotations?". Adopting the protocol in Sec. 3.3, for the same video, we treat one annotation as ground truth and another annotation as detection result. Since we expect consistent annotations to have very close boundaries in time, we do not use relative distance; instead, we evaluate F1 score based on the absolute distance between two boundaries, varying the threshold from 0.2s to 1s with a step of 0.2s, and calculate the average F1 score. By averaging the F1 score over all pairs of annotations for the same video, we can obtain its consistency score. If all raters make very similar annotations, the consistency score will be high i.e. towards 1; otherwise low i.e. towards 0.

Fig. 3 shows that the majority of videos have consistency scores higher than 0.5. This indicates that given our designed task definition and annotation guideline, humans are able to reach decent degree of consensus, taking into account the factors that (1) often due to different human perception manners, a video can have multiple correct segmentations, and (2) sometimes annotators make mistakes.

To understand how the frequency of annotation mistakes (i.e. annotation quality) correlates with the consistency score, in Table 3, we randomly sample 5 non-rejection videos for each consistency score range and conduct manual auditing according to the protocol in Sec. 4.3 to get the average rating for each range. As the consistency score becomes low, the rating gets worse. Recall that the cutoff for the rating to determine qualified annotators is 1.3, which corresponds to 0.5 consistency score here.

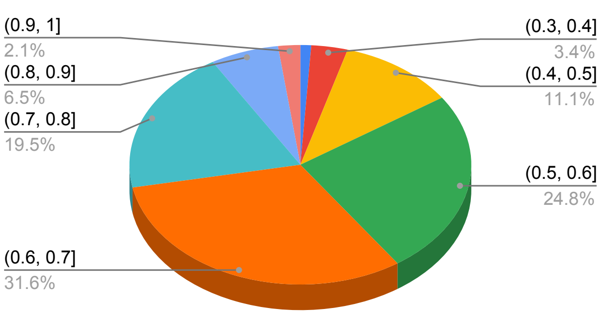

Figure 3: The number of videos percentage (below line) for each range of the consistency score (above line) on our Kinetics-GEBD Val set when the video is not rejected by any annotators.

Consistency	(0.4,0.5]	(0.5,0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,1]
Rating	1.4	1.24	1.20	1.16	1.04

Table 3: Average audit rating vs. average F1 consistency score on our Kinetics-GEBD Val set.

4.6. Post-processing for the Annotations

Given the raw annotations, we conduct the following steps to construct our Kinetics-GEBD benchmark. (1) To ensure annotation quality and remove very ambiguous videos, we exclude videos that have lower than 0.3 consistency score. (2) To capture the diversity of human perception, we only keep videos that receive at least 3 annotations. During evaluation, the detection is compared against each ground truth annotation and the highest F1 score is treated as the final result. (3) For each annotation, if two boundaries are very close (i.e. less than 0.1s), we merge them into one. Note that this includes the case that one Timestamp boundary falls into a Range, or one Range boundary overlaps with another Range boundary. We remove any boundaries from the initial and final 0.3s of each video. More details in Supp.

4.7. Statistics

For the raw Kinetics-GEBD annotation, the average number of boundaries per video per annotation is 5.48 (std dev 2.76, range [1, 33]). The average time between boundaries is 1.47s (std dev 1.24, range [0,10.01]). The number and average length of the time-range boundaries is 265K and 0.71s. The number of timestamp-only boundaries is 1232K.

For the Kinetics-GEBD benchmark (after post-processing raw annotations), the average number of boundaries per video per annotation is 4.77 (std dev 2.24, range [0,14], distribution plot as Fig. 4(a)). The average time between boundaries is 1.65s (std dev 1.25, range [0.023, 10.08], distribution plot as Fig. 4(e)).

Furthermore, the left column of Fig. 4 shows the distribution of #boundaries per video per annotation, #boundaries per video and duration per segment, respectively. To show how these compared on the base Kinetics-400 classes, we rank all Kinetics classes from high to low and highlight 3 classes, as shown in the right column of Fig. 4.

5. Experimental Results of GEBD Methods

5.1. Dataset

In addition to our Kinetics-GEBD, we also experiment on the recent TAPOS dataset [38] containing Olympics sport videos with 21 actions. The training set contains 13,094 action instances and the validation set contains 1,790 action instances. The authors manually defined

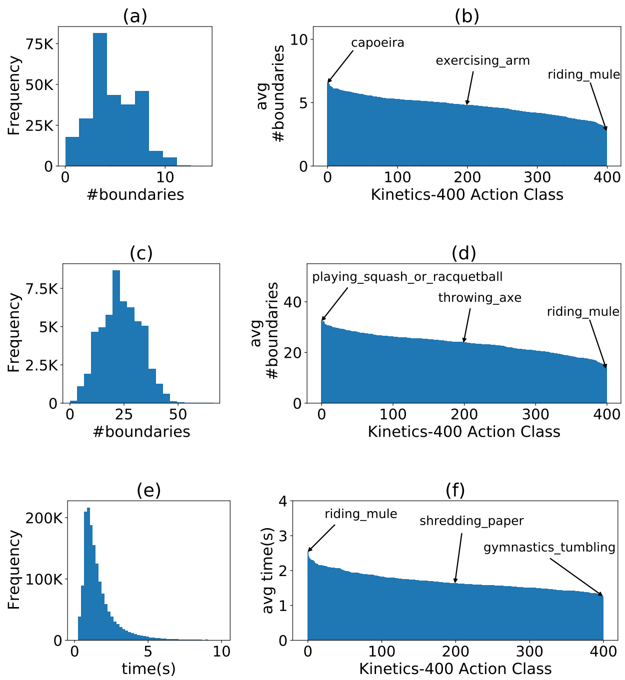

Figure 4: Statistics on Kinetics-GEBD. #boundaries per video per annotation: (a) distribution (b) average over each Kinetics class and then sorted by class; #boundaries per video: (c) distribution (d) average over each Kinetics class and then sorted by class; duration per segment: (e) distribution (f) average over each Kinetics class and then sorted by class.

how to break each action into sub-actions during annotation. While not taxonomy-free, the TAPOS boundaries between sub-actions are analogous to GEBD action boundaries. Thus, we can re-purpose TAPOS for our GEBD task by trimming each action instance with its action label hidden (can be as long as 5mins) and conducting GEBD on each action instance. Note that in TAPOS only 1 rater's annotation has been released and thereby used as ground truth.

5.2. Supervised Methods for GEBD

We directly quote the results of supervised methods from [38] on TAPOS (i.e. the below **#1-3**). Since [38] has not published codes, we implement the below **#4-6** methods by ourselves on our Kinetics-GEBD:

#1. Temporal parsing model: **TransParser** [38] proposes a pattern miner trained with a local loss based on the sub-action boundary supervision and a global loss trained with the action instance label supervision.

#2. Temporal action segmentation models: Connectionist Temporal Modeling (**CTM**) [17] and Iterative Soft Boundary Assignment (**ISBA**) [10] are supervised by the order of occurrence of a set of pre-defined sub-actions.

#3. Action boundary detection model: Temporal Convolution Network (**TCN**) [24, 31] trains a binary classifier to distinguish the frames around boundaries against other frames.

#4. Pairwise boundary Classifier (PC): At each candidate boundary position time t, we use the same backbone network to extract a feature pair: the average feature of frames before and the average feature of frames after t. We conduct global pooling over space for each feature and then we concatenate these two paired features together as the input to a linear binary classifier, which is trained to predict the probability of time t is a boundary. **PC** is trained end-to-end to fine-tune the backbone network pre-trained on ImageNet; training with the backbone fixed does not converge. We watershed the probability sequence to obtain internals above 0.5. Each internal's center is treated as an event boundary.

#5. Temporal action proposal model: to understand how well an class-agnostic action boundary proposal model can detect generic event boundaries, we train a BMN model [29] on THUMOS'14 [19] and test it on Kinetics-GEBD to generate action proposals. We denote **BMN** as treating both the start and end of each action proposal as event boundary. Alternatively, since one intermediate step in BMN is to evaluate two probability scores of respectively being action start and end, we watershed each probability sequence to obtain internals above 0.5 and treat the center of each internal as an event boundary. We take the union of all these centers and denote this method as **BMN-StartEnd**.

#6. Cross-dataset GEBD method **TCN-TAPOS**: to confirm the need of Kinetics-GEBD which is more challenging than TAPOS, we conduct testing on Kinetics-GEBD using the TCN model trained on TAPOS.

5.3. Unsupervised Methods for GEBD

This direction is intriguing because it can potentially handle any kind of events, without the need to annotate a large amount of event boundary labels.

#1. SceneDetect[1]: an online popular library for detecting classical shot changes.

#2. PA - Random: we randomly swap the detection results of the below **PA** method among all videos. The position of each boundary is mapped to the new video with its relative position in the original video unchanged.

#3. PredictAbility (PA): Event Segmentation Theory indicates that the moment people perceive event boundary is where future activity is least predictable [23, 36, 52]. This motivates us to develop a **PA**-based method which first 1) computationally assesses the predictability score over time and then 2) locates the event boundaries by detecting the local minima of the predictability sequence.

1) *Predictability Assessment*: To quantify the predictability at time t, we compute the average feature of

[1] https://github.com/Breakthrough/PySceneDetect

Rel.Dis. threshold		0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	avg
Unsuper.	SceneDetect	0.035	0.045	0.047	0.051	0.053	0.054	0.055	0.056	0.057	0.058	0.051
	PA - Random	0.158	0.233	0.273	0.310	0.331	0.347	0.357	0.369	0.376	0.384	0.314
	PA	**0.360**	**0.459**	**0.507**	**0.543**	**0.567**	**0.579**	**0.592**	**0.601**	**0.609**	**0.615**	**0.543**
Super.	ISBA	0.106	0.170	0.227	0.265	0.298	0.326	0.348	0.369	0.382	0.396	0.302
	TCN	0.237	0.312	0.331	0.339	0.342	0.344	0.347	0.348	0.348	0.348	0.330
	CTM	0.244	0.312	0.336	0.351	0.361	0.369	0.374	0.381	0.383	0.385	0.350
	TransParser	0.289	0.381	0.435	0.475	0.500	0.514	0.527	0.534	0.540	0.545	0.474
	PC	**0.522**	**0.595**	**0.628**	**0.646**	**0.659**	**0.665**	**0.671**	**0.676**	**0.679**	**0.683**	**0.642**

Table 4: F1 results on TAPOS for various supervised and unsuperivsed GEBD methods.

Rel.Dis. threshold		0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	avg
Unsuper.	SceneDetect	0.275	0.300	0.312	0.319	0.324	0.327	0.330	0.332	0.334	0.335	0.318
	PA - Random	0.336	0.435	0.484	0.512	0.529	0.541	0.548	0.554	0.558	0.561	0.506
	PA	**0.396**	**0.488**	**0.520**	**0.534**	**0.544**	**0.550**	**0.555**	**0.558**	**0.561**	**0.564**	**0.527**
Super.	BMN	0.186	0.204	0.213	0.220	0.226	0.230	0.233	0.237	0.239	0.241	0.223
	BMN-StartEnd	0.491	0.589	0.627	0.648	0.660	0.668	0.674	0.678	0.681	0.683	0.640
	TCN-TAPOS	0.464	0.560	0.602	0.628	0.645	0.659	0.669	0.676	0.682	0.687	0.627
	TCN	0.588	0.657	0.679	0.691	0.698	0.703	0.706	0.708	0.710	0.712	0.685
	PC	**0.625**	**0.758**	**0.804**	**0.829**	**0.844**	**0.853**	**0.859**	**0.864**	**0.867**	**0.870**	**0.817**

Table 5: F1 results on Kinetics-GEBD for various supervised and unsuperivsed GEBD methods.

frames preceding and the average feature of frames succeeding t. Then, we compute their squared L2 norm feature distance to obtain the inverse predictability $\phi(t)$; lower distance implies greater predictability.

2) *Boundaries from Predictability*: Given $\phi(t)$, a natural method is to propose temporal boundaries at the local maxima of ϕ. This is similar to the classical blob detection problem, and thus we apply the classical Laplacian of Gaussian (LoG) filter [32] to our 1D temporal problem. We apply the 1D LoG filter to compute $L(t) = \mathrm{LoG}(\phi(t))$, and compute its derivative $L'(t)$. We detect temporal boundaries at the negative-to-positive zero-crossings of L', which correspond to local maxima of ϕ.

5.4. Implementation Details

The following settings are used for all experiments conducted by ourselves unless explicitly specified otherwise: 2 GP100 NVIDIA cards are used. For each video, we sample 1 frame for every 3 frames. The inputs are RGB images resized to 224x224. To make fair comparisons, all models implemented by ourselves, i.e. **PC**, **TCN**, **TCN-TAPOS**, **PA**, **BMN**, **BMN-StartEnd**, build on ResNet-50 [16] backbone. **PC** is trained end-to-end while others simply use the off-the-shelf ImageNet pretrained feature. Our **PC**, **TCN**, **TCN-TAPOS** and **PA** all use 5 frames before and 5 frames after a candidate boundary as the model input. For **PA**, we tune the sigma in the LoG filter on the Train set and set it to 15. During evaluation, we follow TAPOS [38] to vary the Relative Distance (Rel.Dis.) threshold indicated in Sec. 3.3 from 5% to 50% with a step of 5%.

5.5. Results Comparisons

TAPOS val set F1 results are shown in Table 4. Detailed results of precision and recall are in Supp. The predictability-based **PA** method is clearly better than the random guess. It is quite encouraging to see that our unsupervised method **PA** even outperforms all previous supervised methods i.e. ISBA, TCN, CTM, TransParser. **SceneDetect** achieves high precision while quite low recall because it only fires at the very salient boundaries.

Kinetics-GEBD val set F1 results are shown in Table 5. Detailed results of precision and recall are in Supp. Among unsupervised methods, **PA** is clearly better than shot change detection method **SceneDetect** and the random guess **PA - Random**, in particular when the threshold is strict on Kinetics-GEBD. Comparing **PA** with the supervised method **TCN** which also uses the same fixed backbone feature, the gap is not large, indicating that un-supervised or semi-supervised GEBD methods are worthwhile researching in the future. **PC** clearly outperforms others, indicating that the event boundaries cannot be comprehensively represented by off-the-shelf feature while can be better learned by the backbone. For the class-agnostic action proposal methods, directly detecting action proposals (i.e. **BMN**) is not a good GEBD approach but accessing the probability of being boundary (i.e. **BMN-StartEnd**) is effective. **BMN-StartEnd** is still worse than **PC** due to only detecting action change boundaries while ignoring other generic event boundaries like subject change. For the similar reason, on Kinetics-GEBD, a GEBD model trained on TAPOS (i.e. **TCN-TAPOS**) underperforms the same model directly trained on Kinetics-GEBD (i.e. **TCN**). These again confirm the challenging nature of generic event boundaries and the need of our new benchmark Kinetics-GEBD.

6. Applications of Video Event Boundaries

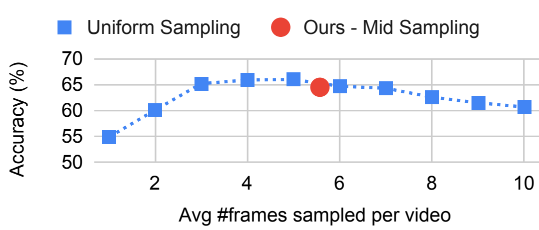

Figure 5: To classify a video, it is difficult to tell what is the optimal number of frames for uniform sampling. Our event boundaries provide cue about how many frames shall be sampled.

6.1. Video-level classification

We test the classification accuracy on videos that receive at least 3 annotations in Kinetics-GEBD Val set. We use the public implementation[2] of the TSN [50] model which uniformly samples K frames, applies ResNet-50 backbone on each frame, and finally average the predictions to get the video-level prediction. Fig. 5 shows that the video-level classification accuracy for uniform sampling (blue curve) increases and then decreases as K varies from 1 to 10. Thus, give a video, how can we determine K?

Despite GEBD is not designed to select discriminative frames, *our boundaries provide cue about how to set K in uniform sampling in order to achieve high classification accuracy* Based on our annotated boundaries, we can break the video into segments and each segment might only need one frame to be sampled. To validate this hypothesis, we sample the middle frame of each segment. Fig. 5 shows that this (the red dot) uses in average 5.5 frames per video and achieves accuracy close to the best achieved by uniform sampling. This is useful in practice when the video content is diverse and thus we do not know what is the best K.

[2]https://github.com/mit-han-lab/temporal-shift-module

6.2. Video summarization

Our temporal boundaries provide a natural way to select keyframes for video summarization. We conduct the following two user study tasks to compare **Ours** (sample the middle frame of subparts) and **Uniform** (uniformly sample the same number of frames as Ours). In Task 1, we randomly sample videos from Kinetics-GEBD Val. In Task 2, we select the videos that the frame distance between **Ours** and **Uniform** are the largest. Each task involves around 200-250 videos.

For each video in both task, 20 users are asked "which set of keyframes better summarize the video comprehensively?" and shall vote one out of three options: (1) Set 1 is better; (2) Set 2 is better; (3) Tie (both good/bad summarization). Table 6 shows the percentage of different options winning at vote-level and at the video-level (e.g. out of 20 votes for the same video, if #votes for (1) is the highest, Set 1 wins). We can see that for random samples **Ours** is clearly better than **Uniform** and for samples of large disparity **Ours** significantly outperforms **Uniform**.

Percentage (%)		Uniform	Ours	Tie
Task 1: random samples	Vote-level	33.9	**40.9**	25.1
	Video-level	38.3	**43.7**	17.8
Task 2: large disparity	Vote-level	12.6	**73.0**	14.3
	Video-level	6.0	**90.0**	4.0

Table 6: User study results for video summarization.

7. Conclusion

In this paper, we have introduced the new task of GEBD and resolved ambiguities in the annotation process. A new benchmark, Kinetics-GEBD, has been created along with novel designs for annotation guidelines and quality assurance. we benchmark supervised and un-supervised GEBD approaches on the TAPOS dataset and our Kinetics-GEBD.

We believe our work is an important stepping stone towards long-form video understanding and hope it will enable future work in learning based on temporal event structure. In the future, we plan to address scene changes which usually happen in much longer videos (e.g. move from kitchen to bathroom in 30mins long ADL [34] videos, move from street to restaurant in hours long UT-Ego [25] videos).

8. Acknowledgement

Mike Shou and Stan Lei are supported by the National Research Foundation, Singapore under its NRFF award NRF-NRFF13-2021-0008.

References

[1] Activitynet challenge 2016. http://activity-net.org/challenges/2016/, 2016. 2

[2] Sathyanarayanan N Aakur and Sudeep Sarkar. A perceptual prediction framework for self supervised event segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[3] Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Joint discovery of object states and manipulation actions. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1, 3

[4] Humam Alwassel, Fabian Caba Heilbron, and Bernard Ghanem. Action search: Spotting actions in videos and its application to temporal action localization. In *Proceedings of the European Conference on Computer Vision*, 2018. 2

[5] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Shot and scene detection via hierarchical clustering for reusing broadcast video. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 2015. 3

[6] Roger G Barker and Herbert F Wright. Midwest and its children: The psychological ecology of an american town. 1955. 4

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[8] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[9] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S Davis, and Yan Qiu Chen. Temporal context network for activity localization in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1, 2

[10] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 6

[11] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 3

[12] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1

[14] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Cascaded boundary regression for temporal action detection. In *the British Machine Vision Conference*, 2017. 1, 2

[15] Michael Gygli. Ridiculously fast shot boundary detection with fully convolutional neural networks. In *IEEE International Conference on Content-Based Multimedia Indexing*, 2018. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 7

[17] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *Proceedings of the European Conference on Computer Vision*, 2016. 6

[18] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[19] Y.-G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014. 2, 6

[20] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2, 3, 4

[21] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1, 3

[22] Hilde Kuehne, Juergen Gall, and Thomas Serre. An end-to-end generative framework for video segmentation and recognition. In *Proceedings of IEEE Winter Applications of Computer Vision Conference*, 2016. 3

[23] Christopher A Kurby and Jeffrey M Zacks. Segmentation in the perception and memory of events. *Trends in cognitive sciences*, 2008. 6

[24] Colin Lea, Austin Reiter, René Vidal, and Gregory D. Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *Proceedings of the European Conference on Computer Vision*, 2016. 1, 3, 6

[25] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 8

[26] Peng Lei and Sinisa Todorovic. Temporal deformable residual networks for action segmentation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[27] Jun Li, Peng Lei, and Sinisa Todorovic. Weakly supervised energy-based learning for action segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3

[28] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1

[29] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action pro-

posal generation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 6

[30] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017. 2

[31] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision*, 2018. 2, 6

[32] Tony Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 1998. 7

[33] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[34] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition*, 2012. 8

[35] Hamed Pirsiavash and Deva Ramanan. Parsing videos of actions with segmental grammars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1

[36] Jeremy R Reynolds, Jeffrey M Zacks, and Todd S Braver. A computational model of event segmentation from perceptual prediction. *Cognitive science*, 2007. 6

[37] Alexander Richard, Hilde Kuehne, and Juergen Gall. Action sets: Weakly supervised action segmentation without ordering constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[38] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Intra- and inter-action understanding via temporal action parsing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 3, 4, 5, 6, 7

[39] Hong Shao, Yang Qu, and Wencheng Cui. Shot boundary detection algorithm based on hsv histogram and hog feature. In *International Conference on Advanced Engineering Materials and Technology*. Atlantis Press, 2015. 3

[40] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[41] Zheng Shou, Junting Pan, Jonathan Chan, Kazuyuki Miyazawa, Hassan Mansour, Anthony Vetro, Xavier Giro-i Nieto, and Shih-Fu Chang. Online action detection in untrimmed, streaming videos-modeling and evaluation. In *Proceedings of the European Conference on Computer Vision*, 2018. 4

[42] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[43] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[44] Tomáš Souček, Jaroslav Moravec, and Jakub Lokoč. Transnet: A deep network for fast detection of common shot transitions. *arXiv preprint arXiv:1906.03363*, 2019. 3

[45] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738, 2013. 3

[46] Shitao Tang, Litong Feng, Zhanghui Kuang, Yimin Chen, and Wei Zhang. Fast video shot transition localization with deep structured models. In *Asian Conference on Computer Vision*. Springer, 2018. 3

[47] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 2015. 1

[48] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 1

[49] Barbara Tversky and Jeffrey M Zacks. Event perception. *Oxford handbook of cognitive psychology*, 2013. 1, 2, 3

[50] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision*, 2016. 8

[51] Zehuan Yuan, Jonathan C Stroud, Tong Lu, and Jia Deng. Temporal action localization by structured maximal sums. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[52] Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. Event perception: a mind-brain perspective. *Psychological bulletin*, 2007. 6

[53] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[54] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2