Are current long-term video understanding datasets long-term?

Ombretta Strafforello
TU Delft, TNO
o.strafforello@tudelft.nl

Klamer Schutte
TNO
klamer.schutte@tno.nl

Jan van Gemert
TU Delft
j.c.vangemert@tudelft.nl

Abstract

Many real-world applications, from sport analysis to surveillance, benefit from automatic long-term action recognition. In the current deep learning paradigm for automatic action recognition, it is imperative that models are trained and tested on datasets and tasks that evaluate if such models actually learn and reason over long-term information. In this work, we propose a method to evaluate how suitable a video dataset is to evaluate models for long-term action recognition. To this end, we define a long-term action as excluding all the videos that can be correctly recognized using solely short-term information. We test this definition on existing long-term classification tasks on three popular real-world datasets, namely Breakfast, CrossTask and LVU, to determine if these datasets are truly evaluating long-term recognition. Our study reveals that these datasets can be effectively solved using shortcuts based on short-term information. Following this finding, we encourage long-term action recognition researchers to make use of datasets that need long-term information to be solved.

1. Introduction

Many interesting actions happening in the real world are long-term. That is, they are composed of several short sub-actions, that we refer to as *short-term actions*. For an action to be *long-term*, we deem that recognizing a single-short term action is not enough, and reasoning about the order and the relationship of short-term actions is required. Two examples of long-term actions, shown in Figure 1, are *winning a soccer game* and *shoplifting in the supermarket*. To understand which team is winning a soccer game, it is necessary to recognize and count the goals scored since the beginning of the game. For the other example, recognizing if a person is shoplifting, it is necessary to observe a person storing a product in their pocket *and* leaving the supermarket without paying. In both examples, it is not possible to recognize the actions without reasoning on multiple ordered short-term actions.

Achieving automatic long-term action recognition is important because it can be used to solve real-world problems, from analyzing sports videos, to understanding movies and recognizing threats in surveillance footage. To make it possible, we need purpose-built computer vision models, that are trained and evaluated on datasets that need long-term reasoning to be solved. While working on long-term action recognition, we notice that every video in the Breakfast dataset [25], a go-to choice in long-term video understanding research [16, 17, 27, 47], contains short-term actions that map to a single long-term action. This implies that accurately recognizing a short-term action in a Breakfast video should be sufficient to infer the corresponding long-term action. We analyze the short-term actions of another popular instructional video dataset, CrossTask [49], and find the same occurrence in 97.72% of its primary tasks videos. We illustrate our statistics on the short-term action occurrences in Figure 2. Since deep learning models are known to use shortcuts to solve classification tasks [13], the models trained and tested on these datasets might learn to exploit short-term information, without encoding any long-term relations.

Motivated by this finding, we propose a method to diagnose whether a long-term dataset is suitable to study long-term action recognition, or can be solved using solely short-term information. To this end, we define two requirements for an action to be long-term: (1) The action is *recognizable only from multiple short-term actions* and not from a single short-term action. (2) The action maps to a *single label*. The first requirement makes long-term action recognition impossible without reasoning over an extended time span. Models that lack this capability, for example based on straightforward pooling operations over time [40], cannot recognize long-term actions. The second requirement leads to discarding multi-label action recognition datasets, like Charades [32], MultiTHUMOS [45] and EPIC-Kitchens [11], as long-term action datasets. In these datasets, the task is to recognize each short-term action contained in the videos. This task could be solved by classifying each short-term action one at a time, while here we are interested in the case where the classification can be made only after reasoning over multiple short-term actions together.

Figure 1: Example of truly long-term actions. *Top:* Who is winning this soccer game?[1], *Bottom:* Is this person shoplifting in the supermarket?[2]. In both cases, it is not possible to answer correctly without considering multiple short-term actions together, their order and relations over time. To understand who is winning the soccer game, it is necessary to recognize and count the goals scored since the beginning of the game. To recognize shoplifting, it is not enough to see a person putting a product in their pocket: also the short-term action *leaving without paying* needs to occur.
[1]Source: YouTube; [2]Source: YouTube from movie *Un povero ricco*, by Pasquale Festa Campanile (1983).

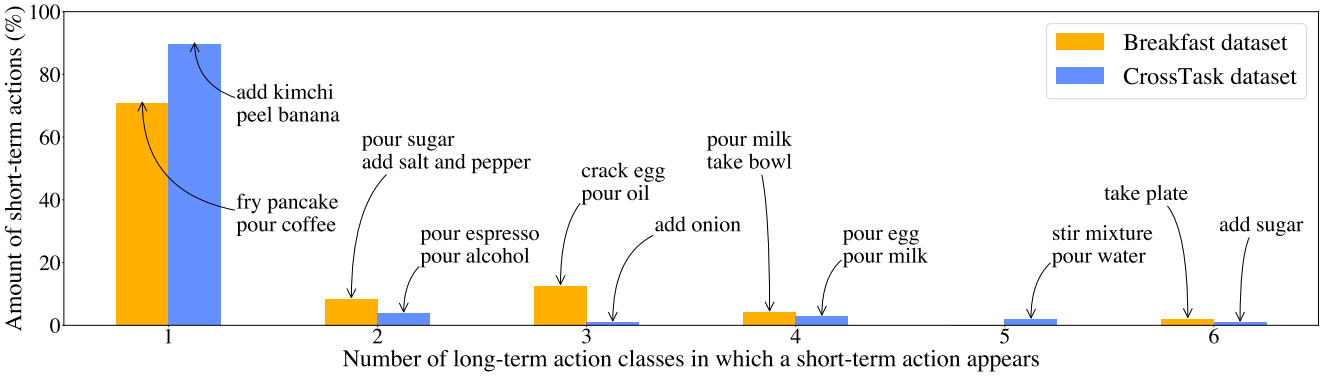

Figure 2: We analyze two popular long-term datasets with long-term and short-term action annotations, Breakfast (coarse annotations) [25] and CrossTask [49] (primary tasks). We count in how many long-term actions the short-term action appears. Recurrent short-term actions, like *pour milk* and *pour egg* appear in four different long-term action classes. More specific short-term actions, like *fry pancake* and *add kimchi*, only occur in one long-term action class. We find that a large percentage of short-term actions (70.8% for Breakfast and 89.5% for CrossTask) appears only in one long-term action class. This implies that recognizing a single short-term action might be sufficient to correctly infer the long-term actions in these datasets.

We design a user study to assess whether a video dataset contains long-term action videos that are not recognizable from a single short-term action. Our study is based on two surveys where users have to watch a video and predict the long-term action being performed in the video. In the *Full Videos Survey*, the users can watch the full video, while in the *Video Segments Survey* a separate group of users can watch only a single short clip extracted from the full video. We measure the average action recognition accuracy of the users per video for each survey. The *Full Videos Survey* gives an upper bound to the user long-term action recognition performance. Comparing the accuracy obtained from the *Video Segments Survey* to the upper bound gives an estimate of how many videos in the dataset require long-term

information to be correctly recognized. If the action recognition performance of the two groups of users is close, we can conclude that most of the videos in the dataset are not suitable to train and evaluate models for long-term action recognition, because they can be recognized solely by exploiting short-term information.

We apply our proposed method to the aforementioned Breakfast and CrossTask datasets and to the Long-form Video Understanding benchmark (LVU) [41], recently proposed for long-term video recognition tasks in movies. We implement the user studies on Amazon Mechanical Turk [1] and collect responses from more than 150 users. Our results show that looking at a single short video segment is sufficient to recognize 90% and 97.2% of the analyzed videos from Breakfast and CrossTask. Similarly, we find that most of the content understanding tasks in LVU can be classified without long-term information, and that some video segments in this dataset are misclassified by users due to annotation noise. We conclude that the aforementioned datasets might not be suitable to develop new methods for long-term action recognition in videos, because they can be solved by ignoring long-term information. We recommend long-term video understanding researchers to be careful when using these datasets and encourage the community to collect more representative video datasets.

In summary, the contributions of our study can be outlined as follows: (1) We provide a definition of long-term action datasets that should prevent long-term action recognition models to use traditional short-term action recognition as a shortcut to solve the task. (2) We introduce a method to investigate whether a video dataset meets this definition of long-term action. (3) We find that short-term information is, in most cases, sufficient to solve long-term video understanding tasks in three commonly used datasets. Thus, we recommend against using these datasets in further research on long term action recognition models. The code and responses from our user study are publicly available[1].

2. Related work

2.1. Action recognition with deep learning

The progress of deep learning (DL) has brought significant advancements in automatic action recognition. DL-based models learn to extract discriminative spatial and temporal features directly from the RGB frames of the training videos. Current action recognition models are composed of 3D convolutional networks [23], like I3D [8], C3D [38], Slow-Fast [12]. More recently, attention-based architectures have also shown competitive performance on action recognition tasks. Examples include ViViT [3], TimeSformer [5] and Video Swin Transformer [28]. When pretrained on sufficiently large datasets, like Kinetics [8] or ActivityNet [7], these models can achieve state-of-the art action recognition on *short* videos datasets, like UCF101 [33], HMDB51 [26] and Something-Something [15]. However, they are not suitable to learn long-term dynamics in long videos, either due to their limited temporal receptive field or the high computational requirements.

2.2. Long-term action recognition

Long-term action recognition refers to the task of recognizing and understanding human actions composed of several short-term actions, possibly involving multiple objects and movements [47]. Examples include cooking a recipe [25], performing a medical surgery [31] or playing a sport game [45]. Usually, long-term actions require an extended period of time to be executed, e.g. above one minute [17]. Several works that tackled the problem of long-term action recognition use different names and definitions for the same concepts. In fact, long-term actions can also be referred to in the literature as *long-range activities* [19, 20] or *complex activities* [16, 17]. Being composed of multiple steps, the activities in *instructional videos* share the same properties of long-term actions [27, 29, 48] and can be comprised into this category. Finally, also *long-form* video understanding involves reasoning over human-object interactions in long videos [41, 44] and can be considered as an instance of long-term action recognition.

Traditional DL-based action recognition models [8, 12, 38, 40] are deemed insufficient to capture discriminative spatio-temporal features that encode long-term information and the semantic relations between the sub-actions. A variety of models have been proposed to overcome this limitation. Hussein *et al.* [17] proposed to capture long-term information with multi-scale temporal convolution. Yu *et al.* [46] used Recurrent Neural Networks to model long video sequences capturing temporal information at different rhythms. Ballan *et al.* [4] showed that explicitly focusing on the actor performing the long-term action improves the recognition performance. Different approaches showed that long-term action recognition can be tackled using graph-based representations, where the nodes correspond to short-term entities and the edges to their interaction over space and time [18, 22, 47]. Finally, Transformer architectures have been designed to model long-term information in a compute- [21, 42] and data-efficient [16] fashion.

Despite their success, DL-based action recognition models can find shortcuts in the data that lets them solve action recognition without learning semantic features, for example classifying the action based on the background scene [10, 13, 43]. In this work, we try to address this problems by analyzing whether commonly used video datasets for long-term action recognition are representative for training DL models, or can be solved using short-term shortcuts.

[1] https://github.com/ombretta/longterm_datasets

2.3. Long-term video datasets

Several datasets have been proposed in the literature to study long-term video understanding tasks. CATER [14] is an ideal example of a dataset that requires long-term information. It involves tracking geometrical shapes that move in a 3D space over time. Sometimes bigger shapes incorporate smaller shapes, rendering their localization impossible without continuous reasoning about past information. As a consequence, models that are not truly long-term fail on this dataset. Unfortunately, the CATER dataset is highly synthetic and cannot be used to train models for real-world applications.

Real-world datasets mostly include cooking [11, 25, 34, 48], home activities [32, 39], sports [45] and instructional videos [2, 36, 48, 49]. A comprehensive overview of long-term video understanding datasets is provided in Table 1. Many of these datasets, for example Charades [32], Epic Kitchens [11] and MultiTHUMOS [45], contain long videos annotated with fine-grained, short-term actions. They can be used for multi-label action recognition, where the task is to predict every short-term action occurring in the video, or for fine-grained action localization. Differently, here we are interested in the single-label classification case, where a global label describes the long-term activity happening in the video. The single label should be recognizable only by reasoning over multiple short-term actions.

Previous work showed that video datasets are sometimes biased towards appearance and more easily recognizable by static information over temporal information [6]. Similarly, in this work we explore whether the global labels of datasets proposed for long-term video understanding tasks can be predicted without long-term information. We choose for a study three popular datasets that include single, video-level labels and cover different long-term dataset categories: Breakfast, CrossTask and LVU. Breakfast [25] is a *complex action recognition* dataset used in several works on long-term video understanding [16, 17, 27, 47]. CrossTask [49] is a dataset of *instructional videos*, which are composed of several short-term steps that contribute to the completion of a long-term task. Finally, the *Long-form Video Understanding* (LVU) dataset [41] was proposed to learn complex long-term relationships, in contrast to short-term patterns, in video clips extracted from movies.

Dataset	#Videos	Length	#L.T.	#S.T.
COFFEE [2]	150	2	5	51
Epic-Kitchens [11]	432	7.5	-	149, 323
Breakfast [25]	2k	2.3	10	48
Composite [30]	212	1-23	41	218
Charades [32]	10k	0.5	-	157
50-Salads [34]	54	6.4	-	17
COIN [36]	11.8k	2.4	180	778
IKEA FA [37]	101	2-4	-	12
DAHLIA [39]	51	39	7	-
LVU - Content understanding [41]	226	1-3	4	-
	1.3k	1-3	5	-
	723	1-3	6	-
Multi-THUMOS [45]	413	3	-	65
YouCookII [48]	2k	5.3	89	-
CrossTask [49]	4.7k	3-6	83	517

Table 1: Overview of current real-world datasets proposed for long-term video understanding tasks. We report the (approximate) number of videos, the average video length in minutes, the number of global *long-term* (L.T.) and *short-term* (S.T.) action recognition classes, if it applies.

3. Assessing long-term action recognition datasets

3.1. User study

According to our definition, an action is long-term if it cannot be classified from a single short video segment. We design a user study to test whether current long-term video understanding datasets respect this property. Our user study consists of two surveys. In the *Full Videos Survey*, the users are presented with the full-length videos from the datasets. In the *Video Segments Survey*, the users are presented with a short video segment extracted from a full-length video. In both surveys, the users are instructed to watch the video clip and express what action is being performed in the full video, in their opinion. The users are provided with a list of possible actions, which correspond to the classes from the analyzed long-term action datasets, and have to select exactly one action class from the list. We include the additional option "*I am not sure*", to let the users express uncertainty when they are in doubt about which action to select.

From the collected user votes in the *Full Videos Survey* and the *Video Segments Survey*, we calculate and compare the action recognition accuracy. If the users from the two groups perform similarly, we can conclude that the videos do not contain long-term actions, as they can be recognized from single short-term actions comparably well than looking at the full videos. We also calculate the user agreement per survey, measured with Krippendorff's α [24], which gives an indication of how subjective the prediction task is. We expect that the more a video is difficult to classify, the more subjective the choice will be, thus resulting in low agreement.

3.2. Measuring recognition accuracy

From the *Full Videos Survey*, we collect user votes per class for each full-length video. In each full video, we express the votes in percentages ($\%user\_votes_v(c)$), which we obtain by dividing the votes per class by the amount of

votes collected for the full video. As formalized in Equation 1, given $\mathcal{C}$ classes from the evaluated dataset, excluding the *I am not sure* option, we assign to the full video prediction ($pred(v)$) the class voted by the majority of the users. The long-term action recognition accuracy is given by the number of full videos assigned with the correct class over the number of full videos considered in the study for the dataset.

$$pred(v) = \arg\max_{c \in \mathcal{C}} \quad \%user\_votes_v(c) \tag{1}$$

In the *Video Segments Survey*, we collect user votes for every segment s_v in a full video. Again, for each segment we calculate the percentage of votes per class $\%user\_votes(c)$. Then, we extract the full video prediction from the votes of a single segment. To do this, we select the segment s_v^* with highest percentage of votes for a single class, excluding the *I am not sure* option. This approach is formalized in Equation 2. In the example in Figure 3, the full video is assigned the class *Making scrambled eggs*, which is voted by 86% of users in *Segment 5*, which is the maximum ratio of votes for one class across the video segments. According to our definition, if the full-length video is long-term, there should be no video segments that lead to the right predicted class. The accuracy is given by the number of full videos assigned with the correct label over the number of full videos considered in the study.

$$pred(v) = pred(s_v^*), \tag{2}$$

$$\text{where } s_v^* = \arg\max_{s_v \in v} \quad \{\max_{c \in \mathcal{C}} \quad \%user\_votes_{s_v}(c)\},$$
$$pred(s_v^*) = \arg\max_{c \in \mathcal{C}} \quad \%user\_votes_{s_v^*}(c).$$

4. Results

We include in our study a representative dataset from complex action recognition, Breakfast [25], one instructional video dataset, CrossTask [49], and the Long-Form Video Understanding (LVU) dataset [41]. We implement the user study on Amazon Mechanical Turk [1] and collect responses from 167 users. We collect, on average, 12.09±1.62 votes for each video and video segment, which is proved to be a proper amount [9]. Table 2 provides an overview of the results from the *Full Videos Survey* and the *Video Segments Survey*, discussed in the following sections.

4.1. Breakfast

Breakfast [25] is a collection of third-person videos of actors cooking a breakfast recipe, like scrambled eggs, coffee, cereals and milk. Each video has a global label, which corresponds to the recipe being made, for a total of 10 classes. The classification task consists in correctly recognizing the recipe.

Dataset	**Classification accuracy (%)**	
	Full Videos	Video Segments
Breakfast	**93.33**	90.0
CrossTask	**100.0**	97.2
LVU – Relationship	**88.89**	**88.89**
LVU – Scene	**100.0**	**100.0**
LVU – Speaking	**80.0**	60.0

Table 2: Average video recognition accuracy obtained from the *Full Videos Survey* and *Video Segments Survey* on the Breakfast [25], CrossTask [49] and LVU [41] datasets. The results suggest that long-term information is helpful but not necessary in the majority of the evaluated datasets.

Dataset	**User agreement**		
	Full Videos	Video Segments	Selected Segments
Breakfast	**0.717**	0.386	0.593
CrossTask	0.671	0.462	**0.767**
LVU – relationship	0.499	0.340	**0.523**
LVU – scene	**0.755**	0.481	0.686
LVU – speaking	0.159	0.191	**0.265**

Table 3: Overview of the user agreement in our user studies, measured terms of Krippendorff's α [24]. We find that the users tend to agree in the *Full Videos Surveys* and when selecting the segments with highest amount of votes for a class. Recognizing the actions in the *Video Segments Survey* is generally harder then when looking at the full video, resulting in more variability in the users predictions and, consequently, in lower agreement.

For our study, we select a representative subset of 30 videos, corresponding to 3 randomly selected videos per class. The full videos have average duration of 2.44 ± 2.18 minutes. For the *Video Segments Survey*, we segment the video according to the short-term action timesteps (*coarse segmentation*) provided in the dataset. We remove segments that are shorter than 5 seconds, as we deem those segments highly uninformative, and we obtain 154 segments in total, of average duration 29 ± 39 seconds, where ∼ 56% of the segments last less than 15 seconds. The large standard deviation is due to some repetitive short-term actions that can last above a minute, e.g. *stir dough* or *fry egg*.

The results in Table 2 show that the recognition accuracy from the *Full Videos Survey* (93.33%) and the *Video Segments Survey* (90.0%) are close. This suggests that, although having access to the full long-term information in the video helps, looking at single short segments is sufficient to infer the right recipe class for the majority of the

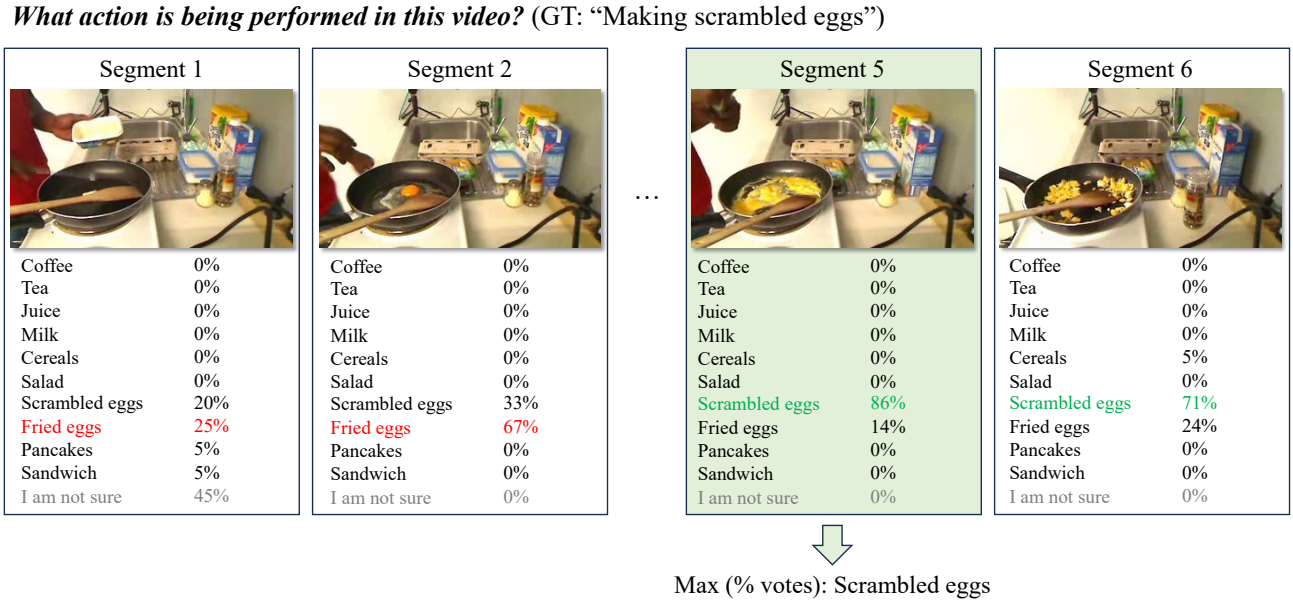

Figure 3: In the *Video Segments Survey*, users have to understand what is happening in a long video by looking only at one short segment. We ask the users to vote for a video class and obtain predictions per segment. We assign to the full video the segment prediction with the highest percentage of votes for one class. In the example, taken from the Breakfast dataset [25], *Segment 5* determines the video prediction *Scrambled eggs*.

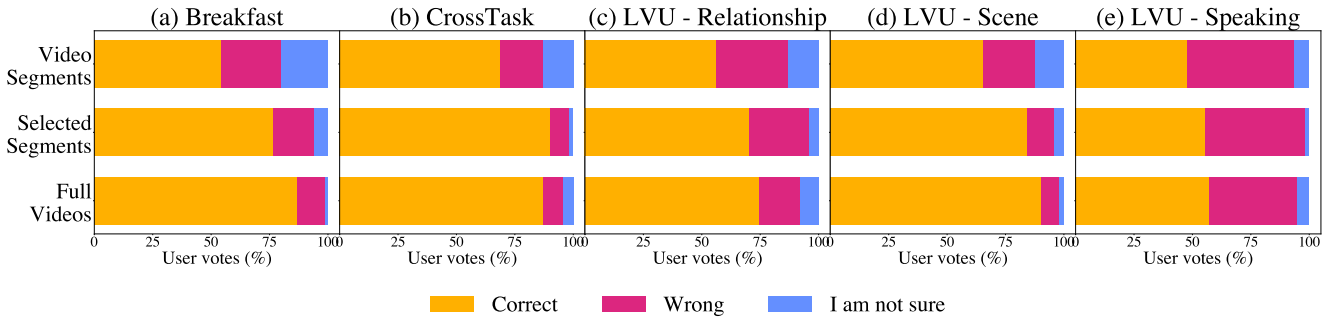

Figure 4: Overview of the user votes (correct, wrong and *I am not sure*) collected in our study. We compare the results from the *Full Videos*, all the *Video Segments*, and the Selected Segments with highest percentage of votes for one class. The amount of correct votes in the Selected Segments is significantly higher than for all the *Video Segments*, and comparable, or even higher, to the amount of correct votes obtained watching the full videos. N.b., the user votes reported in this figure do not have to match the accuracies in Table 2. While the accuracy shows the percentage of videos correctly classified, the user votes are aggregated without considering the votes distributions within the specific videos.

videos. From this result we conclude that the Breakfast dataset is not a proper long-term action dataset, according to our definition.

We analyze the amount of correct user votes, wrong votes and *I am not sure* votes obtained in the user study and illustrated in Figure 4 (a). We obtained 86.78% of correct votes in the *Full Videos Survey* and 54.47% in the *Videos Segments Survey*. However, if we consider only the segments with the highest percentage of votes for one class, the amount of correct votes reaches 76.36%. A similar trend occurs in the user agreement in Table 3. By further inspecting the results from the *Video Segments Survey*, we notice that users are generally more uncertain classifying the video segments early in the video, with a higher portion of *I am not sure* votes compare to the later segments. In particular, 63.57% of *I am not sure* votes are obtained in from the first two video segments in chronological order. We argue that breakfast dishes are usually better recognizable towards the

end of the video, when the recipe is complete.

4.2. CrossTask

CrossTask [49] is an instructional video dataset of $\sim$ 4.7k videos, covering themes like auto repair, cooking and DIY. The instructional videos show how to perform a *tasks* (e.g., *Make a Latte*) through a list of *steps* (e.g., *add coffee*, *press coffee*, *pour water*, *pour espresso*, *steam milk*, *pour milk*). It contains 18 primary tasks with steps annotations and 65 related tasks with unlabeled steps. The dataset is meant to be used to learn steps in a weakly supervised learning setup. Here, we evaluate whether predicting the *task* illustrated in an instructional video also fits our definition of long-term action recognition. We collect results from 36 video clips (2 random videos per primary task) of average duration 4.50 $\pm$ 2.14 minutes. Similarly to Breakfast, we extract 260 segments from the videos according to the timesteps provided with the dataset. In CrossTask, the segments are significantly shorter than Breakfast, with average duration of 10 $\pm$ 11 seconds and $\sim$ 81% of the segments being shorter than 15 seconds.

In Table 2, we compare the task recognition accuracy from the *Full Videos Survey*, 100%, and the *Video Segments Survey*, 97.2%. In both cases, users can recognize the task with high accuracy. Only one video (YouTube id *kReUYklvjnc*) is misclassified in the *Video Segments Survey*, despite 5/8 of its video segments being correctly classified. Considering the user agreement (Table 3) and correct votes by the users (Figure 4, b), we find that both quantities are marginally higher in the Selected Segments over the Full Videos. This result shows that users tend to make the same mistakes (as for video *kReUYklvjnc*) while confirming that most of the tasks are generally recognizable both from short video segments and full videos. It is worth noting that the results reported in Table 2 and Figure 4 are not necessarily the same. The accuracy corresponds to the percentage of videos correctly classified, while the user votes are aggregated without considering the votes distributions within the specific videos. Because of the high task recognition accuracy obtained from the *Video Segments Survey*, we conclude that the videos in CrossTask do not contain long-term actions. We recommend to use this dataset for the other video understanding tasks that is supports, like captioning and action localization.

4.3. LVU

The Long-Form Video Dataset (LVU) [41] has been recently proposed to study complex relationships in video clips extracted from movies. It provides three tasks, related to content understanding, user engagement prediction and movie metadata prediction and contains over 11k videos. Similarly to previous work [35], we select the task of *Content Understanding*, which involves classifying the *relationship* among the characters, where the *scene* is taking place and the characters *speaking* style, from video clips of $\sim$2.5 minutes. The respective annotations consist in a global label per video. We assess whether predicting *Relationship*, *Scene* and *Speaking* is a form of long-term action recognition, according to our definition. We select videos from the test set and manually extract segments for each of the three classification tasks. We obtain 9 videos (3 per class) for *Relationship*, 12 videos (2 per class) for *Scene* and 10 videos (2 per class) for *Speaking*, and a total of 140 segments of $\sim$30 seconds.

Table 2 shows the classification accuracies obtained from the *Full Videos Survey* and *Video Segments Survey*. Comparing the results, we find no difference for *Relationship* and *Scene*. In particular, *Scene* classification is performed with 100% accuracy, indicating that this prediction task is easy for humans. We identify a problem associated with LVU - *Relationship*. The labels husband-wife, friends, boyfriend-girlfriend are associated with specific characters in the movie, but other characters might appear within the same video clip. For example, in Figure 5 (a), the ground-truth label for the movie in the first row is *Husband-Wife*. However, a third male character appears in the scene in addition to the *husband and wife*. Therefore, the labels only correctly apply to a specific subset of the characters in the scene, or to a precise time window when only the target characters appear. As a result, the full videos are classified with a high percentage of wrong votes, while some of the video segments that do not include the characters corresponding to the label are completely misclassified. This justifies the large portion of wrong votes in Figure 4 (c) and relatively low agreement in Table 3.

We find a similar annotation problem in LVU - *Speaking*. Also in this case, the global label only applies to a subset of the characters in the scene. In the example in Figure 5 (c), the label *Threatens* only applies to the man with the gun. This explains the difference in performance when comparing the accuracies from the *Full Videos Survey* and *Video Segments Survey* in Table 2, the large amount of wrong votes in Figure 4 (e) and low agreement in Table 3. Because of the problem with the annotations and the equal recognition performance of 88.89% obtained from the *Full Videos Survey* and *Video Segments Survey* (reported in Table 2), we conclude that LVU - *Relationship* is not a long-term video understanding task. Similar conclusions apply for LVU - *Scene*, with perfect classification scores resulting from both surveys. Finally, the labels in LVU - *Speaking* are not truly long-term, as they apply to a subset of characters speaking only during some relatively short time-windows.

5. Conclusion

We propose a method to assess whether an action is *long-term*. We apply our method to three current long-term video

Figure 5: Examples of correct (green) and wrong (red) classification results collected from the *Video Segments* (V.S.) and *Full Videos* (F.V.) surveys on the Long-form Video Understanding (LVU) - Relationship (a), Scene (b) and Speaking(c) dataset [41]. Users correctly classify a large portion of video segments. Other segments result misclassified due to annotation noise.

understanding datasets, Breakfast, CrossTask and LVU. Our results show that long-term information might help but is *not necessary* in the majority of videos from the analyzed datasets. In fact, the long-term actions in these videos can be correctly classified by humans by looking solely at a single short video segment. This result suggests that deep learning models trained and tested on these datasets might pick short-term shortcuts and still show correct recognition performance, without actually learning any long-term information. Following our findings, we urge researchers who are investigating automatic long-term action recognition to use datasets that need long-term information to be solved.

Acknowledgements. This work is part of the research program Efficient Deep Learning (EDL), which is (partly) financed by the Dutch Research Council (NWO).

References

[1] Amazon mechanical turk. https://www.mturk.com/. Accessed: 2023-07-05. 3, 5

[2] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Learning from narrated instruction videos. *IEEE transactions on pattern analysis and machine intelligence*, 40(9):2194–2208, 2017. 4

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 3

[4] Luca Ballan, Ombretta Strafforello, and Klamer Schutte. Long-term behaviour recognition in videos with actor-focused region attention. In *VISIGRAPP (5: VISAPP)*, pages 362–369, 2021. 3

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 3

[6] Petr Byvshev, Pascal Mettes, and Yu Xiao. Are 3d convolutional networks inherently biased towards appearance? *Computer Vision and Image Understanding*, 220:103437, 2022. 4

[7] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 3

[8] 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*, pages 6299–6308, 2017. 3

[9] Arthur Carvalho, Stanko Dimitrov, and Kate Larson. How many crowdsourced workers should a requester hire? *Annals of Mathematics and Artificial Intelligence*, 78:45–72, 2016. 5

[10] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 1, 4

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 3

[13] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1, 3

[14] Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning. *arXiv preprint arXiv:1910.04744*, 2019. 4

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 3

[16] Hongji Guo, Hanjing Wang, and Qiang Ji. Uncertainty-guided probabilistic transformer for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20052–20061, 2022. 1, 3, 4

[17] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Timeception for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2019. 1, 3, 4

[18] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Videograph: Recognizing minutes-long human activities in videos. *arXiv preprint arXiv:1905.05143*, 2019. 3

[19] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Pic: Permutation invariant convolution for recognizing long-range activities. *arXiv preprint arXiv:2003.08275*, 2020. 3

[20] Noureldien Hussein, Mihir Jain, and Babak Ehteshami Bejnordi. Timegate: Conditional gating of segments in long-range activities. *arXiv preprint arXiv:2004.01808*, 2020. 3

[21] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022. 3

[22] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 3

[23] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012. 3

[24] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011. 4, 5

[25] 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*, pages 780–787, 2014. 1, 2, 3, 4, 5, 6

[26] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 3

[27] Muheng Li, Lei Chen, Yueqi Duan, Zhilan Hu, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Bridge-prompt: Towards ordinal action understanding in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19880–19889, 2022. 1, 3, 4

[28] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 3

[29] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 3

[30] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 144–157. Springer, 2012. 4

[31] Aidean Sharghi, Helene Haugerud, Daniel Oh, and Omid Mohareri. Automatic operating room surgical activity recognition for robot-assisted surgery. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 385–395. Springer, 2020. 3

[32] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016. 1, 4

[33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3

[34] 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. 4

[35] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pre-training with multimodal temporal contrastive learning. *arXiv preprint arXiv:2210.06031*, 2022. 7

[36] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 4

[37] Sam Toyer, Anoop Cherian, Tengda Han, and Stephen Gould. Human pose forecasting via deep markov models. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2017. 4

[38] 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*, pages 4489–4497, 2015. 3

[39] Geoffrey Vaquette, Astrid Orcesi, Laurent Lucat, and Catherine Achard. The daily home life activity dataset: A high semantic activity dataset for online recognition. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 497–504, 2017. 4

[40] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 1, 3

[41] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021. 3, 4, 5, 7, 8

[42] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022. 3

[43] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. 3

[44] Xitong Yang, Fu-Jen Chu, Matt Feiszli, Raghav Goyal, Lorenzo Torresani, and Du Tran. Relational space-time query in long-form videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6408, 2023. 3

[45] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126:375–389, 2018. 1, 3, 4

[46] Tianshu Yu, Yikang Li, and Baoxin Li. Rhyrnn: Rhythmic rnn for recognizing events in long and complex videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 127–144. Springer, 2020. 3

[47] Jiaming Zhou, Kun-Yu Lin, Haoxin Li, and Wei-Shi Zheng. Graph-based high-order relation modeling for long-term action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8993, 2021. 1, 3, 4

[48] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 3, 4

[49] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019. 1, 2, 4, 5, 7