Supplementary Material: Large Scale Holistic Video Understanding

Ali Diba[1,5⋆], Mohsen Fayyaz[2,⋆], Vivek Sharma[3,⋆],
Manohar Paluri, Jürgen Gall[2], Rainer Stiefelhagen[3], Luc Van Gool[1,4,5]

[1]KU Leuven, [2]University of Bonn,[3]KIT, Karlsruhe, [4]ETH Zürich, [5]Sensifai
{firstname.lastname}@kuleuven.be, {lastname}@iai.uni-bonn.de,
{firstname.lastname}@kit.edu, Balamanohar@gmail.com

Appendix: This document provides supplementary material as mentioned in the main paper.

A HVU Dataset

A.1 Human Annotation Details

The row machine generated annotations consist almost 8K labels. The initial stage of human verification on validation set resulted in 4378 labels. And the final stage of complete human verification/modification process ended up in 3142 labels. In human annotation process, 80 new labels are added by human annotators.

In specific for the HVU human verification task, we employed three different teams (Team-A, Team-B and Team-C) of 55 human annotators. Team-A works on the taxonomy of the dataset. This team builds the taxonomy based on the visual meaning and definition of the tags obtained from APIs prediction. Team-B and Team-C are the verification teams and perform four tasks. The tasks they performs are: (a) verify the tags of videos by watching each video and flag false tags; (b) review the tags by watching the videos of each tag and flag the wrong videos; (c) add tags to the videos if some tags are missing; and (d) they suggest modification on tags such as, renaming or merging.

To make sure both Team-B and Team-C have a clear understanding of the tags and the corresponding videos, we ask them to use the provided tags definition from Team-A. For the aforementioned four tasks, Team-B goes through all the videos and provides the first round of clean annotations. Followed by this, Team-C reviews the annotations from Team-B to guarantee an accurate and cleaner version of annotations. Finally, Team-A reviews the suggestions provided from tasks (c) and (d) and apply them to the dataset. The verification process takes ∼100 seconds on average per video clip for a trained worker. It took about 8500 person-hours to firstly clean the machine-generated tags and remove errors and secondly add any possible missing labels from the dictionary. By incorporating the machine generated tags and human annotation, the HVU

⋆Ali Diba, Mohsen Fayyaz and Vivek Sharma contributed equally to this work and listed in alphabetical order.

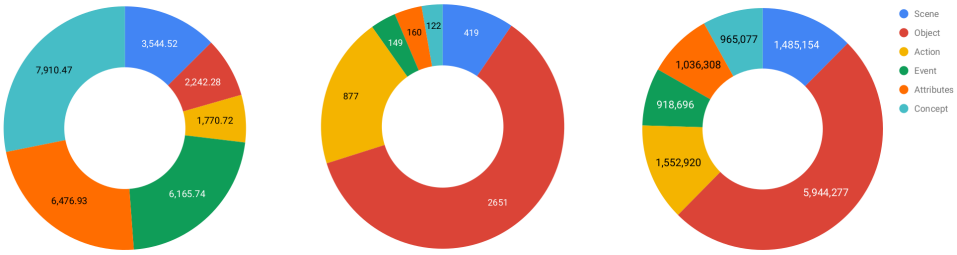


Fig. 1: Left: Average number of samples per label in each of main categories. Middle: Number of labels for each main category. Right: Number of samples per main category. All statistics are for the machine generated tags of HVU training set.

Task Category	Scene	Object	Action	Event	Attribute	Concept	Total
#Labels	419	2651	877	149	160	122	4378
#Annotations	1,485,154	5,944,277	1,552,920	918,696	1,036,308	965,077	11,902,432
#Videos	366,941	480,821	481,418	320,428	368,668	375,664	481,418

Table 1: Statistics of machine generated tags of HVU training set for different categories. The category with the highest number of labels and annotations is the object category.

dataset covers a diverse set of tags with clean annotations. Using machine generated tags in the first step helps us to cover larger number of tags than a human can remember and label it in a reasonable time.

To make sure that we have a balanced distribution of samples per tag, we consider a minimum number of 50 samples.

To provide more details regarding the HVU human annotation process, we report the statistics of the different stages of the annotation process. Table 1 shows the statistics of the machine generated annotations of training set. Note, that the labels and categories are result of the initial human annotation process over the validation set of the dataset. The category with the highest number of labels and annotations is the object category. Concept is the category with the lowest number of labels. To have a better understanding of the statistics of the annotations we depict the distribution of categories with respect to the number of annotations, labels, and annotations per label in Figure 1. We can observe that the object category has the highest quota of labels and annotations, which is due to the abundance of objects in video. Despite having the highest quota of the labels and annotations, the object category does not have the highest annotations per label ratio. Figure 2 shows the percentage of the different subsets of the main categories. There are 50 different sets of videos based on assigned semantic categories. About 36% of the videos have all of the categories.

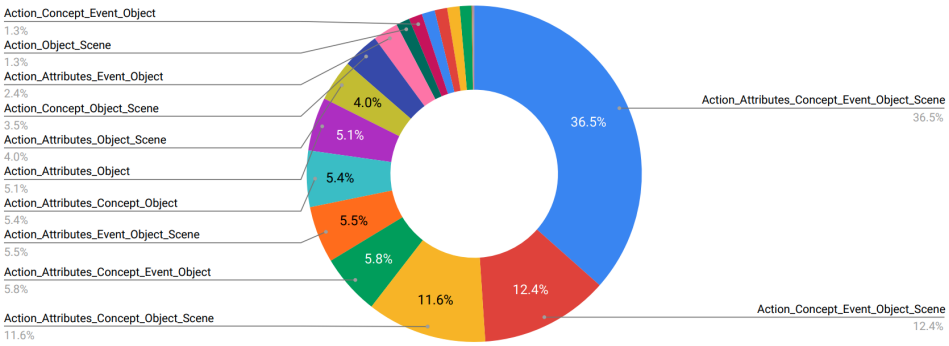


Fig. 2: Coverage of different subsets of the 6 main semantic categories in videos. 16.4% of the videos have annotations of all categories. All statistics are for the machine generated tags of HVU training set.

Dataset	Scene	Object	Action	Event	Attribute	Concept	HVU Overall %
Machine-Generated HVU	46.3	22.4	43.8	31.4	25.3	20.1	31.6
Human-Annotation HVU	50.1	27.9	46.7	35.7	29.2	23.2	35.4

Table 2: Performance comparison between machine generated and human-verified tags of HVU. This evaluation shows how human annotation process is crucial to have a more efficient dataset. The CNN model which is used for this experiment is 3D-ResNet18.

A.2 Effect of Human Annotation

To present the impact of human annotation process, we have evaluated both versions of the HVU with machine-generated tags and human-annotated tags. We have trained two 3D-ResNet18 for each set and the comparison came in Table 2.

A.3 HVU Samples

We present some samples of videos and their corresponding tags in Fig 3 and Fig 4.

A.4 Effect of Additional Categories on Kinetics

One of our arguments in our paper is about how more semantic categories like object, scene, etc can lead to learn effective video representation. We have shown results on the HVU dataset in the paper. Here, we provided the similar experiment for the Kinetics-600 as a subset of our HVU. We have compared performance of a 3D-ResNet18 trained on Kinetics videos with its action labels versus

Training Labels	Action Recognition Performance
Action	65.6
Action + HVU	68.8

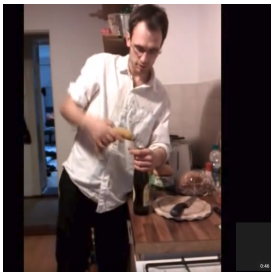
Table 3: Evaluation of training Kinetics with HVU labels.

trained on full HVU labels for the same videos. For the evaluation, we have measured the performance on Kinetics action labels. It can be seen in Table 3 that having more semantic labels in the training for Kinetics, improves the action classification performance. It is due to the fact that HVU can bring more capabilities to the deep models for learning new visual features for understanding videos.

forest,musician,flutist,music,musical_instrument,brass_instrument,wind_instrument,flautist,recreation,musical_instrument_accessory,plant,playing_flute,tree

string_instrument,musician,man,,guitarist,plucked_string_instruments,music,tapping_guitar,bass,musical_instrument_accessory,performance,string_instrument_accessory,electric_guitar,sitting,monochrome_photography,musical_instrument,guitar_accessory,resonator

sport_venue,shoe,outdoor_shoe,joint,foot,ball,grass,knee,human_leg,fun,football_player,ball_game,green,footwear,football,player,sports_equipment,juggling_soccer_ball,soccer,plant,soccer_ball,sports,play



opening_bottle_not_wine_,joint,muscle,service,finger,distilled_beverage,fun,taste,standing,arm,t_shirt,glass,alcohol,drink,hand,bottle,photograph,cooking

smile,nose,textile,cheek,thigh,mouth,girl,diaper,finger,baby_products,human_leg,fun,playing_xylophone,infant,toy,facial_expression,skin,child,hand,sitting,human_hair_color,daytime,play,toddler

coast,watercourse,plant,wetland,terrain,floodplain,marsh,wading_through_mud,boulder,tree,water,natural_resources,river,rock,waterway,outcrop,shore,creek

Fig. 3: Video frame samples from HVU with corresponding tags of different categories.

mopping_floor,wood_stain,sleeve,design,man,wood,wood_flooring,gentleman,standing,swab,facial_hair,shirt,outerwear,tartan,flooring,laminate_flooring,floor,dress_shirt,plaid,angle

individual_sports,indoor_games_and_sports,joint,games,combat_sport,weapon_combat_sports,leisure,net,fun,recreation,martial_arts,epee,striking_combat_sports,fencing,competition,contact_sport,fencing_sport_,flooring,fencing_weapon,floor,sports,play

italian_food,food,pizza,making_pizza,appetizer,cuisine,pizza_cheese,prosciutto,vegetable,darkness,sicilian_pizza,recipe,rectangle,european_food,flatbread

charcoal,campfire,shovel,smoke,outdoor_grill,fire,animal_source_foods,barbecue_grill,grilling,winter,fun,ice,meat,cooking_on_campfire,roasting,grass,barbequing

hand,multimedia,server,electronics,electronic_device,finger,computer_hardware,technology,assembling_computer,computer_case,arm,magenta,text

bee_keeping,human,grass,backyard,outdoor_structure,wood,tree,forest,human_behavior,beekeeper,leaf,garden,bee,yard,apiary,t_shirt,plant,beehive,male

Fig. 4: More examples of video frame samples from HVU with corresponding tags of different categories.