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

Abstract. Video recognition has been advanced in recent years by benchmarks with rich annotations. However, research is still mainly limited to human action or sports recognition - focusing on a highly specific video understanding task and thus leaving a significant gap towards describing the overall content of a video. We fill this gap by presenting a large-scale "Holistic Video Understanding Dataset" (HVU). HVU is organized hierarchically in a semantic taxonomy that focuses on multi-label and multi-task video understanding as a comprehensive problem that encompasses the recognition of multiple semantic aspects in the dynamic scene. HVU contains approx. 572k videos in total with 9 million annotations for training, validation and test set spanning over 3142 labels. HVU encompasses semantic aspects defined on categories of scenes, objects, actions, events, attributes and concepts which naturally captures the real-world scenarios.

We demonstrate the generalisation capability of HVU on three challenging tasks: 1.) Video classification, 2.) Video captioning and 3.) Video clustering tasks. In particular for video classification, we introduce a new spatio-temporal deep neural network architecture called "Holistic Appearance and Temporal Network" (HATNet) that builds on fusing 2D and 3D architectures into one by combining intermediate representations of appearance and temporal cues. HATNet focuses on the multi-label and multi-task learning problem and is trained in an end-to-end manner. Via our experiments, we validate the idea that holistic representation learning is complementary, and can play a key role in enabling many real-world applications. https://holistic-video-understanding.github.io/

1 Introduction

Video understanding is a comprehensive problem that encompasses the recognition of multiple semantic aspects that include: a scene or an environment, objects, actions, events, attributes, and concepts. Even if considerable progress is made in video recognition, it is still rather limited to action recognition - this is due to the fact that there is no established video benchmark that integrates joint

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

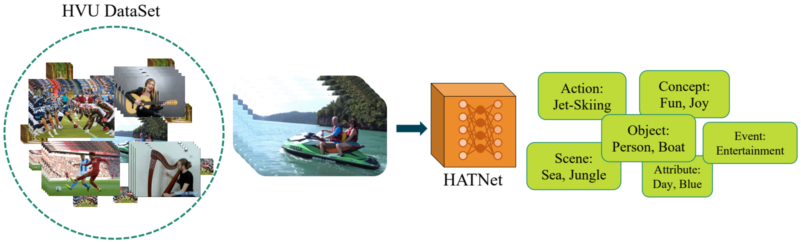

Fig. 1: Holistic Video Understanding Dataset: A multi-label and multi-task fully annotated dataset and HATNet as a new deep ConvNet for video classification.

recognition of multiple semantic aspects in the dynamic scene. While Convolutional Networks(ConvNets) have caused several sub-fields of computer vision to leap forward, one of the expected drawbacks of training the ConvNets for video understanding with a single label per task is insufficiency to describe the content of a video. This issue primarily impedes the ConvNets to learn a generic feature representation towards challenging holistic video analysis. To this end, one can easily overcome this issue by recasting the video understanding problem as multi-task classification, where multiple labels are assigned to a video from multiple semantic aspects. Furthermore, it is possible to learn a generic feature representation for video analysis and understanding. This is in line with image classification ConvNets trained on ImageNet that facilitated the learning of generic feature representation for several vision tasks. Thus, training ConvNets on a multiple semantic aspects dataset can be directly applied for holistic recognition and understanding of concepts in video data, which makes it very useful to describe the content of a video.

To address the above drawbacks, this work presents the "Holistic Video Understanding Dataset" (**HVU**). HVU is organized hierarchically in a semantic taxonomy that aims at providing a multi-label and multi-task large-scale video benchmark with a comprehensive list of tasks and annotations for video analysis and understanding. HVU dataset consists of 476k, 31k and 65k samples in train, validation and test set, and is a sufficiently large dataset, which means that the scale of dataset approaches that of image datasets. HVU contains approx. 572k videos in total, with ∼7.5M annotations for training set, ∼600K for validation set, and ∼1.3M for test set spanning over 3142 labels. A full spectrum encompasses the recognition of multiple semantic aspects defined on them including 248 categories for scenes, 1678 for objects, 739 for actions, 69 for events, 117 for attributes and 291 for concepts, which naturally captures the long tail distribution of visual concepts in the real world problems. All these tasks are supported by rich annotations with an average of 2112 annotations per label. The HVU action categories builds on action recognition datasets [23, 27, 29, 47, 64] and further extend them by incorporating labels of scene, objects, events, attributes, and concepts in a video. The above thorough annotations enable developments of strong algorithms for a holistic video understanding to describe the content of a video. Table 1 shows the dataset statistics.

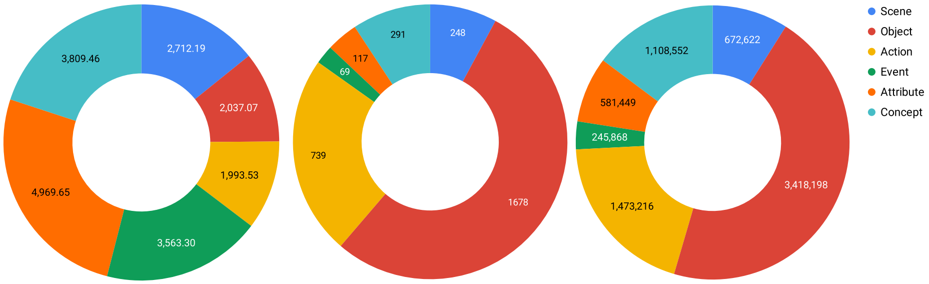

Fig. 2: Left: Average number of samples per label in each of main categories. Middle: Number of labels for each main category. Right: Number of samples per main category.

In order to show the importance of holistic representation learning, we demonstrate the influence of HVU on three challenging tasks: video classification, video captioning and video clustering. Motivated by holistic representation learning, for the task of video classification, we introduce a new spatio-temporal architecture called "Holistic Appearance and Temporal Network" (HATNet) that focuses on the multi-label and multi-task learning for jointly solving multiple spatio-temporal problems simultaneously. HATNet fuses 2D and 3D architectures into one by combining intermediate representations of appearance and temporal cues, leading to a robust spatio-temporal representation. Our HATNet is evaluated on challenging video classification datasets, namely HMDB51, UCF101 and Kinetics. We experimentally show that our HATNet achieves outstanding results. Furthermore, we show the positive effect of training models using more semantic concepts on transfer learning. In particular, we show that pre-training the model on HVU with more semantic concepts improves the fine-tuning results on other datasets and tasks compared to pre-training on single semantic category datasets such as, Kinetics. This shows the richness of our dataset as well as the importance of multi-task learning. Furthermore, our experiments on video captioning and video clustering demonstrates the generalisation capability of HVU on other tasks by showing promising results in comparison to the state-of-the-art.

2 Related Work

Video Recognition with ConvNets: As to prior hand-engineered [8, 28, 30, 39, 55, 61] and low-level temporal structure [18, 19, 35, 58] descriptor learning there is a vast literature and is beyond the scope of this paper.

Recently ConvNets-based action recognition [16, 26, 46, 50, 59] has taken a leap to exploit the appearance and the temporal information. These methods operate on 2D (individual image-level) [12, 14, 20, 48, 49, 59, 63] or 3D (video-clips or snippets of K frames) [16, 50, 51, 53]. The filters and pooling kernels for these architectures are 3D (x, y, time) i.e. 3D convolutions ($s \times s \times d$) [63] where d is the kernel's temporal depth and s is the kernel's spatial size. These 3D ConvNets are intuitively effective because such 3D convolution can be used to directly extract spatio-temporal features from raw videos. Carreira *et al.*proposed inception [25]

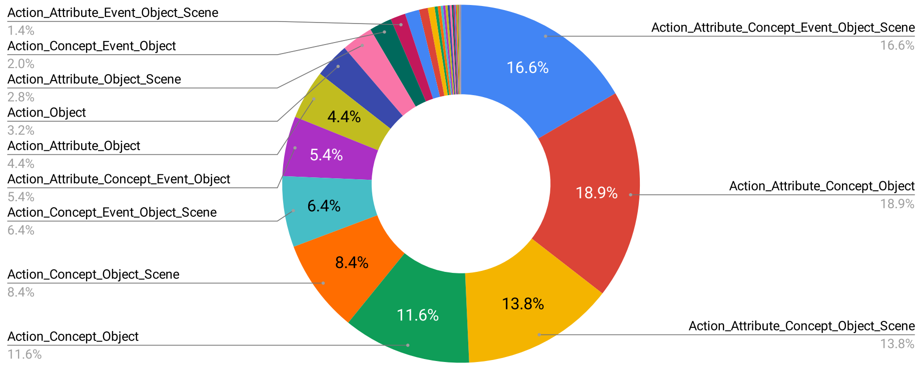

Fig. 3: Coverage of different subsets of the 6 main semantic categories in videos. 16.6% of the videos have annotations of all categories.

based 3D CNNs, which they referred to as I3D [6]. More recently, some works introduced temporal transition layer that models variable temporal convolution kernel depths over shorter and longer temporal ranges, namely T3D [11]. Further Diba *et al.* [10] propose spatio-temporal channel correlation that models correlations between channels of a 3D ConvNets wrt. both spatial and temporal dimensions. In contrast to these prior works, our work differs substantially in scope and technical approach. We propose an architecture, HATNet, that exploits both 2D ConvNets and 3D ConvNets to learn an effective spatio-temporal feature representation. Finally, it is worth noting the self-supervised ConvNet training works from unlabeled sources [21, 42, 44], such as Fernando *et al.* [17] and Mishra *et al.* [33] generate training data by shuffling the video frames; Sharma *et al.* [37, 40, 41, 43] mines labels using a distance matrix or clustering based on similarity although for video face clustering; Wei *et al.* [60] predict the ordering task; Ng *et al.* [34] estimates optical flow while recognizing actions; Diba *et al.* [13] predicts short term future frames while recognizing actions. Self-supervised and unsupervised representation learning is beyond the scope of this paper.

The closest work to ours is by Ray *et al.* [36]. Ray *et al.*concatenate pre-trained deep features, learned independently for the different tasks, scenes, object and actions aiming to the recognition, in contrast our HATNet is trained end-to-end for multi-task and multi-label recognition in videos.

Video Classification Datasets: Over the last decade, several video classification datasets [4, 5, 29, 38, 47] have been made publicly available with a focus on action recognition, as summarized in Table 2. We briefly review some of the most influential action datasets available. The HMDB51 [29] and UCF101 [47] has been very important in the field of action recognition. However, they are simply not large enough for training deep ConvNets from scratch. Recently, some large action recognition datasets were introduced, such as ActivityNet [5] and Kinetics [27]. ActivityNet contains 849 hours of videos, including 28,000 action instances. Kinetics-600 contains 500k videos spanning 600 human action classes

Task Category	Scene	Object	Action	Event	Attribute	Concept	Total
#Labels	248	1678	739	69	117	291	3142
#Annotations	672,622	3,418,198	1,473,216	245,868	581,449	1,108,552	7,499,905
#Videos	251,794	471,068	479,568	164,924	316,040	410,711	481,417

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

Dataset	Scene	Object	Action	Event	Attribute	Concept	#Videos	Year
HMDB51 [29]	-	-	51	-	-	-	7K	'11
UCF101 [47]	-	-	101	-	-	-	13K	'12
ActivityNet [5]	-	-	200	-	-	-	20K	'15
AVA [23]	-	-	80	-	-	-	57.6K	'18
Something-Something [22]	-	-	174	-	-	-	108K	'17
HACS [64]	-	-	200	-	-	-	140K	'19
Kinetics [27]	-	-	600	-	-	-	500K	'17
EPIC-KITCHEN [9]	-	323	149	-	-	-	39.6K	'18
SOA [36]	49	356	148	-	-	-	562K	'18
HVU (**Ours**)	248	1678	739	69	117	291	572K	'20

Table 2: Comparison of the HVU dataset with other publicly available video recognition datasets in terms of #labels per category. Note that SOA is not publicly available.

with more than 400 examples for each class. The current experimental strategy is to first pre-train models on these large-scale video datasets [5, 26, 27] from scratch and then fine-tune them on small-scale datasets [29, 47] to analyze their transfer behavior. Recently, a few other action datasets have been introduced with more samples, temporal duration and the diversity of category taxonomy, they are HACS [64], AVA [23], Charades [45] and Something-Something [22]. Sports-1M [26] and YouTube-8M [3] are the video datasets with million-scale samples. They consist quite longer videos rather than the other datasets and their annotations are provided in video-level and not temporally stamped. YouTube-8M labels are machine-generated without any human verification in the loop and Sports-1M is just focused on sport activities.

A similar spirit of HVU is observed in SOA dataset [36]. SOA aims to recognize visual concepts, such as scenes, objects and actions. In contrast, HVU has several orders of magnitude more semantic labels(6 times larger than SOA) and not just limited to scenes, objects, actions only, but also including events, attributes, and concepts. Our HVU dataset can help the computer vision community and bring more attention to holistic video understanding as a comprehensive, multi-faceted problem. Noticeably, the SOA paper was published in 2018, however the dataset is not released while our dataset is ready to become publicly available.

Motivated by efforts in large-scale benchmarks for object recognition in static images, i.e. the Large Scale Visual Recognition Challenge (ILSVRC) to learn a generic feature representation is now a back-bone to support several related vision tasks. We are driven by the same spirit towards learning a generic feature representation at the video level for holistic video understanding.

3 HVU Dataset

The HVU dataset is organized hierarchically in a semantic taxonomy of holistic video understanding. Almost all real-wold conditioned video datasets are targeting human action recognition. However, a video is not only about an action which provides a human-centric description of the video. By focusing on human-centric descriptions, we ignore the information about scene, objects, events and also attributes of the scenes or objects available in the video. While SOA [36] has categories of scenes, objects, and actions, to our knowledge it is not publicly available. Furthermore, HVU has more categories as it is shown in Table 2. One of the important research questions which is not addressed well in recent works on action recognition, is leveraging the other contextual information in a video. The HVU dataset makes it possible to assess the effect of learning and knowledge transfer among different tasks, such as enabling transfer learning of object recognition in videos to action recognition and vice-versa. In summary, HVU can help the vision community and bring more interesting solutions to holistic video understanding. Our dataset focuses on the recognition of scenes, objects, actions, attributes, events, and concepts in user generated videos. Scene, object, action and event categories definition is the same and standard as in other image and datasets. For attribute labels, we target attributes describing scenes, actions, objects or events. The concept category refers to any noun and label which present a grouping definition or related higher level in the taxonomy tree for labels of other categories.

3.1 HVU Statistics

HVU consists of **572k** videos. The number of video-clips for train, validation, and test set are **481k**, **31k** and **65k** respectively. The dataset consists of trimmed video clips. In practice, the duration of the videos are different with a maximum of 10 seconds length. HVU has 6 main categories: scene, object, action, event, attribute, and concept. In total, there are 3142 labels with approx. 7.5M annotations for the training, validation and test set. On average, there are ∼2112 annotations per label. We depict the distribution of categories with respect to the number of annotations, labels, and annotations per label in Fig. 2. 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. However, the average number of ∼2112 annotations per label is a reasonable amount of training data for each label. The scene category does not have a large amount of labels and annotations which is due to two reasons: the trimmed videos of the dataset and the short duration of the videos. This distribution is somewhat the same for the action category. The dataset statistics for each category are shown in Table 1 for the training set.

3.2 Collection and Annotation

Building a large-scale video understanding dataset is a time-consuming task. In practice, there are two main tasks which are usually most time consuming for creating a large-scale video dataset: (a) data collection and (b) data annotation. Recent popular datasets, such as ActivityNet, Kinetics, and YouTube-8M are collected from Internet sources like YouTube. For the annotation of these datasets, usually a semi-automatic crowdsourcing strategy is used, in which a human manually verifies the crawled videos from the web. We adopt a similar strategy with difference in the technical approach to reduce the cost of data collection and annotation. Since, we are interested in the user generated videos, thanks to the taxonomy diversity of YouTube-8M [3], Kinetics-600 [27] and HACS [64], we use these datasets as main source of the HVU. By using these datasets as the source, we also do not have to deal with copyright or privacy issues so we can publicly release the dataset. Moreover, this ensures that none of the test videos of existing datasets is part of the training set of HVU. Note that, all of the aforementioned datasets are action recognition datasets.

Manually annotating a large number of videos with multiple semantic categories (i.e thousands of concepts and tags) has two major shortcomings, (a) manual annotations are error-prone because a human cannot be attentive to every detail occurring in the video that leads to mislabeling and are difficult to eradicate; (b) large scale video annotation in specific is a very time consuming task due to the amount and temporal duration of the videos. To overcome these issues, we employ a two-stage framework for the HVU annotation. In the first stage, we utilize the Google Vision API [1] and Sensifai Video Tagging API [2] to get rough annotations of the videos. The APIs predict 30 tags per video. We keep the probability threshold of the APIs relative low ($\sim 30\%$) as a guarantee to avoid false rejects of tags in the video. The tags were chosen from a dictionary with almost 8K words. This process resulted in almost 18 million tags for the whole dataset. In the second stage, we apply human verification to remove any possible mislabeled noisy tags and also add possible missing tags missed by the APIs from some recommended tags of similar videos. The human annotation step resulted in 9 million tags for the whole dataset with $\sim$3500 different tags.

We provide more detailed statistics and discusion regarding the annotation process in the supplementary materials.

3.3 Taxonomy

Based on the predicted tags from the Google and the Sensifai APIs, we found that the number of obtained tags is approximately $\sim$8K before cleaning. The services can recognize videos with tags spanning over categories of scenes, objects, events, attributes, concepts, logos, emotions, and actions. As mentioned earlier, we remove tags with imbalanced distribution and finally, refine the tags to get the final taxonomy by using the WordNet [32] ontology. The refinement and pruning process aims to preserve the true distribution of labels. Finally, we

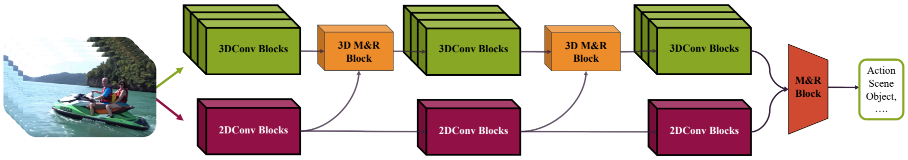

Fig. 4: HATNet: A new 2D/3D deep neural network with 2DConv, 3DConv blocks and merge and reduction (M&R) block to fuse 2D and 3D feature maps in intermediate stages of the network. HATNet combines the appearance and temporal cues with the overall goal to compress them into a more compact representation.

ask the human annotators to classify the tags into 6 main semantic categories, which are scenes, objects, actions, events, attributes and concepts.

In fact, each video can be assigned to multiple semantic categories. Almost 100K of the videos have all of the semantic categories. In comparison to SOA, almost half of HVU videos have labels for scene, object and action together. Figure 3 shows the percentage of the different subsets of the main categories.

4 Holistic Appearance and Temporal Network

We first briefly discuss state-of-the-art 3D ConvNets for video classification and then propose our new proposed “Holistic Appearance and Temporal Network” (HATNet) for multi-task and multi-label video classification.

4.1 3D-ConvNets Baselines

3D ConvNets are designed to handle temporal cues available in video clips and are shown to be efficient performance-wise for video classification. 3D ConvNets exploit both spatial and temporal information in one pipeline. In this work, we chose 3D-ResNet [51] and STCnet [10] as our 3D CNNs baseline which have competitive results on Kinetics and UCF101. To measure the performance on the multi-label HVU dataset, we use mean average precision (mAP) over all labels. We also report the individual performance on each category separately. The comparison between all of the methods can be found in Table 3. These networks are trained with binary cross entropy loss.

4.2 Multi-Task Learning 3D-ConvNets

Another approach which is studied in this work to tackle the HVU dataset is to have the problem solved with multi-task learning or a joint training method. As we know the HVU dataset consists of high-level categories like objects, scenes, events, attributes, and concepts, so each of these categories can be dealt like separate tasks. In our experiments, we have defined six tasks, scene, object, action, event, attribute, and concept classification. So our multi-task learning network is trained with six objective functions, that is with multi-label classification for each task. The trained network is a 3D-ConvNet which has separate Conv layers as separate heads for each of the tasks at the end of the network.

For each head we use the binary cross entropy loss since it is a multi-label classification for each of the categories.

4.3 2D/3D HATNet

Our "Holistic Appearance and Temporal Network" (HATNet) is a spatio-temporal neural network, which extracts temporal and appearance information in a novel way to maximize engagement of the two sources of information and also the efficiency of video recognition. The motivation of proposing this method is deeply rooted in a need of handling different levels of concepts in holistic video recognition. Since we are dealing with still objects, dynamic scenes, different attributes and also different human activities, we need a deep neural network that is able to focus on different levels of semantic information. We propose a flexible method to use a 2D pre-trained model on a large image dataset like ImageNet and a 3D pre-trained model on video datasets like Kinetics to fasten the process of training but the model can be trained from scratch as it is shown in our experiments as well. The proposed HATNet is capable of learning a hierarchy of spatio-temporal feature representation using appearance and temporal neural modules.

Appearance Neural Module. In HATNet design, we use 2D ConvNets with 2D Convolutional (2DConv) blocks to extract static cues of individual frames in a video-clip. Since we aim to recognize objects, scenes and attributes alongside of actions, it is necessary to have this module in the network which can handle these concepts better. Specifically, we use 2DConv to capture the spatial structure in the frame.

Temporal Neural Module. In HATNet architecture, the 3D Convolutions (3DConv) module handles temporal cues dealing with interaction in a batch of frames. 3DConv aims to capture the relative temporal information between frames. It is crucial to have 3D convolutions in the network to learn relational motion cues for efficiently understanding dynamic scenes and human activities. We use ResNet18/50 for both the 3D and 2D modules, so that they have the same spatial kernel sizes, and thus we can combine the output of the appearance and temporal branches at any intermediate stage of the network.

Figure 4 shows how we combine the 2DConv and 3DConv branches and use merge and reduction blocks to fuse feature maps at the intermediate stages of HATNet. Intuitively, combining the appearance and temporal features are complementary for video understanding and this fusion step aims to compress them into a more compact and robust representation. In the experiment section, we discuss in more detail about the HATNet design and how we apply merge and reduction modules between 2D and 3D neural modules. Supported by our extensive experiments, we show that HATNet complements the holistic video recognition, including understanding the dynamic and static aspects of a scene and also human action recognition. In our experiments, we have also performed tests on HATNet based multi-task learning similar to 3D-ConvNets

Model	Scene	Object	Action	Event	Attribute	Concept	HVU Overall %
3D-ResNet	50.6	28.6	48.2	35.9	29	22.5	35.8
3D-STCNet	51.9	30.1	50.3	35.8	29.9	22.7	36.7
HATNet	**55.8**	**34.2**	**51.8**	**38.5**	**33.6**	**26.1**	**40**

Table 3: MAP (%) performance of different architecture on the HVU dataset. The backbone ConvNet for all models is ResNet18.

Model	Scene	Object	Action	Event	Attribute	Concept	Overall
3D-ResNet (Standard)	50.6	28.6	48.2	35.9	29	22.5	35.8
HATNet (Standard)	55.8	34.2	51.8	38.5	33.6	26.1	40
3D-ResNet (Multi-Task)	51.7	29.6	48.9	36.6	31.1	24.1	37
HATNet (Multi-Task)	**57.2**	**35.1**	**53.5**	**39.8**	**34.9**	**27.3**	**41.3**

Table 4: Multi-task learning performance (mAP (%) comparison of 3D-ResNet18 and HATNet, when trained on HVU with all categories in the multi-task pipeline. The backbone ConvNet for all models is ResNet18.

based multi-task learning discussed in Section 4.2. HATNet has some similarity to the SlowFast [15] network but there are major differences. SlowFast uses two 3D-CNN networks for a slow and a fast branch. HATNet has one 3D-CNN branch to handle motion and dynamic information and one 2D-CNN to handle static information and appearance. HATNet also has skip connections with M&R blocks between 3D and 2D convolutional blocks to exploit more information.

2D/3D HATNet Design. The HATNet includes two branches: first is the 3D-Conv blocks with merging and reduction block and second branch is 2D-Conv blocks. After each 2D/3D blocks, we merge the feature maps from each block and perform a channel reduction by applying a $1 \times 1 \times 1$ convolution. Given the feature maps of the first block of both 2DConv and 3DConv, that have 64 channels each. We first concatenate these maps, resulting in 128 channels, and then apply $1 \times 1 \times 1$ convolution with 64 kernels for channel reduction, resulting in an output with 64 channels. The merging and reduction is done in the 3D and 2D branches, and continues independently until the last merging with two branches.

We employ 3D-ResNet and STCnet [10] with ResNet18/50 as the HATNet backbone in our experiments. The STCnet is a model of 3D networks with spatio-temporal channel correlation modules which improves 3D networks performance significantly. We also had to make a small change to the 2D branch and remove pooling layers right after the first 2D Conv to maintain a similar feature map size between the 2D and 3D branches since we use 112×112 as input-size.

5 Experiments

In this section, we demonstrate the importance of HVU on three different tasks: video classification, video captioning and video clustering. First, we introduce

Pre-Training Dataset	UCF101	HMDB51	Kinetics
From Scratch	65.2	33.4	65.6
Kinetics	89.8	62.1	-
HVU	**90.5**	**65.1**	**67.8**

Table 5: Performance (mAP (%)) comparison of HVU and Kinetics datasets for transfer learning generalization ability when evaluated on different action recognition dataset. The trained model for all of the datasets is 3D-ResNet18.

Method	**Pre-Trained Dataset**	**CNN Backbone**	**UCF101**	**HMDB51**	**Kinetics-400**	**Kinetics-600**
Two Stream (spatial stream) [46]	Imagenet	VGG-M	73	40.5	-	
RGB-I3D [6]	Imagenet	Inception v1	84.5	49.8	-	
C3D [50]	Sport1M	VGG11	82.3	51.6	-	
TSN [59]	Imagenet,Kinetics	Inception v3	93.2	-	72.5	
RGB-I3D [6]	Imagenet,Kinetics	Inception v1	95.6	74.8	72.1	
3D ResNext 101 (16 frames) [24]	Kinetics	ResNext101	90.7	63.8	65.1	
STC-ResNext 101 (64 frames) [10]	Kinetics	ResNext101	96.5	74.9	68.7	
ARTNet [57]	Kinetics	ResNet18	93.5	67.6	69.2	
R(2+1)D [53]	Kinetics	ResNet50	96.8	74.5	72	
ir-CSN-101 [52]	Kinetics	ResNet101	-	-	76.7	
DynamoNet [13]	Kinetics	ResNet101	-	-	76.8	
SlowFast 4×16 [15]	Kinetics	ResNet50	-	-	75.6	78.8
SlowFast 16×8* [15]	Kinetics	ResNet101	-	-	78.9*	81.1
HATNet (32 frames)	Kinetics	ResNet50	96.8	74.8	77.2	80.2
HATNet (32 frames)	HVU	ResNet18	96.9	74.5	74.2	77.4
HATNet (16 frames)	HVU	ResNet50	96.5	73.4	76.3	79.4
HATNet (32 frames)	HVU	ResNet50	**97.8**	**76.5**	**79.3**	**81.6**

Table 6: State-of-the-art performance comparison on UCF101, HMDB51 test sets and Kinetics validation set. The results on UCF101 and HMDB51 are average mAP over three splits, and for Kinetics(400,600) is Top-1 mAP on validation set. For a fair comparison, here we report the performance of methods which utilize only RGB frames as input. *SlowFast uses multiple branches of 3D-ResNet with bigger backbones.

the implementation details and then show the results of each mentioned method on multi-label video recognition. Following, we compare the transfer learning ability of HVU against Kinetics. Next, as an additional experiment, we show the importance of having more categories of tags such as scenes and objects for video classification. Finally, we show the generalisation capability of HVU for video captioning and clustering tasks. For each task, we test and compare our method with the state-of-the-art on benchmark datasets. For all experiments, we use RGB frames as input to the ConvNet. For training, we use 16 or 32 frames long video clips as single input. We use PyTorch framework for implementation and all the networks are trained on a machine with 8 V100 NVIDIA GPUs.

5.1 HVU Results

In Table 3, we report the overall performance of different simpler or multi-task learning baselines and HATNet on the HVU validation set. The reported performance is mean average precision on all of the labels/tags. HATNet that exploits both appearance and temporal information in the same pipeline achieves the best performance, since recognizing objects, scenes and attributes need an appearance module which other baselines do not have. With HATNet, we show that combining the 3D (temporal) and 2D (appearance) convolutional blocks one can learn a more robust reasoning ability.

5.2 Multi-Task Learning on HVU

Since the HVU is a multi-task classification dataset, it is interesting to compare the performance of different deep neural networks in the multi-task learning paradigm as well. For this, we have used the same architecture as in the previous experiment, but with different last layer of convolutions to observe multi-task learning performance. We have targeted six tasks: scene, object, action, event, attribute, and concept classification. In Table 4, we have compared standard training without multi-task learning heads versus multi-task learning networks.

The simple baseline multi-task learning methods achieve higher performance on individual tasks as expected, in comparison to standard networks learning for all categories as a single task. Therefore this initial result on a real-world multi-task video dataset motivates the investigation of more efficient multi-task learning methods for video classification.

5.3 Transfer Learning: HVU vs Kinetics

Here, we study the ability of transfer learning with the HVU dataset. We compare the results of pre-training 3D-ResNet18 using Kinetics versus using HVU and then fine-tuning on UCF101, HMDB51 and Kinetics. Obviously, there is a large benefit from pre-training of deep 3D-ConvNets and then fine-tune them on smaller datasets (i.e. HVU, Kinetics ⇒ UCF101 and HMDB51). As it can be observed in Table 5, models pre-trained on our HVU dataset performed notably better than models pre-trained on the Kinetics dataset. Moreover, pre-training on HVU can improve the results on Kinetics also.

5.4 Benefit of Multiple Semantic Categories

Here, we study the effect of training models with multiple semantic categories, in comparison to using only a single semantic category, such as Kinetics which covers only action category. In particular, we designed an experiment by having the model trained in multiple steps by adding different categories of tags one by one. Specifically, we first train 3D-ResNet18 with action tags of HVU, following in second step we add tags from object category and in the last step we add tags from the scene category. For performance evaluation, we consider action category of HVU. In the first step the gained performance was 43.6% accuracy and after second step it was improved to 44.5% and finally in the last step it raised to 45.6%. The results show that adding high-level categories to the training, boosts the performance for action recognition in each step. As it was also shown in Table 4, training all the categories together yields 47.5% for the action category which is ∼4% gain over action as single category for training. Thus we can

Model	Pre-Training Dataset	BLEU@4
SA(VGG+C3D) [62]	ImageNet+Sports1M	36.6
M3(VGG+C3D) [56]	ImageNet+Sports1M	38.1
SibNet(GoogleNet) [31]	ImageNet	40.9
MGSA(Inception+C3D) [7]	ImageNet+Sports1M	42.4
I3D+M [54]	Kinetics	41.7
3D-ResNet50+M	Kinetics	41.8
3D-ResNet50+M	HVU	**42.7**

Table 7: Captioning performance comparisons of [54] with different models and pre-training datasets. M denotes the motion features from optical flow extracted as in the original paper.

infer from this that an effective feature representation can be learned by adding additional categories, and also acquire knowledge for an in-depth understanding of the video in holistic sense.

5.5 Comparison on UCF, HMDB, Kinetics

In Table 6, we compare the HATNet performance with the state-of-the-art on UCF101, HMDB51 and Kinetics. For our baselines and HATNet, we employ pre-training in two separate setups: one with HVU and another with Kinetics, and then fine-tune on the target datasets. For UCF101 and HMDB51, we report the average accuracy over all three splits. We have used ResNet18/50 as backbone model for all of our networks with 16 and 32 input-frames. HATNet pre-trained on HVU with 32 frames input achieved superior performance on all three datasets with standard network backbones. Note that on Kinetics, HATNet even with ResNet18 as a backbone ConvNet performs almost comparable to SlowFast which is trained by dual 3D-ResNet50. In Table 6, however while SlowFast has better performance using dual 3D-ResNet101 architecture, HATNet obtains comparable results with much smaller backbone.

5.6 Video Captioning

We present a second task that showcases the effectiveness of our HVU dataset, we evaluate the effectiveness of HVU for video captioning task. We conduct experiments on a large-scale video captioning dataset, namely MSR-VTT [62]. We follow the standard training/testing splits and protocols provided originally in [62]. For video captioning, the performance is measured using the BLEU metric.

Method and Results: Most of the state-of-the-art video captioning methods use models pre-trained on Kinetics or other video recognition datasets. With this experiment, we intend to show another generalisation capability of HVU dataset where we evaluate the performance of pre-trained models trained on HVU against Kinetics. For our experiment, we use the Controllable Gated Network [54] method, which is to the best of our knowledge the state-of-the-art for captioning task.

Model	Pre-Training Dataset	Clustering Accuracy (%)
3D-ResNet50	Kinetics	50.3
3D-ResNet50	HVU	53.5
HATNet	HVU	**54.8**

Table 8: Video clustering performance: evaluation based on extracted features from networks pre-trained on Kinetics and HVU datasets.

For comparison, we considered two models of 3D-ResNet50, pre-trained on (i) Kinetics and (ii) HVU. Table 7 shows that the model trained on HVU obtained better gains in comparison to Kinetics. This shows HVU helps to learn more generic video representation to achieve better performance in other tasks.

5.7 Video Clustering

With this experiment, we evaluate the effectiveness of generic features learned using HVU when compared to Kinetics.

Dataset: We conduct experiments on ActivityNet-100 [5] dataset. For this experiment we provide results when considering 20 action categories with 1500 test videos. We have selected ActivityNet dataset to make sure there are no same videos in HVU and Kinetics training set. For clustering, the performance is measured using clustering accuracy [41].

Method and Results: We extract features using 3D-ResNet50 and HATNet pre-trained on Kinetics-600 and HVU for the test videos and then cluster them with KMeans clustering algorithm with the given number of action categories. Table 8 clearly shows that the features learned using HVU is far more effective compared to features learned using Kinetics.

6 Conclusion

This work presents the “Holistic Video Understanding Dataset” (HVU), a large-scale multi-task, multi-label video benchmark dataset with comprehensive tasks and annotations. It contains 572k videos in total with 9M annotations, which is richly labeled over 3142 labels encompassing scenes, objects, actions, events, attributes and concepts categories. Through our experiments, we show that the HVU can play a key role in learning a generic video representation via demonstration on three real-world tasks: video classification, video captioning and video clustering. Furthermore, we present a novel network architecture, HATNet, that combines 2D and 3D ConvNets in order to learn a robust spatio-temporal feature representation via multi-task and multi-label learning in an end-to-end manner. We believe that our work will inspire new research ideas for holistic video understanding. For the future plan, we are going to expand the dataset to 1 million videos with similar rich semantic labels and also provide annotations for other important tasks like activity and object detection and video captioning.

Acknowledgements: This work was supported by DBOF PhD scholarship & GC4 Flemish AI project, and the ERC Starting Grant ARCA (677650). We also would like to thank Sensifai for giving us access to the Video Tagging API for dataset preparation.

References

1. Google vision ai api. cloud.google.com/vision
2. Sensifai video tagging api. www.sensifai.com
3. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. arXiv:1609.08675 (2016)
4. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR (2014)
5. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR (2015)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
7. Chen, S., Jiang, Y.G.: Motion guided spatial attention for video captioning. In: AAAI (2019)
8. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: ECCV (2006)
9. Damen, D., Doughty, H., Maria Farinella, G., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV (2018)
10. Diba, A., Fayyaz, M., Sharma, V., Arzani, M.M., Yousefzadeh, R., Gall, J., Van Gool, L.: Spatio-temporal channel correlation networks for action classification. In: ECCV (2018)
11. Diba, A., Fayyaz, M., Sharma, V., Karami, A.H., Arzani, M.M., Yousefzadeh, R., Van Gool, L.: Temporal 3d convnets using temporal transition layer. In: CVPR Workshops (2018)
12. Diba, A., Sharma, V., Van Gool, L.: Deep temporal linear encoding networks. In: CVPR (2017)
13. Diba, A., Sharma, V., Van Gool, L., Stiefelhagen, R.: Dynamonet: Dynamic action and motion network. In: ICCV (2019)
14. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR (2015)
15. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. ICCV (2019)
16. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: CVPR (2016)
17. Fernando, B., Bilen, H., Gavves, E., Gould, S.: Self-supervised video representation learning with odd-one-out networks. In: CVPR (2017)
18. Fernando, B., Gavves, E., Oramas, J.M., Ghodrati, A., Tuytelaars, T.: Modeling video evolution for action recognition. In: CVPR (2015)
19. Gaidon, A., Harchaoui, Z., Schmid, C.: Temporal localization of actions with actoms. PAMI (2013)
20. Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., Russell, B.: Actionvlad: Learning spatio-temporal aggregation for action classification. In: CVPR (2017)
21. Girdhar, R., Tran, D., Torresani, L., Ramanan, D.: Distinit: Learning video representations without a single labeled video. In: ICCV (2019)
22. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: ICCV (2017)

23. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: CVPR (2018)
24. Hara, K., Kataoka, H., Satoh, Y.: Learning spatio-temporal features with 3d residual networks for action recognition. In: ICCV (2017)
25. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
26. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014)
27. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset. arXiv:1705.06950 (2017)
28. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC (2008)
29. Kuehne, H., Jhuang, H., Stiefelhagen, R., Serre, T.: Hmdb51: A large video database for human motion recognition. In: High Performance Computing in Science and Engineering (2013)
30. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR (2008)
31. Liu, S., Ren, Z., Yuan, J.: Sibnet: Sibling convolutional encoder for video captioning. In: ACMM (2018)
32. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM (1995)
33. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: ECCV (2016)
34. Ng, J.Y.H., Choi, J., Neumann, J., Davis, L.S.: Actionflownet: Learning motion representation for action recognition. In: WACV (2018)
35. Niebles, J.C., Chen, C.W., Fei-Fei, L.: Modeling temporal structure of decomposable motion segments for activity classification. In: ECCV (2010)
36. Ray, J., Wang, H., Tran, D., Wang, Y., Feiszli, M., Torresani, L., Paluri, M.: Scenes-objects-actions: A multi-task, multi-label video dataset. In: ECCV (2018)
37. Roethlingshoefer, V., Sharma, V., Stiefelhagen, R.: Self-supervised face-grouping on graph. In: ACMMM (2019)
38. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: ICPR (2004)
39. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: ACM MM (2007)
40. Sharma, V., Sarfraz, S., Stiefelhagen, R.: A simple and effective technique for face clustering in tv series. In: CVPR workshop on Brave New Motion Representations (2017)
41. Sharma, V., Tapaswi, M., Sarfraz, M.S., Stiefelhagen, R.: Self-supervised learning of face representations for video face clustering. In: International Conference on Automatic Face and Gesture Recognition (2019)
42. Sharma, V., Tapaswi, M., Sarfraz, M.S., Stiefelhagen, R.: Video face clustering with self-supervised representation learning. IEEE Transactions on Biometrics, Behavior, and Identity Science (2019)
43. Sharma, V., Tapaswi, M., Sarfraz, M.S., Stiefelhagen, R.: Clustering based contrastive learning for improving face representations. In: International Conference on Automatic Face and Gesture Recognition (2020)
44. Sharma, V., Tapaswi, M., Stiefelhagen, R.: Deep multimodal feature encoding for video ordering. In: ICCV workshop on Holistic Video Understanding (2019)

45. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: ECCV (2016)
46. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS (2014)
47. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402 (2012)
48. Sun, L., Jia, K., Yeung, D.Y., Shi, B.E.: Human action recognition using factorized spatio-temporal convolutional networks. In: ICCV (2015)
49. Tang, P., Wang, X., Shi, B., Bai, X., Liu, W., Tu, Z.: Deep fishernet for object classification. arXiv:1608.00182 (2016)
50. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015)
51. Tran, D., Ray, J., Shou, Z., Chang, S.F., Paluri, M.: Convnet architecture search for spatiotemporal feature learning. arXiv:1708.05038 (2017)
52. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks. ICCV (2019)
53. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR (2018)
54. Wang, B., Ma, L., Zhang, W., Jiang, W., Wang, J., Liu, W.: Controllable video captioning with pos sequence guidance based on gated fusion network. In: ICCV (2019)
55. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV (2013)
56. Wang, J., Wang, W., Huang, Y., Wang, L., Tan, T.: M3: Multimodal memory modelling for video captioning. In: CVPR (2018)
57. Wang, L., Li, W., Li, W., Van Gool, L.: Appearance-and-relation networks for video classification. In: CVPR (2018)
58. Wang, L., Qiao, Y., Tang, X.: Video action detection with relational dynamic-poselets. In: ECCV (2014)
59. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: towards good practices for deep action recognition. In: ECCV (2016)
60. Wei, D., Lim, J., Zisserman, A., Freeman, W.T.: Learning and using the arrow of time. In: CVPR (2018)
61. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: ECCV (2008)
62. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: CVPR (2016)
63. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: CVPR (2015)
64. Zhao, H., Yan, Z., Torresani, L., Torralba, A.: Hacs: Human action clips and segments dataset for recognition and temporal localization. arXiv:1712.09374 (2019)