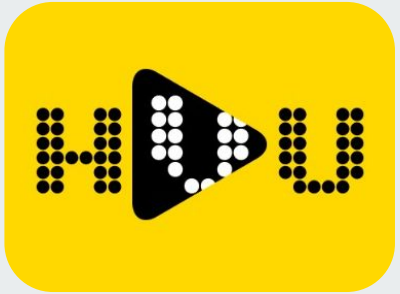
UNIVERSITÄT BONN

Holistic Large Scale Video Understanding

Ali Diba*, Mohsen Fayyaz*, Vivek Sharma*, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, Luc Van Gool
*contributed equally to this work and listed in alphabetical order.



KU LEUVEN

UNIVERSITÄT BONN

KIT Karlsruher Institut für Technologie

facebook

ETH zürich

Video Recognition

- Datasets
  - Human Action Recognition
  - Sports Recognition
- Methods
  - Action Recognition

Datasets - HMDB51

The dataset contains 6849 clips divided into 51 action categories, each containing a minimum of 101 clips.

H. Kuehne, H. Jhuang, R. Stiefelhagen, and T. Serre. Hmdb51: A large video database for human motion recognition. In High Performance Computing in Science and Engineering. 2013

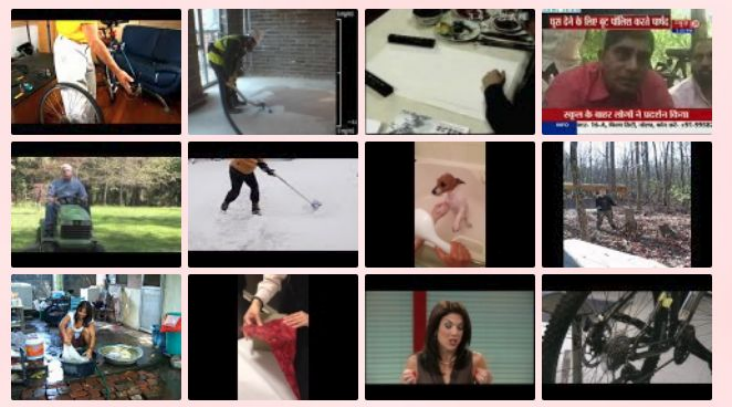
Datasets - UCF101

UCF101 is an action recognition data set of realistic action videos, collected from YouTube, having 101 action categories with 13320 videos from 101 action categories

K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402, 2012

Datasets - ActivityNet

ActivityNet provides samples from 203 activity classes with an average of 137 untrimmed videos per class and 1.41 activity instances per video, for a total of 849 video hours.



F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In CVPR, 2015.

Datasets - AVA

The AVA dataset densely annotates 80 atomic visual actions in 430 15-minute movie clips, where actions are localized in space and time, resulting in 1.62M action labels with multiple labels per human occurring frequently.

C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik. Ava: A video dataset of spatiotemporally localized atomic visual actions. In CVPR, 2018.

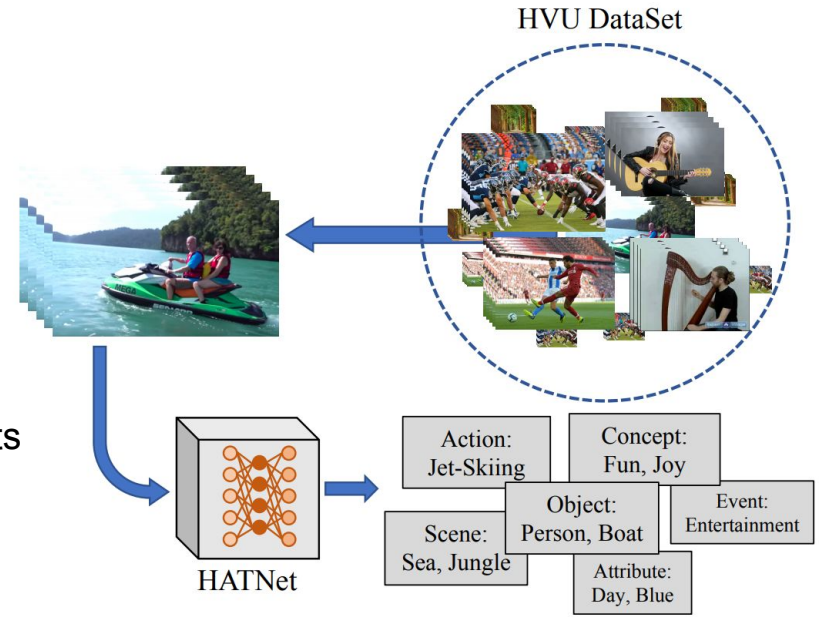
Datasets - Kinetics

Kinetics-700 includes a diverse range of human focused actions. The dataset consists of approximately 650,000 video clips, and covers 700 human action classes with at least 600 video clips for each action class. Each clip lasts around 10 seconds and is labeled with a single class.

W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. arXiv:1705.06950, 2017.

Holistic Video Understanding Dataset - HVU

- Multi-label and multi-task video understanding
- 577k videos
- 13M annotations for training and validation set spanning over 4378 classes
- Main categories: scenes, objects, actions, events, attributes and concepts



HVU Statistics

Train	Validation	Test
481k	31k	65k

HVU dataset statistics i.e. #videos-clips for train, validation, and test sets.

Task Category	Scene	Object	Action	Event	Attribute	Concept	Total
#Classes	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

Statistics of the HVU training set for different categories. The category with most number of classes and annotations is the object category.

HVU Collecting and Annotation

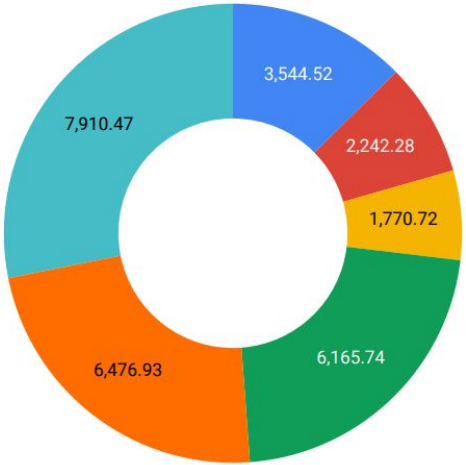
Collecting Videos	Annotation	Verification
Thanks to the category taxonomy diversity of Youtube8M, Kinetics-600 and HACS, we have used these datasets as main source of the HVU.	We have employed a semi-automatic method for annotation. We have used the Sensifai Video Tagging API to get rough annotations of the videos, which predicts multiple tags (or class labels) for each video.	Expert human annotators verify the relevance of the tags to their corresponding video for the validation and test sets.

HVU Taxonomy

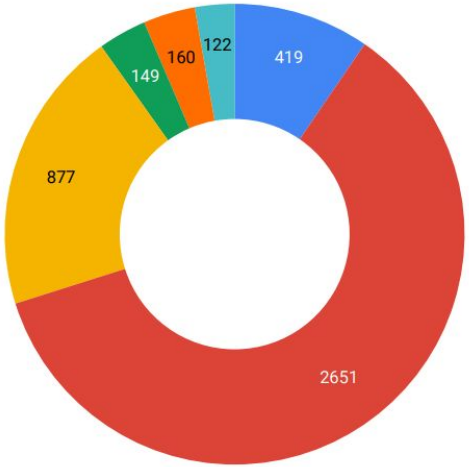
- We prune tags with imbalanced distribution and finally, refine the tags to get the final taxonomy.
- The refinement and pruning process was aimed to preserve the true distribution of labels.
- Finally, we ask the human annotators to classify the tags in to 6 main semantic categories, they are scenes, objects, actions, events, attributes and concepts.
- Moreover, it is important to note that each video may be assigned to multiple semantic categories.
- About 36% of the videos have all of the categories.

t-SNE visualization of semantically related categories tend to co-occur on HVU. This embedding is purely based on class co-occurrence, without using video content.

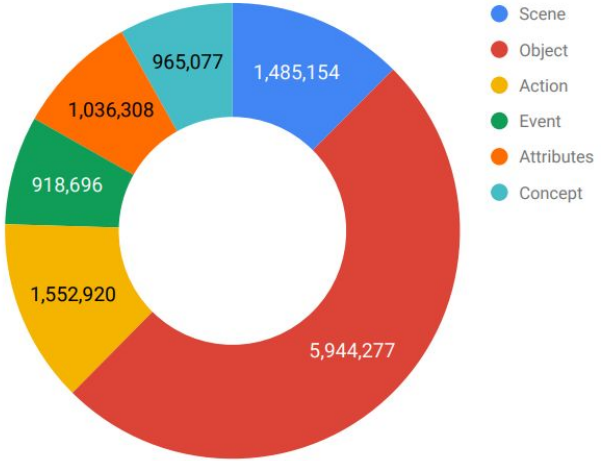
HVU Main Semantic Categories



Average #annotations per class of each cat.

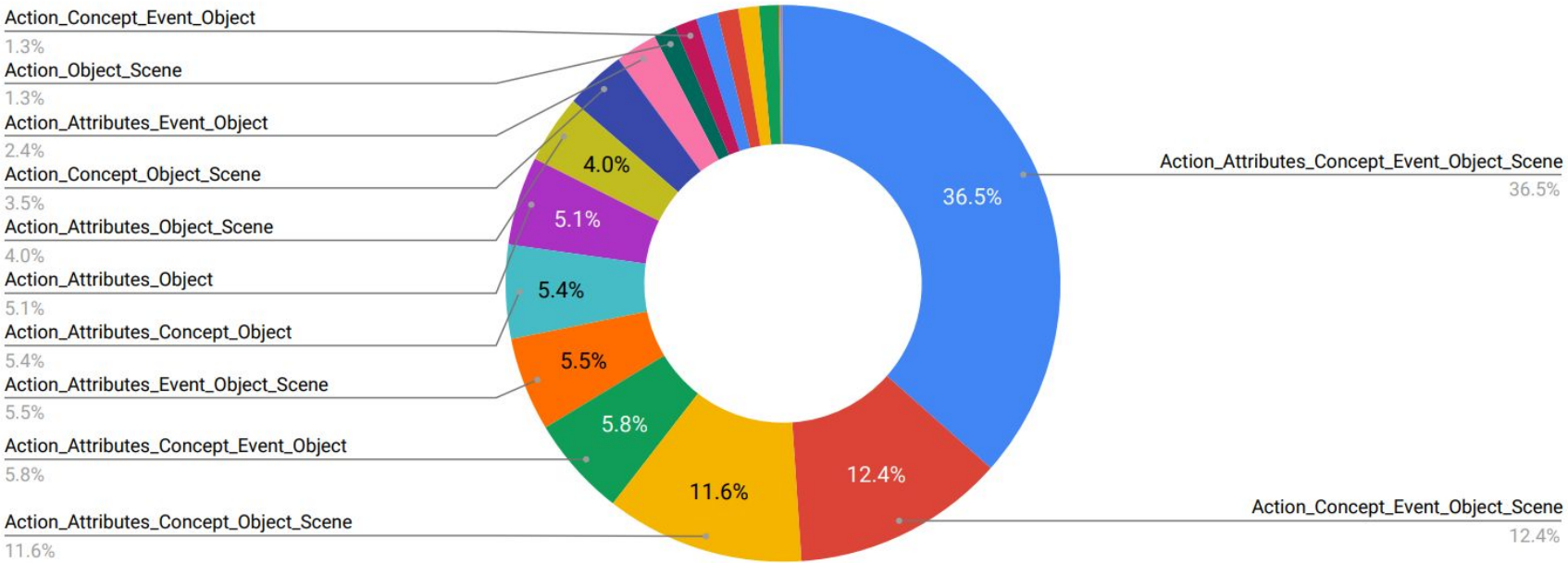


#Classes of each category over all classes



#Annotations of each cat. over all annotations

HVU - Coverage of Semantic Categories



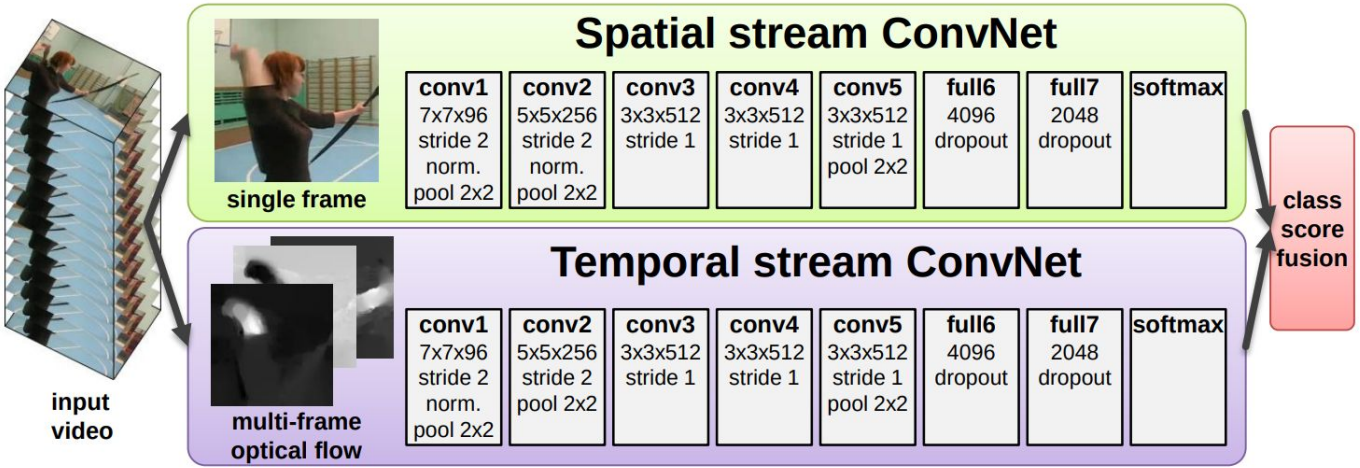
Coverage of 16 different subsets of the 6 main semantic categories in videos. 36.5% of the videos have annotations of all categories.

Comparison of HVU with other Datasets

Dataset	Scene	Object	Action	Event	Attribute	Concept	#Videos	Year
HMDB51 [24]	-	-	51	-	-	-	7K	'11
UCF101 [37]	-	-	101	-	-	-	13K	'12
ActivityNet [4]	-	-	200	-	-	-	20K	'15
AVA [18]	-	-	80	-	-	-	57.6K	'18
Something-Something [17]	-	-	174	-	-	-	108K	'17
HACS [52]	-	-	200	-	-	-	140K	'19
Kinetics [22]	-	-	600	-	-	-	500K	'17
SOA [32]	49	356	148	-	-	-	562K	'18
HVU	419	2651	877	149	160	122	577K	'19

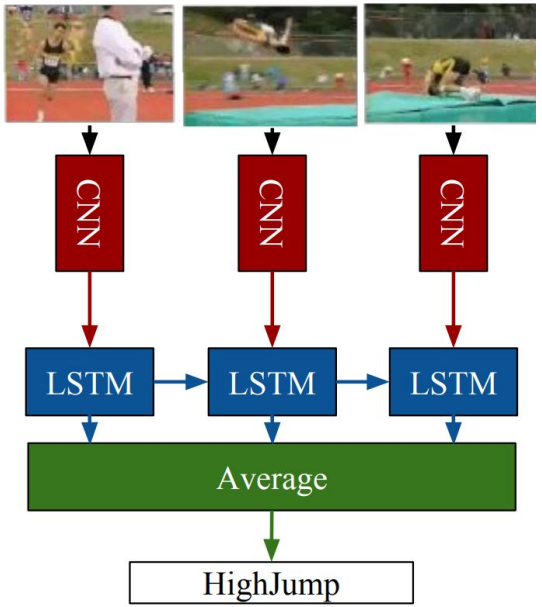
Comparison of the HVU dataset with other publicly available video recognition datasets in term of #classes per category. Note that SOA is not publicly available at this moment.

Methods - Two-Stream Convolutional Networks for Action Recognition in Videos



K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, 2014.

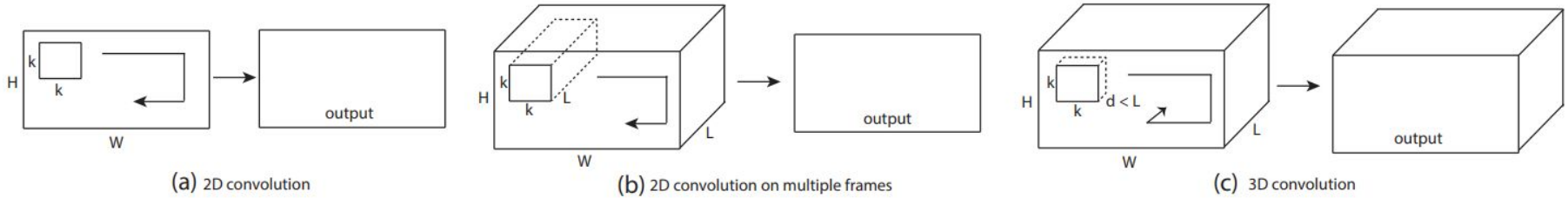
Methods - Long-term Recurrent Convolutional Networks (LRCNs)



J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In CVPR, 2015.

Methods - C3D

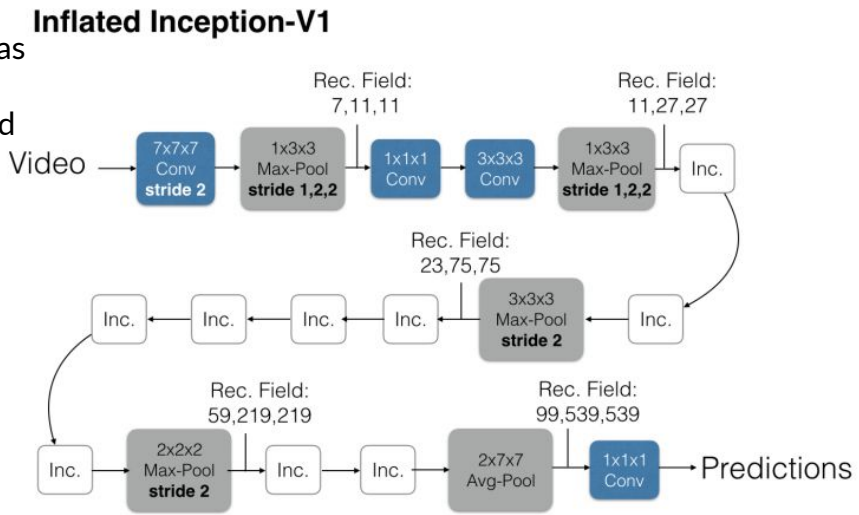
- Using 3D convolutions for video recognition

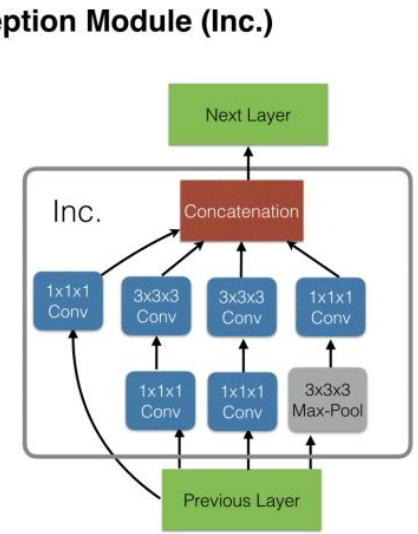


D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In ICCV, 2015.

Methods - I3D

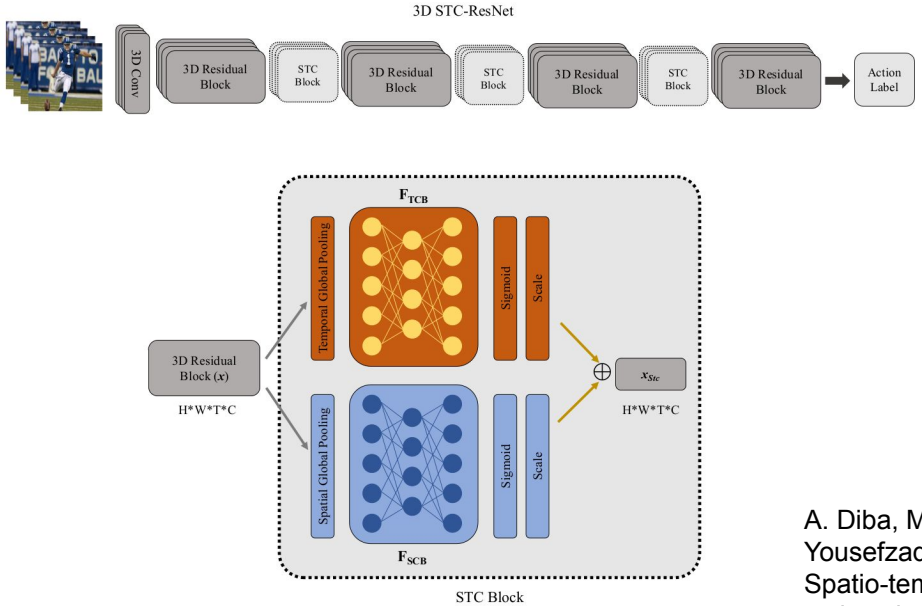
- Using inflation as a transfer learning method





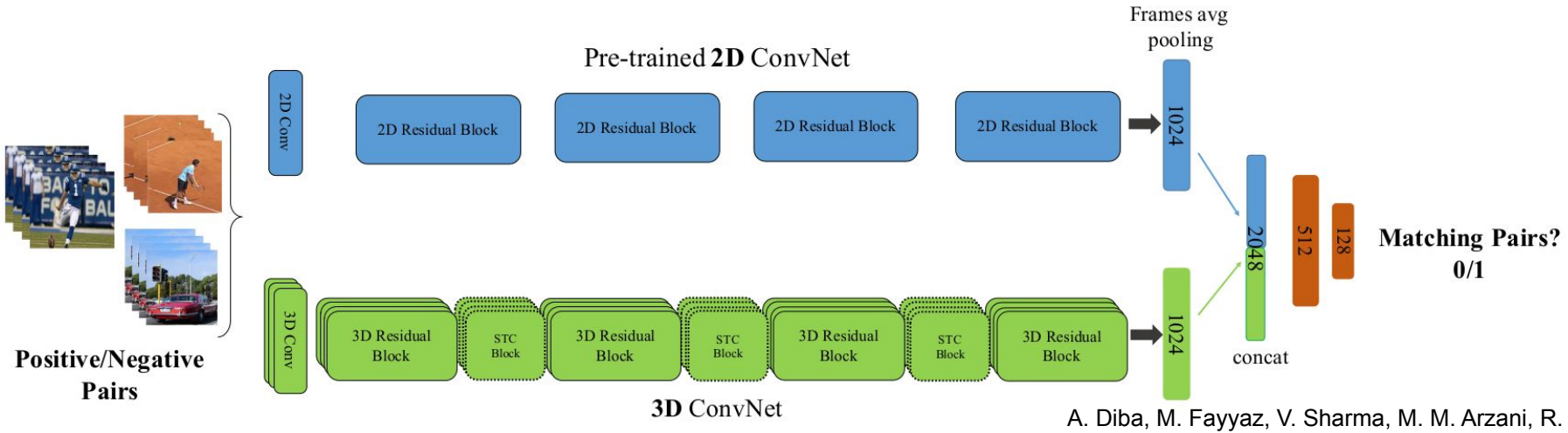
J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR, 2017.

Methods - STCNet



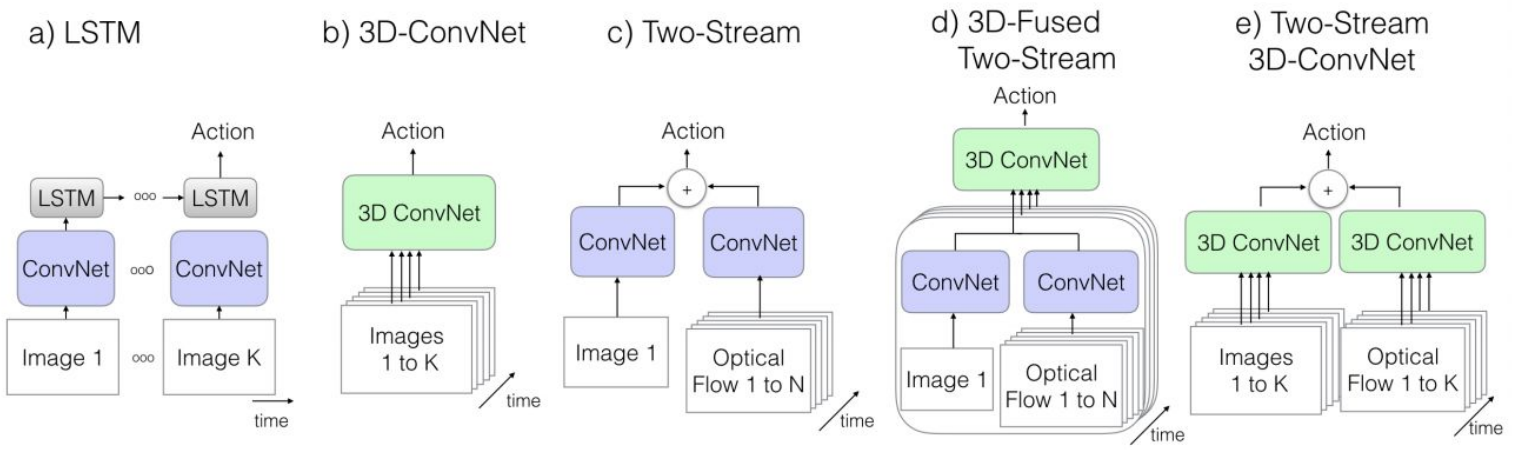
A. Diba, M. Fayyaz, V. Sharma, M. M. Arzani, R. Yousefzadeh, J. Gall, and L. Van Gool. Spatio-temporal channel correlation networks for action classification. In ECCV, 2018.

Methods - STCNet - Transfer Learning



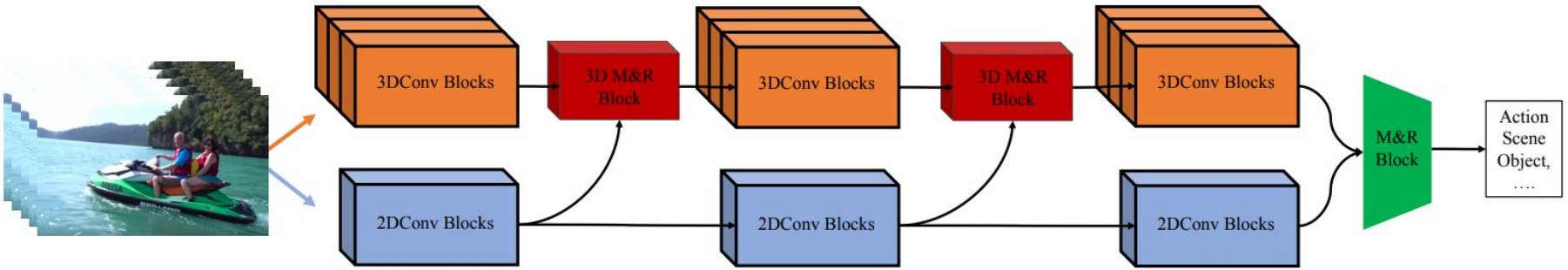
A. Diba, M. Fayyaz, V. Sharma, M. M. Arzani, R. Yousefzadeh, J. Gall, and L. Van Gool. Spatio-temporal channel correlation networks for action classification. In ECCV, 2018.

Action Recognition General Models



J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR, 2017.

Holistic Appearance and Temporal Network



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.

Comparison on HVU

Model	Scene	Object	Action	Event	Attribute	Concept	HVU Overall %
3D-ResNet	58.5	38.4	53.2	35.3	32.1	24.4	40.3
3D-STCNet	59.1	38.7	57.1	37.5	33.6	25.7	41.9
HATNet	**62**	**43.4**	**58.5**	**41.9**	**34.5**	**27.6**	**44.7**

Different architecture mAP (%) performance comparison when trained on HVU dataset. The backbone ConvNet for all models is ResNet18.

Comparison with State-of-the-Arts

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 is Top-1 mAP on validation set. For a fair comparison, in this table we report the performance of methods which utilize only RGB frames as input

Method	Pre-Trained Dataset	CNN Backbone	UCF101	HMDB51	Kinetics
Two Stream (spatial stream) [36]	Imagenet	VGG-M	73	40.5	-
Conv+LSTM [10]	Imagenet	AlexNet	68.2	-	-
TDD+FV [47]	Imagenet	VGG-M	90.3	63.2	-
RGB-I3D [5]	Imagenet	Inception v1	84.5	49.8	-
TSN [48]	Imagenet	Inception v2	86.4	53.7	-
LTC [43]	Sport1M	VGG11	82.4	48.7	-
C3D [40]	Sport1M	VGG11	82.3	51.6	-
TSN [48]	Imagenet,Kinetics	Inception v3	93.2	-	72.5
RGB-I3D [5]	Imagenet,Kinetics	Inception v1	95.6	74.8	72.1
RGB-I3D [5]	Kinetics	Inception v1	95.6	74.8	71.6
3D ResNet 101 (16 frames) [19]	Kinetics	ResNet101	88.9	61.7	62.8
3D ResNext 101 (16 frames) [19]	Kinetics	ResNext101	90.7	63.8	65.1
STC-ResNext 101 (16 frames) [7]	Kinetics	ResNext101	92.3	65.4	66.2
STC-ResNext 101 (64 frames) [7]	Kinetics	ResNext101	96.5	74.9	68.7
C3D [45]	Kinetics	ResNet18	89.8	62.1	65.6
ARTNet [45]	Kinetics	ResNet18	93.5	67.6	69.2
R(2+1)D [42]	Kinetics	ResNet50	96.8	74.5	72
SlowFast [11]	Kinetics	ResNet50	-	-	75.6
HATNet (16 frames)	Kinetics	ResNet18	94.1	69.2	70.4
3D-ResNet18 (16 frames)	HVU	ResNet18	90.4	65.1	66.9
3D-ResNet18 (32 frames)	HVU	ResNet18	90.9	66.6	67.3
HATNet (16 frames)	HVU	ResNet18	95.4	72.2	71.8
HATNet (32 frames)	HVU	ResNet18	96.9	74.5	73.9
HATNet (16 frames)	HVU	ResNet50	**96.5**	**73.4**	**74.6**
HATNet (32 frames)	HVU	ResNet50	**97.7**	**76.2**	**76.3**

https://holistic-video-understanding.github.io/workshops/iccv2019.html

HVU Workshop in ICCV'19

Speakers

Rahul Sukthankar

Kristen Grauman

Carl Vondrick

Manohar Paluri

Organizers

Vivek Sharma

Mohsen Fayyaz

Ali Diba

Luc Van Gool

Juergen Gall

Rainer Stiefelhagen

Manohar Paluri

Thanks for your attention