Machine Learning for Human Activity Recognition from Video

Shikhar Shrestha
Stanford University, CA
shikhars@stanford.edu

Abstract

Human activity recognition is a very important problem in computer vision that is still largely unsolved. While recent advances in areas such as deep learning have given us great results on image related tasks, it is still unclear as to what a good feature representation is for recognizing activities from videos. A large part of the problem is that not many clean and big enough data sets are available for this task that fairly represent real-world conditions. Recently a new dataset has been made available to the research community by Allen Institute that has enough content to effectively train a deep network and extract a suitable representation for video-native tasks. This work implements state-of-the-art deep video feature extraction on this dataset and then trains a classifier to perform Activity Recognition. The results are then compared and benchmarked.

1. Introduction

Human Activity Recognition from video is one of the most important fundamental problems in computer vision that is still largely unsolved. Robots, Drones etc. operating in real-world setting have to be able to understand what is happening in their surroundings to successfully co-operate with humans.

Recent advances in deep learning for image recognition have been very promising but performance on video native tasks like activity recognition is still not up to the mark. In this project, the motivation was to explore deep representations for video for recognizing activities using a newly open-sourced dataset and benchmark its performance against more tradition features like IDT.

Handling video data and training networks on them is a complex task due to compute and memory constraints and hence a late fusion type method was finally implemented as it could be trained in a reasonable time.

The results obtained while promising are still not at the level required for real-world deployment but provide important insights into the task and its specific requirements. Those insights have been presented as conclusions of this work along with recommendations for future directions to solve the problem.

2. Related Work

Currently available methods for activity recognition in video includes two stream networks [4] that use two simultaneous streams of data RGB frame and optical flow to incorporate temporal information with a CNN using siamese architecture. Three dimensional convolutional networks i.e. C3D can inherently incorporate multiple frames [2]. CNN + RNN methods [5] have recently performed well on video based tasks but are notoriously difficult to train because of memory constraints. Fusion based architectures [1] combine features obtained from CNN early or late depending on the method used. The method implemented in this work in inspired from the fusion methods but incorporate four instead of two frames from the video sequence.

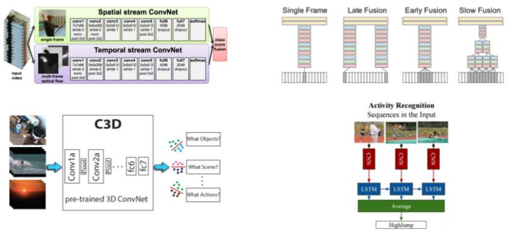


Fig. 1 Available Methods

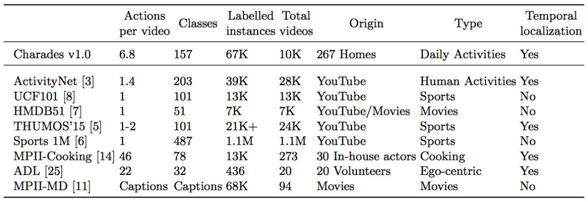
3. Dataset

Currently available datasets for human activity recognition in video like ActivityNet, UCF101, Sports 1M have high bias. Most of these datasets are built by collecting video from sites like YouTube that poorly reflect real-world conditions and under-represent boring, daily activities. Some researchers have tried to mitigate this bias by collecting video activity datasets from movies which are also an example of production video. Other datasets like the KTH action dataset have very little scene variability which is going to be a common aspect of any intelligent system operating in the real-world.

Recently the Allen Institute has made available the Charades v1.0 dataset that tries to overcome some of these biases by leveraging crowd-sourced data collection and annotation for daily activities with high scene variability.

The table below shows a comparison of the Charades dataset with other commonly known video activity recognition datasets. It consists of 157 activity classes that represent common household activities in a bank of 10K videos.

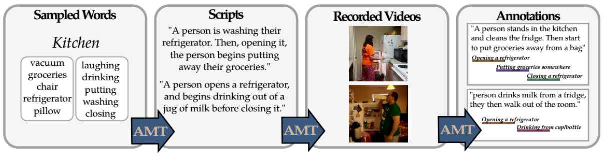
The novel approach followed for the collection of this dataset is very promising as it scales very well and would allow for the development of algorithms that more robust and precise in real-world conditions.



	Actions per video	Classes	Labelled instances	Total videos	Origin	Type	Temporal localization
Charades v1.0	6.8	157	67K	10K	267 Homes	Daily Activities	Yes
ActivityNet [3]	1.4	203	39K	28K	YouTube	Human Activities	Yes
UCF101 [8]	1	101	13K	13K	YouTube	Sports	No
HMDB51 [7]	1	51	7K	7K	YouTube/Movies	Movies	No
THUMOS'15 [5]	1-2	101	21K+	24K	YouTube	Sports	Yes
Sports 1M [6]	1	487	1.1M	1.1M	YouTube	Sports	No
MPII-Cooking [14]	46	78	13K	273	30 In-house actors	Cooking	Yes
ADL [25]	22	32	436	20	20 Volunteers	Ego-centric	Yes
MPII-MD [11]	Captions	Captions	68K	94	Movies	Movies	No

Fig. 2 Activity Recognition Datasets

Workers on AMT are provided with prompts/scripts to record video which are then verified and annotated in a distributed fashion with redundancy. The approach is depicted in Fig. 3.



"A person is washing their refrigerator. Then, opening it, the person begins putting away their groceries."

"A person opens a refrigerator, and begins drinking out of a jug of milk before closing it."

"A person stands in the kitchen and cleans the fridge. Then start to put groceries away from a bag"

"person drinks milk from a fridge, they then walk out of the room."

Fig. 3 Dataset Workflow

The 157 activity classes were identified by analyzing triplets of (verb, proposition, noun) contained in the scripts. In total, the dataset is built by combining 40 objects and 30 actions in 15 scenes.

Fig. 4 shows some sample key frames from the dataset compared to frames from other activity datasets.

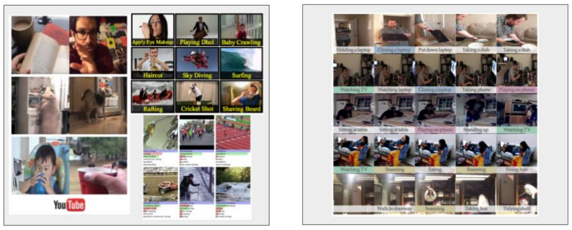


Fig. 4 Real-world video data

4. Methods

Multi-Stream late fusion architecture was implemented in this project. Imagenet pre-trained GoogleNet CNN was used to extract image features from four equidistant frames which are then pooled together using fully connected layers. During training, last two layers of the CNN was fine-tuned along with training the FC layer weights.

4.1 Implementation Framework

In this project. Google's TensorFlow platform was used for the implementation of the deep network. Several deep learning frameworks were evaluated including Caffe and Torch however, TensorFlow seems to be very well supported and now has a decent model library that includes the GoogleNet model pre-trained on ImageNet used in this project shown in Fig. 5.

This was a huge advantage as it drastically cuts down training time which usually extends into weeks. TensorFlow is based on performing graph operations on tensors.

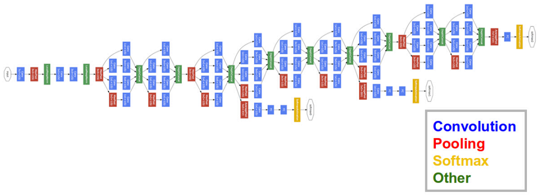


Fig. 5 GoogleNet

4.2 Model Architecture

The multi-stream late fusion model shown in Fig. 6 was trained using stochastic gradient descent solver with momentum set to 0.9. The training was run for 20 hours on AWS and then the FC layers were stripped away and features of 1024 elements were extracted.

The learned representation is then used for activity recognition with a Linear SVM model. Using the SVM improves the obtained mAP compared to softmax output from the FC layer.

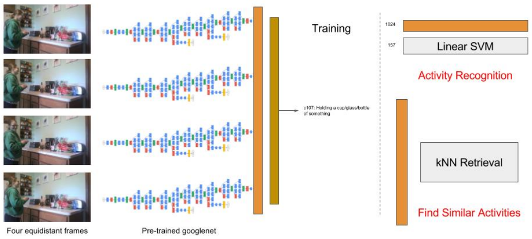


Fig. 6 Multi-Stream Late Fusion

The main benefit of learning representations for activity is that the features can be used for auxiliary tasks. In this project, a retrieval task was also evaluated using the learned features. kNN retrieval allows for extracting videos with similar activities compared to the probe video and is fairly easy to implement on the extracted features. Retrieval also helps evaluate quality of the learned features and their separation in the embedding space. Results obtained are presented in the next section.

Several deep learning model architectures have been evaluated on the Charades dataset primarily for benchmarking its performance against dense trajectory features. The observation was that none of these perform as well as IDT (Improved dense trajectory) and the mAP values obtained were generally very low. The results obtained for this method do not outperform IDT but provide an interesting approach to using deep learning for learning a good representation for video activity recognition that might yield better results in the future.

5. Results

Fig.7 shows qualitative results on the activity recognition task. The first two images on the left show video keyframes where the action class was correctly recognized. The images on the right show keyframes where the action was incorrectly recognized. Observing the results shows that temporal pooling of information using the fusion architecture is not effective in recognizing activity sequences. Most of the results are based on correctly identifying objects in the keyframes.

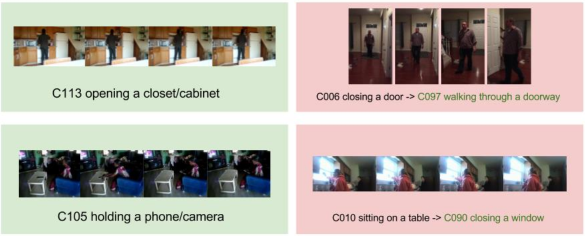

C113 opening a closet/cabinet
C006 closing a door -> C097 walking through a doorway
C105 holding a phone/camera
C010 sitting on a table -> C090 closing a window

Fig. 7 Qualitative Results - Recognition

Fig. 8 shows results on the retrieval task for the same video as before. The learned features have semantic content and are sensible as the retrieved videos do belong to similar action classes.

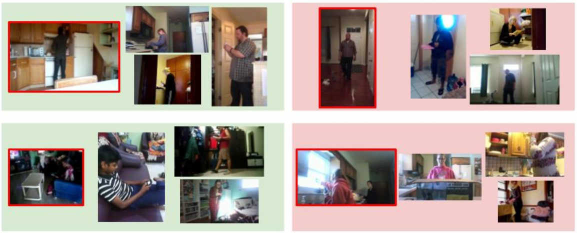
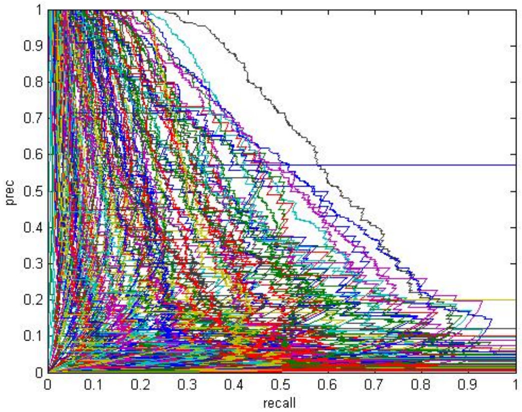


Fig. 8 Qualitative Results - Retrieval

A more quantitative measure of performance is depicted in Fig. 9 and Fig. 10. The precision-recall curves for all actions classes (157) are shown. Its clear that there is very high variance in the recognition performance across the classes. The right most curve depicts actions classes with tight object affordances. For ex- when a TV is detected in the video, it can only be the “watching TV” action class and hence a higher mAP is achieved. Left most curves depict the most ambiguous action classes.



`

Fig. 9 Precision-Recall Curve Classwise

Fig. 10 shows class-wise average precision for each action class. Again the observations was that the high precisions action classes had salient objects that could be easily recognized and tight affordance.

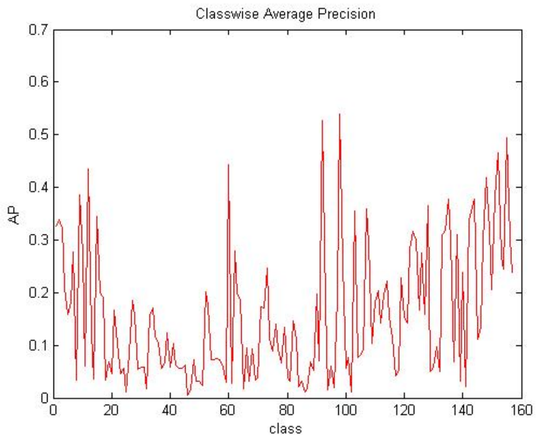


Fig. 10 Average Precision for each Class

The table below shows a comparison of results against other methods implemented on Charades. The current method performs significantly better than C3D and Two-Stream network configurations. The balanced Two-Stream (handles class imbalance in dataset) however achieves a mAP of 14.3% which is close to the achieved result. IDT still outperforms all deep learning methods as a feature to represent activities in video. mAP of 15.6% was achieved with the multi-stream late fusion method.

Method	C3D	Two-Stream	IDT	THIS
mAP (%)	10.9	11.9/14.3	17.2	15.6

6. Conclusion

The key conclusion of this work was that static CNN based approaches used for activity recognition from video that use some sort of temporal information pooling work well only with salient objects present in the scene and where the affordances on object-activity interactions we very tight. For objects where the likely number of action classes is large, the ambiguity is very high and cannot be resolved without a method that properly utilizes temporal evolution of the features.

Handling video data proved to be significantly difficult due the size of the dataset and limitations of memory and compute resources. Training a CNN-RNN with high resolution video data is still limited with even the state-of-the-art GPU hardware. Fusion based methods are very useful as they can reuse static image features of a CNN and extend the technique to video however they do not pool temporal information effectively even though they computational efficient.

As a future direction, this dataset will lead to many interest techniques for activity recognition. The dataset is well curated and large enough to train a deep learning model effectively. It will specially interesting to evaluate the performance of the recently developed Structural RNN method on this method as we are still far from having a method that can achieve results ready for in-field deployment.

References

[1] Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014.

[2] Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015.

[3] Sigurdsson, Gunnar A., et al. "Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding." *arXiv preprint arXiv:1604.01753*(2016).

[4] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." *Advances in Neural Information Processing Systems*. 2014.

[5] Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. "Unsupervised learning of video representations using lstms." *CoRR, abs/1502.04681* 2 (2015).