**Main Question**

“Can web action images be leveraged to train better CNN models and to reduce the burden of curating large amounts of training videos?”

**Motivation**

- Labeled web images tend to contain **discriminative action poses**, which highlight discriminative portions of a video’s temporal progression.
- n images contain more unique content compared to n video frames, and images are easier to collect.

Clearly, there exists a compromise between temporal information available in videos and discriminative poses and variety of unique content in images.

**Images beneficial irrespective of CNN depth?**

Three CNN models are used for action recognition on the dataset UCF101 split 1. All architectures benefit.

**Which classes benefit most?**

For UCF101 split 1, the top 25 classes benefiting from adding images are presented (absolute improvement).

**BU 101 Dataset**

We collect action images that correspond with the 101 action classes in the UCF101 video dataset. We manually filter for duplicate and irrelevant images eg. drawings or cartoons.

We test our approach by collecting a crawled unfiltered dataset for the larger scale dataset ActivityNet: ~800 hrs of video.

We obtain state-of-the-art performance when adding images and using motion features: Improved dense trajectories.

**State-of-the-art performance on UCF101**

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDT-FV (Wang et al. ICCVW'13)</td>
<td>85.9</td>
</tr>
<tr>
<td>Two-stream CNN (Simonyan et al. NIPS'14)</td>
<td>88.0</td>
</tr>
<tr>
<td>RCNN using LSTM (Ng et al. arXiv'15)</td>
<td>88.6</td>
</tr>
<tr>
<td>VGG16 + Images + IDT-FV</td>
<td>91.1</td>
</tr>
</tbody>
</table>

**Scalability: ActivityNet**

We test our approach by collecting a crawled unfiltered dataset for the larger scale dataset ActivityNet: ~800 hrs of video.

- State-of-the-art results on ActivityNet.
- Replacing 16.2M frames by 393K images obtains comparable accuracy.

**Conclusion**

We proposed a filtering technique for data of action videos, thereby reducing the amount of curated training videos needed.