Mining YouTube - Learning action classes from unlabeled videos

Hilde Kuehne, Alexander Richard, Juergen Gall
University of Bonn, Germany
{kuehne,richard,gall}@iai.uni-bonn.de

1. Introduction

Action recognition has become an important topic in academic and industrial research with most approaches so far relying on fully supervised training. Thus, for any action class to be trained, preclipped videos or a label with temporal information is needed. But the acquisition of such data is very time consuming as it needs some kind of human supervision, e.g. in form of Mechanical Turk workers. This puts a natural limit to the size of current benchmarks as well as to the transfer of current methods to real life application as collecting hand-annotated training data is often not feasible here. To address this issue, we propose a new benchmark focusing on the problem of learning actions from real-live videos without human supervision. To this end we follow the conventions of a real life setting and start with the target videos, a densely annotated set of 250 cooking videos with realistic action labels as they occur in the instructions given in the video or in textual recipes. The resulting annotations comprise 512 action classes with 10412 labeled action instances. As the annotation of a training set for this amount of classes would be very time consuming, we mine the respective training clips automatically. To this end, we collect a set of 191K samples from 21K videos by searching class related attributes on YouTube using the transcripts of the audio stream to generate weak labels from the training examples as e.g. shown in Figure 1. This approach follows and continues the idea of earlier approaches such as [5, 4, 1] and extends their ideas to large scale real live conditions. We benchmark the proposed dataset with respect to current features and architectures showing the challenging problems as well as the the benefits of this type of data.

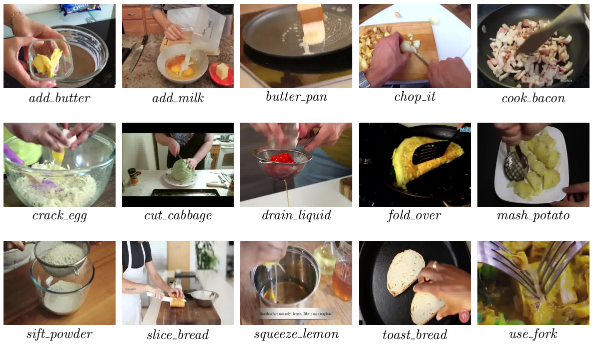

Figure 1. Examples of frames and related classes from YouTube cooking videos. Similar to [5], we follow the idea of automatically mining large scale training data from videos and subtitles without the need for human intervention.

2. Learning actions without annotation

The presented benchmark is based on the idea that it should be possible to learn action classes and concepts from unannotated videos. To do so, we follow [5, 4, 1] and make use of the spoken language within in the videos to extract possible class labels. Especially in context of instructional videos, people usually explain and comment their actions to the viewers, so the performed actions are named during execution. To gather the training data we run various searches related to the target classes. To receive a textual representation of the video content, we lend on the close-captioning function. It shows that from the list of possible search results 36K videos are available for download with respective subtitles, 5K of them with manually added and 30K videos with automatically generated captions. We parse the given subtitles for possible target classes by word appearances resulting in 191K action instances (33M frames) for training. As this set includes both hand edited and automatically generated subtitles, we assess how both sets contribute to the overall dataset. We analyze the properties of the edited as well as the automatic training data separately, see Table 1, also using human annotators to evaluate the hit rate for both on 2000 videos. It shows that the hitrate differs between the edited and the automatically generated subtitles by only 8.3% with 46.2% accuracy in case of edited subtitles and 37.9% for the automatically generated data. Note that this number only refers to the number of clips in which the labeled action was present, but that usually not all frames correspond to the labeled activity.

3. Evaluation

To extract features we use the Temporal Segment Network framework as proposed by [7] and the BNInception

	Comparison of training data	
	Edited	Automatic
Source Videos	5842	30557
Mined instances	$65k$	$125k$
Mined frames	$10M$	$23M$
Hitrate	46.2%	37.9%

Table 1. Evaluation of edited and the automatically generated subtitles. Hitrate describes the amount of videos which were considered as correct by a human annotator i.a. the labeled action was present in the video.

models pretrained on the Kinetics dataset provided on the website of the authors. As we do not have any reliable training data, we omit fine tuning and only extract features from the output of the last global pooling layer of the architecture. For evaluation we consider the task of temporal alignment of frames to a given ordered set of action classes. For this task, the transcript of a video as well as the video itself is available at test time and the goal is to temporally align the video frames accordingly. It has been introduced by [1] and has so far been used for most weak learning evaluations *e.g*. [2, 3, 6]. As performance measure, we use the Jaccard index computed as intersection over union (IoU) as well as intersection over detection (IoD) over all frames for each class and report the mean over all classes. To keep the computation feasible, we run the following experiments with a subset of 100k samples. To this end, we consider two shallow network architectures, a multi layer perceptron (MLP) and a network with one layer of gated recurrent units (GRUs). We vary the size both networks by using 1024, 2048, 4096 and 8192 units respectively. It shows the the MLP network with 2048 units outperforming all other configurations with an IoD of 14.02 and an IoU of 9.02. Additionally, all setups are able to perform consistently better than a random ($IoD = 9.11$, $IoU = 5.27$) or uniform alignment ($IoD = 9.12$, $IoU = 5.45$) on the test set.

3.1. Comparison to web-crawled video data

The first question raised by the proposed approach of mining samples from subtitles is the question if the same task could also be achieved by simply using web-crawled videos for the respective classes. To asses the difference of both methods, we use the videos retrieved by the class based search query and trained the best performing system, MLP with 2048 units, with the respective frame-based features. Instead of using only snippets, we use features from the whole video and train the network with the respective class labels. Looking at the results in Table 2 it becomes clear that for the here targeted, fine grained actions, a webly supervised search and training procedure is obviously not enough.

	Web-crawled	Subtitle
Jacc. IoD	4.03	14.02
Jacc. IoU	2.05	9.02

Table 2. Results for the training with web-crawled videos compared to samples based on subtitle mining. It shows that simple web crawling is not suitable for the targeted fine grained actions

3.2. Extension to large data sizes

As one of the main advantages of automatic mining is the fact that it scales, we extend the experiments from the so far used validation set by doubling the amount of training data and using the full size dataset. To compensate for the imbalanced distribution, we only use up to 500 clips for each action. Overall we use 190k video clips. It shows that doubling the amount of training samples leads to an noticeable increase for all configurations, with the best performing one reaching IoD rate of 16.86 and aun IoU rate of 10.37.

4. Conclusion

It shows that learning from automatically labeled data is not only possible, it also helps to gather more training samples, even if a large amount of noise is included, to increase the overall performance on the test data. We hope that this benchmark will allow to make the whole process of automatically gathering training samples and evaluating respective systems reproducible by providing all data necessary and, thus, lead to new ideas and insights in this field.

References

[1] P. Bojanowski, R. Lajugie, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Weakly supervised action labeling in videos under ordering constraints. In *European Conf. on Computer Vision*, pages 628–643, 2014. 1, 2

[2] D.-A. Huang, L. Fei-Fei, and J. C. Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *European Conf. on Computer Vision*, pages 137–153, 2016. 2

[3] H. Kuehne, A. Richard, and J. Gall. Weakly supervised learning of actions from transcripts. *arXiv preprint arXiv:1610.02237*, 2016. 2

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2008. 1

[5] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabinovich, and K. Murphy. What's cookin'? Interpreting cooking videos using text, speech and vision. In *Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015. 1

[6] A. Richard and J. Gall. Temporal action detection using a statistical language model. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016. 2

[7] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Val Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1