Action Recognition Based on Divide-and-conquer

湖南大學 HUNAN UNIVERSITY
信息科學與工程學院

Guanghua Tan, PhD, Rui Miao, MD and Yi Xiao, PhD

INTRODUCTION

Abstract

Deep convolutional neural networks have made great breakthroughs in the field of action recognition. Since sequential video frames have a lot of redundant information, compared with dense sampling, sparse sampling network can also achieve good results. Due to sparse sampling's limitation of access to information, this paper mainly discusses how to further improve the learning ability of the model based on sparse sampling. We proposed a model based on divide-and-conquer, which use a threshold α to determine whether action data require sparse sampling or dense local sampling for learning. Finally, our approach obtains the state-the-of-art performance on the datasets of HMDB51 (72.4%) and UCF101 (95.3%)

PROPOSED METHOD

Framework

The model based on divide-and-conquer contains two modules, including sparse sampling module and dense sampling module, as shown in figure 1. Threshold α determines which sampling strategy is appropriate for different action data. The final output consists of two parts: the first part is the output of the sparse sampling, and the second part is the fusion of the sparse sampling and the dense sampling..

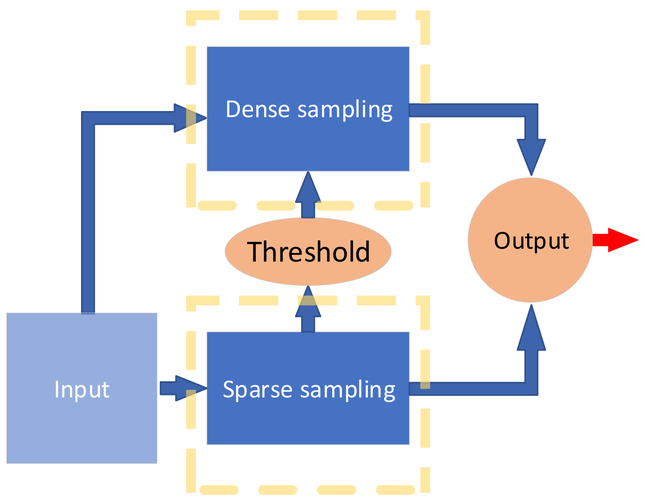

Fig. 1. The model based on divide-and-conquer.

Sparse Sampling

Sparse sampling extracts short snippets over a long video sequence with a sparse sampling scheme, where the samples distribute uniformly along the temporal dimension, composed of spatial stream ConvNets and temporal stream ConvNets, as shown in figure 2.

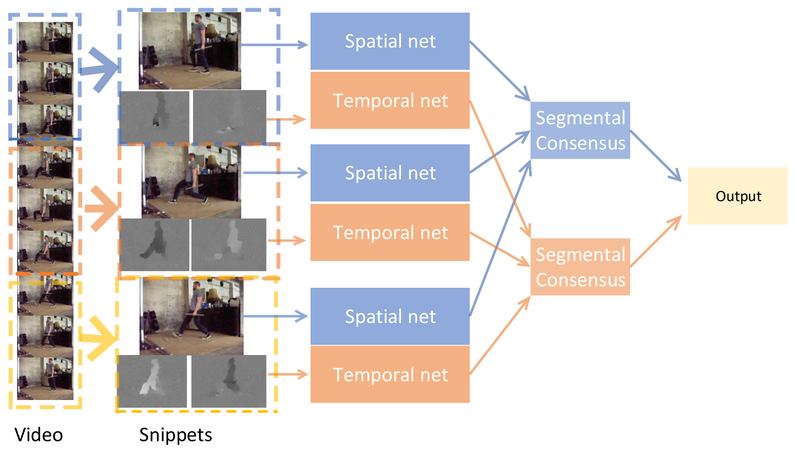

Fig. 2. Sparse sampling network.

Threshold Setting

$$f = (Pmax - Psec) / Pmax$$

where Pmax is the maximum classification probability of the prediction result, and Psec is the second largest classification probability. When f is less than α, we think that sparse sampling cannot distinguish this action very well, so the prediction result must be merged with the results of the external algorithm. Table 1 includes all the actions whose f is less than α in UCF-101.

Table 1. Actions whose prediction of *f* is less than threshold *α* in sparse sampling

Action	Accuracy	Correct classification
BreastStroke	76.4%	FrontCrawl (23.5%)
BrushingTeeth	71.4%	Hammering (14.2%) ShavingBeard (14.2%)
CricketBowling	71.4%	CricketShot (28.5%)
FieldHockeyPenalty	70.1%	CricketShot (8.7%) Shotput (14%)
ShavingBeard	69.8%	BrushingTeeth (22.6%)
SkateBoarding	80%	Skiing (20%)
ThrowDiscus	77.5%	HammerThrow(12.2%) Shotput(8.1%)

Dense Sampling

Dense sampling algorithm can give sparse sampling more detailed information when f is less than threshold α. For dense sampling, we use a single-frame-based algorithm. Dense sampling is shown in figure 3.

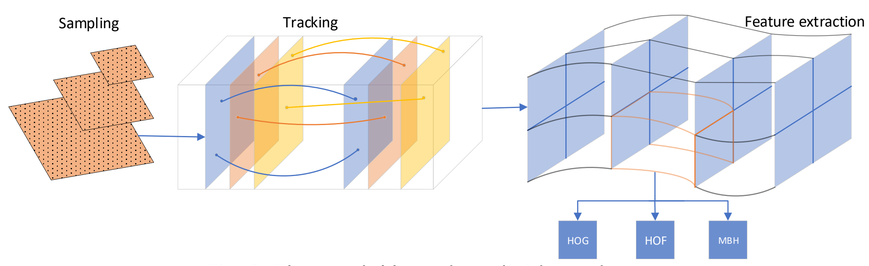

Fig. 3. The model based on divide-and-conquer.

If the dense sampling algorithm is trained with the complete UCF-101 data which contains 101 different kinds of actions, the result is shown in table 2. It can be seen that the predictive accuracy of those chosen actions is not as good as the sparse sam-pling network.

Table 2. For all actions whose *f* are less than *α* in sparse sampling network, dense sampling's performance

Action	Sparse sampling	Dense sampling
BreastStroke	59.2%	76.4%
BrushingTeeth	66.7%	71.4%
CricketBowling	61.6%	71.4%
FieldHockeyPenalty	68.5%	70.1%
ShavingBeard	67.2%	69.8%
SkateBoarding	65.4%	80%
ThrowDiscus	70.7%	77.5%

Performance Analysis

Sometimes the predictive accuracy of one action is improved, but its similar action's predictive accuracy is reduced adversely. When evaluating the predictive effect, similar actions are a whole. So we introduce prediction on similar action sets (include all actions similar to each other) to evaluate one model's ability. Table 3 shows that similar action sets' predictive accuracy of sparse sampling network is generally low compared to the average prediction rate (94%) for all data.

Table 3. Sparse sampling network's prediction accuracy of different similar action sets.

Similar action sets	Accuracy
BreastStroke+FrontCrawl	59.2%
BrushingTeeth+ShavingBeard+Hammering	66.7%
CricketBowling+CricketShot	61.6%
FieldHockeyPenalty+Shotput	68.5%
ThrowDiscus+HammerThrow	67.2%

Since sparse sampling network filters out obfuscated actions, it is not necessary for the dense sampling algorithm to recognize all actions. For example, if sparse sampling network thinks that a video is BreastStroke, the dense sampling algorithm only needs to further determine whether the video is belong to BreastStroke or FrontCrawl because FrontCrawl is only similar to BreastStroke. In this case, the dense sampling algorithm's predictive accuracy is greatly improved, as shown in table 4. Meanwhile, we found that the HOF feature is useless when distinguishing similar actions, so the dense sampling algorithm mainly extracts the HOG feature and the MBH feature of videos.

Table 4. Accuracy comparison for our model and sparse sampling and dense sampling.

Similar action sets	Our model	Sparse sampling	Dense sampling
BreastStroke+FrontCrawl	84.3%	83.1%	67.7%
BrushingTeeth+ShavingBeard+Hammering	78.2%	73.1%	58.9%
CricketBowling+CricketShot	83.3%	64.7%	66.7%
FieldHockeyPenalty+Shotput	89.4%	82.3%	51.5%
ThrowDiscus+HammerThrow	91.6%	92.7%	89.1%

Table 5 shows the final results of our model in UCF-101 and HMDB51, and we compared it with other methods.

Table 5. Accuracy comparison for our model and sparse sampling and dense sampling.

Action	UCF-101	HMDB51
Our model	95.3%	72.4%
TSN	94.2%	69.4%
IDT	85.9%	57.2%
GRP+IDT	92.3%	67.0%
LSTM	93.6%	66.2%
ST-VLMPF	93.6%	69.5%

REFERENCES

[1] . Wang H, Schmid C.: Action Recognition with Improved Trajectories. In: IEEE International Conference on Computer Vision. IEEE, 2014:3551-3558.
[2] Simonyan K, Zisserman A.: Two-Stream Convolutional Networks for Action Recognition in Videos. Computational Linguistics, 1(4):568-576 (2014).
[3] Wang L, Xiong Y, Wang Z, et al. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In: Computer Vision and Pattern Recognition, 2016.IEEE Transactions on Pattern Analysis & Machine Intelligence, PP(99):2999-3007, (2017)
[4] Sun L, Jia K, Yeung D Y, et al.: Human Action Recognition Using Factorized Spatio Temporal Convolutional Networks. In: International Conference on Computer Vision, 2015: 4597-4605.