Action Recognition and Anticipation from egocentric video

Name: Hao Mengkang
CID: 01905746
Supervisor: Jianing Qiu and Dr Benny Lo

HAMLYN SYMPOSIUM
ON MEDICAL ROBOTICS
2021

Abstract

The video action anticipation have become popular in many domains in recent years, because of its wide applications in AR/VR, robot imitation learning, and autonomous driving. This project explores the Siamese network [1] based on the method of Temporal Shift Module [2] (TSM) to recognize the video action and anticipate the next future action given an egocentric video. In this work, we mainly focus on video action recognition. A subset of the popular EPIC-Kitchen dataset [3] is used to evaluate our method. Object masks are fused with RGB frames to enhance the action recognition accuracy, which leads to a 6.25% increase in terms of the top1 test accuracy compared with using RGB inputs along.

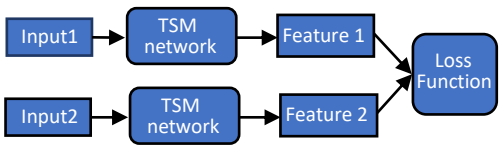
Introduction

- **Task:** Video action recognition and anticipation
- **Aim**: Improving action recognition and anticipation accuracy using multi-modality fusion
- **Dataset:** 40 kinds of actions that have 6 to 10 video segments
  320 total video segments from epic-kitchen 100
  Train set: each action has 2-6 video segments
  160 video segments totally
  Validation set: each action has 2 video segments
  80 video segments totally
  Test set: each action has 2 video segments
  80 video segments totally

Methods

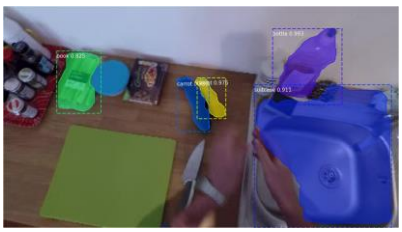
- **Siamese structure**

To decrease the distance between the inputs with same classes and to increase the distance between inputs with different classes.



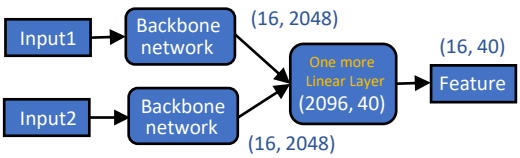
- **Mask and Bounding-box features**

The masks are obtained from Mask RCNN network [4].

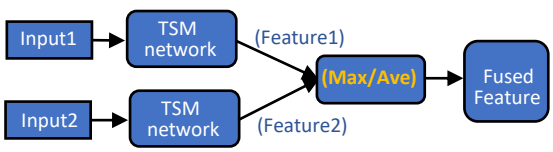


- **Fusion Method**

1. Late-fusion (Modify the TSM network In Linear Layer)



2. Feature-fusion (fuse two features by max/average pooling)



Preliminary Results

Figure 1 are the visualizations of the predicted label when we input RGB and Mask separately.
In order to analyze if the mask modality will improve the result of RGB-frame only, The detailed results of the RGB only, Mask only and three kinds of Fusions has been listed in Table 1.
From the results of Table 1, it can be seen that the Average-late fusion has the best Top1 and Top5 Test accuracy than other fusion methods, and it has a 6.25% and 2.5% improvement in Top1 and Top5 Test accuracy than the results of RGB only.

(a) (b)

Fig 1. (a) The visualization of predicted label using RGB frame alone; (b) The visualization of predicted label using Mask alone

Table 1 – The results with the different modalities and fusion methods

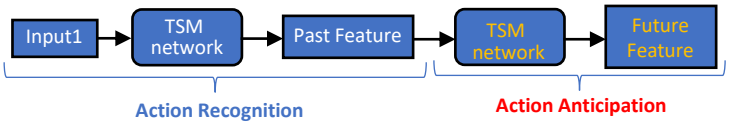
	Top1-Val Acc	Top5-Val Acc	Top1-Test Acc	Top5-Test Acc
RGB only	26.25%	50%	32.5%	52.5%
Mask only	13.75%	42.5%	16.25%	43.75%
Late-Fusion	22.5%	47.5%	30%	55%
Maximum Feature-Fusion	-	-	32.5%	52.5%
Average Feature Fusion	-	-	**38.75%**	**55%**

Future Plan

Up to now, the action recognition of egocentric video has been finished with both RGB and Mask modalities. However, the accuracy could still be improved by potentially fusing more modalities:

- **Audios**
- **Flow-frames**

The next stage of the project will be focused on the action anticipation which is to anticipate the next future action after observing a fixed segment of an egocentric video.



References

[1] Melekhov I, Kannala J, Rahtu E. Siamese network features for image matching[C]//2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016: 378-383..
[2] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7083-7093.
[3] E. K. W. P. J. M. H. D. A. F. G. M. F. Dima Damen, "EPIC-KITCHENS-55," 2020. [Online]. Available: https://epic-kitchens.github.io/2020-55.html#results. [Accessed 10 6 2021].
[4] K. He, G. Gkioxari, P. Dollar and R. Girshick, " Mask R-CNN,"*Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961-2969.