# NTU hTC Deep Learning Competition - Hand Detection

D05943006 葉陽明，R05943080 莊政彥，R05943171 劉承曄，R06942075 許宗嫄

## I. Introduction

近年來，由於穿戴式裝置產品的大幅發展及普及化，造就了大量的資料。其中，具有相機功能的穿戴式裝置產生的資料型態多為第一人稱視角(Egocentric)的資料，影像與使用者真實所見差異極小，這種獨特的資料型態也衍生出了許多新的需求及應用，例如第一人稱視角的臉部追蹤與辨識、關鍵影格及物件粹取(egocentric video summarization) [1]、使用者手部偵測 [2]...等等。其中，由於知曉使用者手部位置後，可以進一步瞭解使用者與物件及場景的互動，也可以藉此進行手勢等辨識，因此使用者手部辨識可以說是各種進階應用中的基礎。

然而，第一人稱視角手部影像資料目前尚未有一個完整、龐大且分類詳細的資料庫，為了因應未來的需求，如何利用少量有限的真實訓練資料做到穩定精準的使用者手部辨識成為了一大議題。

本計畫旨在使用3D模組合成之第一人稱視角影像資料作為訓練資料，輔以極少量的真實訓練資料，找出能穩定且精確找出使用者手部位置的演算法，並輸出其對應的bounding box。資料庫方面，我們使用hTC提供的第一人稱影像進行訓練，其中包含了有合成影像及真實影像的訓練資料，以及完全由真實影像構成的測試資料，資料的分布分別為100,000張有標記之合成影像，及464張有標記與6197張無標記之真實影像，標記包含部分手指骨架座標及手部bounding box等等。

計畫中，我們採用Faster R-CNN [3]在合成資料上先進行簡單的訓練後，再用少部分的真實資料進行transfer learning，在測試資料的mAP上達到了0.85以上的效果。

## II. Methodology

Egocentric hand detection可以被視為總共兩個類別(左手與右手) 的object detection問題，因此我們使用Faster Region-Based Convolution Networks (Faster R-CNN) [3]作為模型架構，在整合的架構之中產生object proposal，提出目標物件可能的位置區域，並且對這些proposal進行classification及refinement，偵測其類別並精修bounding box的座標及長寬。

Faster R-CNN依功能可以細分成兩個模組，Region Proposal Network (RPN) 及Fast R-CNN [4]，RPN負責進行object proposal，Fast R-CNN則對RPN提出的proposal進行object detection，兩個模組在訓練時以end-to-end方式進行最佳化，以下將分述兩個模組的細部架構。

### A. Fast R-CNN

Fast R-CNN [4]的架構如Fig. 1. 所示，input為一張影像及其對應的一組object proposals，影像經過數層convolution layers與maxpooling layers產生一張conv feature map，然後一層region of interest (RoI) pooling layer依據各個object proposal的區域位置，在conv feature map上抽取固定長度的feature vector，經過數層fully-connected (fc) layer，產生兩個output layer分支，一支做classification產生softmax probability估測該object proposal屬於各類別的機率，一支做regression估測bounding box的參數包含頂點座標及長寬。整個模組的loss包括classification loss及regression loss，其中classification loss為negative log loss，regression loss為robust L1 loss。

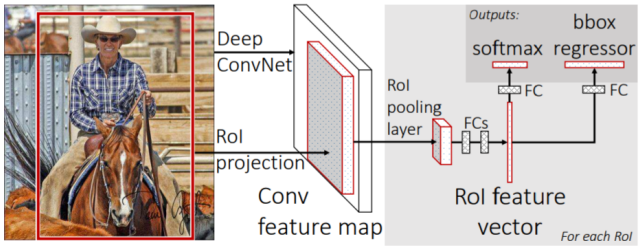


Fig. 1. Fast R-CNN模型架構

### B. Region Proposal Network

Region Proposal Network (RPN)的input為single scaled影像，output則是數個矩形的object proposal以及其分數。在Faster R-CNN中，RPN的架構直接與Fast R-CNN整合，如Fig. 3.所示，兩者共用前面數層convolution layers以減少參數與計算量，然後RPN在最後一層共用的conv feature map上用3x3的window抽取低維度的feature vector (256-d或512-d)，這個feature vector分別進入兩個fully-connected layers，一個做classification，一個做regression。RPN的特點在於這個部分，使用不同尺度與比例的anchors作為參考，如Fig. 2.所示，使得最後預測出來的bounding box大小具有彈性。

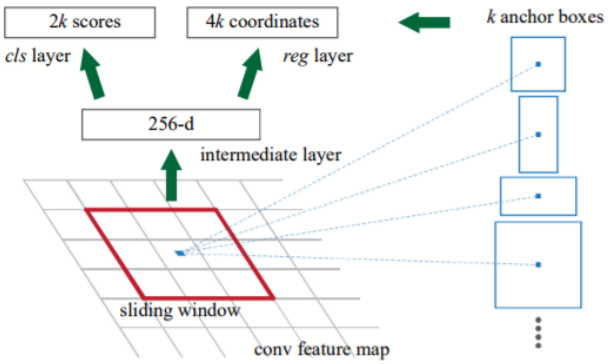


Fig. 2. Region Proposal Network (RPN)

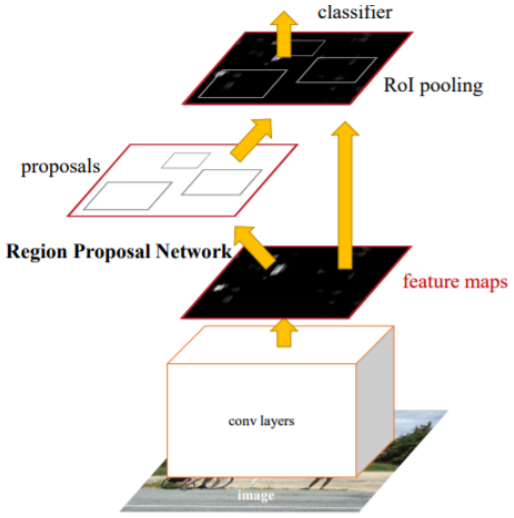


Fig. 3. Faster R-CNN模型架構

### C. Faster R-CNN

Faster R-CNN的訓練方法為，先以pre-trained model，如在ImageNet上pre-trained過的ResNet來初始化RPN裡convolutional layers的參數，然後以RPN產生的object proposal來訓練Fast R-CNN，再將Fast R-CNN內convolutional layers的參數分享給RPN，固定shared layers的參數，最後再回頭訓練Fast R-CNN

## III. Implementation

### A. Model Overview

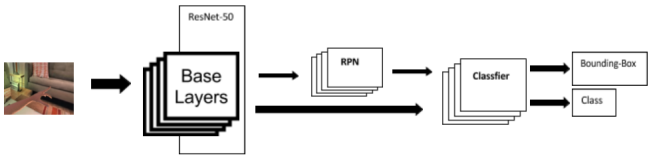


Fig. 4. 我們的Model

### B. Shared Based Layers

仿照Faster-RCNN論文架構，我們使用ResNet-50來當作Base-Layers，讓RPN跟Classifier共享。

### C. Region Proposal Network (RPN)

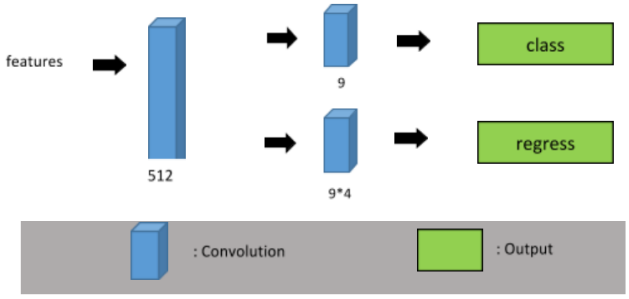


Fig. 5. 我們使用的RPN

### D. Classifier

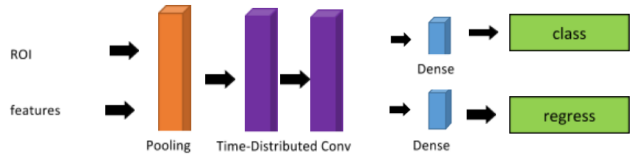


Fig. 6. 我們使用的Fast R-CNN

### E. Training Procedure

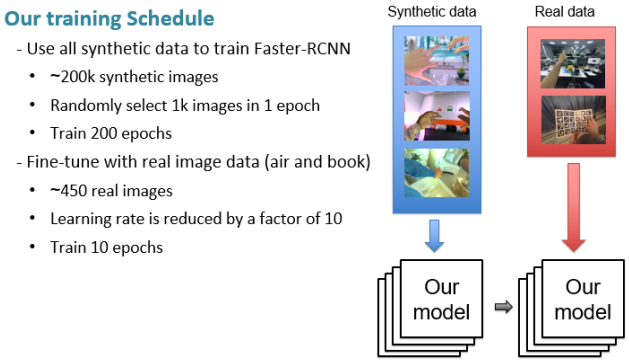


Fig. 7. 我們的訓練流程

## IV. Experiment

我們一共做了兩個實驗，一個是不同epochs數量對結果的影響，另一個是利用Cycle-GAN [5]對原synthetic data轉成real data再做training

### A. Epochs數量對結果的影響

我們測試了Training on Synthetic Data：epochs = 80, 200的組合，測試結果如Fig. 8(a).與Fig. 8(b).可以看到200 epochs的結果較80epochs時，抓到手的機率增加，但似乎有時會在錯誤的區域誤認物件為手。

另外測試了Fine-Tuning on Real Data: epochs = 1,10的組合，測試結果如Fig. 8(c).與Fig. 8(d).可以看到經過fine-tune後，抓到手的機率大增，也不再出現錯誤偵測區域，而當fine-tune的epochs增大時，可以看見，model對於物件內的類別的信心度也增加了。

### B. Cycle-GAN Style Transfer

我們以unlabeled synthetic data與real data訓練了一個Cycle-GAN [5]模型，用這個Cycle-GAN將10000張labeled synthetic data轉到real data domain然後加入training data中，訓練Faster R-CNN，實驗結果顯示這樣做對Faster R-CNN在訓練第一階段有所幫助，但是在第二階段(fine-tune)時，提升則相對有限，整體表現並沒有顯著提升。

## V. Conclusion

使用Faster R-CNN架構，我們最後以訓練200 epochs在synthetic data上，然後縮小learning rate，fine-tune


(a) (b) (c) (d)

Fig. 8. (a) 80 epochs on synthetic data. (b) 200 epochs on synthetic data. (c) 1 epoch fine-tune on real data. (d) 10 epochs fine-tune on real data.

| train with synthetic data | | train with cycle GAN data | |
|---|---|---|---|
| epoch | score | epoch | score |
| 80 | 0.2083 | 200 | 0.4167 |
| 200 | 0.3333 | | |
| 450 | 0.1944 | | |
| **fine-tune with real data** | | **fine-tune with real data** | |
| epoch | score | epoch | score |
| 3 | 0.8489 | 5 | 0.8367 |
| 5 | 0.8501 | | |
| 10 | 0.8488 | | |
| 15 | 0.8428 | | |
| 20 | 0.8379 | | |
| 30 | 0.8322 | | |

TABLE I. NUMBER OF FINE-TUNE EPOCH V.S. JUDGER SCORE(ON-LINE)

5 epochs在real data上得到我們的best model。在實驗過程中我們發現fine-tune對模型表現的影響非常大，顯示synthetic data與real data確實存在不小domain mismatch，以最簡單的transfer learning方式即可大幅提升模型表現。反倒是Cycle-GAN的方法，雖然使用了較複雜的模型架構，但是還不足以將synthetic data轉到real domain，只學到影像的整體風格，如明暗與色調等，對於偵測bounding box沒有顯著的幫助。

另外，我們實作出來的模型大小約100MB左右，與其他組別的方法相較之下，雖然能夠得到不錯的judger score，但是這樣的模型大小，距離穿戴式裝置的需求還有一大段路要走。

## ACKNOWLEDGMENT

經過這次的Project,我們學習到最新的bounding box detection方法。才疏學淺的我們以前只知道要把整張圖分類成一種類別，並不知道還有這種針對多種類別的所在區域所做的研究，透過這類研究的進展，我認為機器對圖像的理解更接近一般人類，藉由針對影像所在區域做的分類，未來許多應用或許可以利用此技術來對特定區域或特定類別進行處理，想必未來會有許多更厲害的研究，挑戰機器學習的極限。

## REFERENCES

[1] Lee, Yong Jae and Ghosh, Joydeep and Grauman, Kristen, *Discovering Important People and Objects for Egocentric Video Summarization*, CVPR, 2012.

[2] Li, Cheng and Kitani, Kris M, *Model Recommendation with Virtual Probes for Egocentric Hand Detection*, CVPR, 2013.

[3] Ren, Shaoqing and He, Kaiming and Girshick, Ross and Sun, Jian, *Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks*, NIPS 2015.

[4] Girshick, Ross, *Fast R-CNN*, ICCV 2015.

[5] Zhu, Jun-Yan and Park, Taesung and Isola, Phillip and Efros, Alexei A, *Unpaired image-to-image translation using cycle-consistent adversarial networks*, ICCV 2017.