國立臺灣大學電機資訊學院資訊工程學系
碩士論文
Department of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National Taiwan University
Master Thesis

# 影像中使用者感興趣區域偵測之資料集
# A Benchmark for Region-of-Interest Detection in Images

莊宜蓁
Yi-Chen Chuang

指導教授：莊永裕博士
Advisor: Yung-Yu Chuang, Ph.D.

中華民國 109 年 6 月
June, 2020

# 國立臺灣大學碩士學位論文
# 口試委員會審定書

## 影像中使用者感興趣區域偵測之資料集
## A Benchmark for Region-of-Interest Detection in Images

本論文係莊宜蓁君 (R07922066) 在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 109 年 6 月 28 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

_____

_____ _____

_____ _____

_____ _____

_____ _____

所　長：_____

# 誌謝

感謝...

# Acknowledgements

I’m glad to thank…

# 摘要

本論文提出了一影像中使用者感興趣區域 (region of interest) 偵測之資料集 (benchmark)。使用者感興趣區域偵測在許多應用中極為有用，過去雖然有許多使用者感興趣區域之自動偵測演算法被提出，然而由於缺乏公開資料集，這些方法往往只測試了各自的小量資料而難以互相比較。從其它領域可以發現，基於公開資料集的可重製實驗與該領域突飛猛進密切相關，因此本論文填補了此領域之不足，我們提出名為「Photoshoot」的遊戲來蒐集人們對於感興趣區域的標記，並以這些標記來建立資料集。透過這個遊戲，我們已蒐集大量使用者對於感興趣區域的標記，並結合這些資料成為使用者感興趣區域模型。我們利用這些模型來量化評估五個使用者感興趣區域偵測演算法，此資料集也可更進一步作為基於學習理論演算法的測試資料，因此使基於學習理論的偵測演算法成為可能。

**關鍵字：** 關鍵字

# Abstract

This thesis presents a benchmark for region of interest (ROI) detection. ROI detection has many useful applications and many algorithms have been proposed to automatically detect ROIs. Unfortunately, due to the lack of benchmarks, these methods were often tested on small data sets that are not available to others, making fair comparisons of these methods difficult. Examples from many fields have shown that repeatable experiments using published benchmarks are crucial to the fast advancement of the fields. To fill the gap, this thesis presents our design for a collaborative game, called Photoshoot, to collect human ROI annotations for constructing an ROI benchmark. With this game, we have gathered a large number of annotations and fused them into aggregated ROI models. We use these models to evaluate five ROI detection algorithms quantitatively. Furthermore, by using the benchmark as training data, learning-based ROI detection algorithms become viable.

**Keywords:** keyword

# Contents

# List of Figures

# List of Tables

# Chapter 1

# 緒論

人臉偵測一直以來都是深度學習和電腦視覺領域中一個重要的議題。經過了幾年的研究，人臉偵測也漸趨成熟，而能夠被應用在更多領域中。其中一個新興的領域便是智慧型汽車這項技術。近年來人們不斷研究如何讓汽車駕駛變得更加方便和安全，因而研發出各項技術，如智慧駕駛輔助、自動駕駛等等。隨著這些技術的成熟，車內影像的應用也漸趨重要，其中一項應用便是對車內人臉的偵測。

但經過測試，一般人臉偵測的模型做在車內人臉上的表現並不是很好。車內人臉偵測的結果對後續的相關應用而言十分重要，因此改善它的結果是相當值得研究的。經過觀察，我們發現車內人臉偵測和一般人臉偵測相比，更容易碰到在不同光線照射下的例子；然而一般人臉偵測的研究較重視在大部分情境下的表現結果而忽略了在較極端光照下的例子，因此其在車內人臉的偵測上便表現得不如預期。

在本研究中，我們試著消除不同光線照射對輸入資料造成的影響，使其在人臉偵測上能獲得較好的表現。我們在輸入資料過人臉偵測器之前先對資料進行正規化處理，並期待能透過消除環境光對資料的影響來改善人臉偵測的結果。

接下來的章節會陸續提到以下內容：在第二章，我們會介紹和本研究相關的其他研究，包含了人臉偵測、低光照下的人臉偵測和影像增強。在第三章，我們會詳細介紹本研究所使用的方法。在第四章，我們會說明訓練和測試的設定，展示視覺和數據上的結果，並以實驗說明方法中各個細節對結果造成的影響。而在第五章，我們會做出本研究的結論。

# Chapter 2

# 相關研究

## 2.1 人臉偵測

[1] 人臉偵測在電腦視覺是個發展成熟的議題，其能被應用在諸多領域上，因此有很大的重要性。從以前到現在，人們使用各種不同的方式來解決這個議題。Viola-Jones 人臉偵測器結合了積分圖 (Integral Image)、哈爾特徵 (Haar Feature)、自適應增強 (AdaBoost) 學習、將數個弱分類器級聯 (Cascade) 等概念，率先做到了實時性 (Real-Time) 高精度人臉偵測。在卷積神經網路興起後，大家對人臉偵測的研究又更加熱烈。FaceBoxes 藉由在卷積層使用較大的步伐 (Stride) 快速將輸入縮小，在盡量不影響結果的情況下減少輸出所需的頻道數，並搭配 Faster R-CNN 中核心的 RPN (Region Proposal Network) 網路和錨點 (Anchor) 的機制來做到高精度實時性的人臉偵測；DSFD 強化圖片中被擷取的特徵並利用這兩組特徵來算出比單一一組特徵更準確的臉部位置；BlazeFace 則使用了輕量化的網路和需要的運算處理較低的架構等，使人臉偵測能夠進一步被用於行動裝置的相關應用上。

## 2.2 低光照下的人臉偵測

人臉偵測並非只被應用於正常光照下的情境，有時候我們也會需要處理夜間等較有挑戰性的情境，而大家對於這樣的情境也各有不同的解決方案。有些人 (Se Woon Cho) 使用低光照下的圖片作為訓練資料，試圖使訓練時的情境接近測試時的情境以提升偵測的準確率；也有人 (Mrinal Kanti Bhowmik 和 Jinwoo Kang) 使用

了紅外光攝影機獲取非可見光照射下的資料，避開了低光照對圖片造成的影響；其他人則選擇了對圖片進行增強，試圖還原在低光照下損失的色彩與細節。

## 2.3 影像增強

影像增強也是一項非常熱門的應用，其大量被用於加強低光照下的圖片使損失的色彩與細節得以被還原。較為傳統並廣為人知的作法包含了直方圖均衡化、珈瑪校正和視網膜增強算法等，前兩者主要透過拉高像素間的對比來達到強調細節的效果，而第三者透過模擬人眼視網膜以計算出圖片增強後的結果。其餘方法多運用深度學習訓練出增強圖片的模型。MSR-net 受啟發於帶色彩恢復的多尺度視網膜增強算法並將其演算法轉換成卷積神經網路；Jianrui Cai 將圖片分為低頻和高頻並分別進行增強，融合兩個結果後再進行第二次增強得到最後結果；Chen Wei 把圖片分為光照和反射兩部分，把光照部分增強後再和反射部分相乘得到最終結果。

# Chapter 3

# 方法

我們的方法為了能使在不同光線照射下的輸入資料獲得較好的偵測結果，會在偵測前先對圖片進行正規化，圖一說明了測試時從輸入資料到獲得偵測結果的流程。為了能更好的消除光線對輸入資料造成的影響，我們在訓練上使用了特別設計的訓練流程，主要分為預訓練和主要訓練。

## 3.1 預訓練

我們做預訓練的對象是負責對圖片進行正規化的正規器，這一步的目的是給正規器一個好的初始權重，以便在後續的主要訓練進行優化。

我們設計了一套特別的訓練架構—三圖一組架構—來訓練正規器 (如圖二)。以下會分別就這套訓練架構中的損失函數、資料和所使用的模型進行說明：

### 3.1.1 損失函數

我們希望設計出的損失函數能使來自不同光照下的圖片被正規化成相似的結果，以便進行後續的人臉偵測。為了達成這個目標，我們想對同一場景在三個不同光照條件下的圖片進行比較。這個想法是受到了 FaceNet 的啟發。它是一篇做人臉辨識的論文，而他們設計的損失函數會使錨點盡量靠近同身份的肯定結果，並使其盡量遠離不同身分的否定結果 (如圖三)，他們透過比較三個點之間距離來修改

參數權重的架構啟發了我們。我們的損失函數定義如下：

$$L_{Total} = \alpha L_{Content} + L_{Light}$$

其中 $L_{Light}$ 意在使不同光照下的輸入在經過處理後能有盡量一致的長相；而為了使學出來的模型能兼顧保留原圖樣貌和消除光線影響，我們加入了 $L_{Content}$ 以保留原圖中的資訊。$\alpha$ 在此作為一個調整兩個損失函數間取捨的超參數。

我們使用三張一組的圖片來實作上述損失函數的細節，每一組圖片包含了同一場景在曝光正常 (圖 4-1)、曝光不足 (圖 4-2)、曝光過度 (圖 4-3) 這三種情況下的圖片 O、D、B。我們將 D 和 B 這兩張圖片經過正規器後輸出的兩張圖 RD 和 RB 算出兩張圖間的損失函數 $L_3$，此處為了使兩張圖之間的距離越小越好而使用了 L2 損失函數。在此同時，為了不使這兩個結果和原圖相差過大，我們將 RD 和 RB 分別和 O 算出兩個損失函數 $L_1$ 和 $L_2$。這三個損失函數和我們在前面定義的損失函數關係如下：

(Lcontent 和 Llight 怎麼等於這些損失函數)

### 3.1.2 資料

由於利用真實資料會有收集上的困難，也較難保證場景中的內容物不變，在此架構中所使用三張一組的資料是透過演算法將原始圖片做調整，模擬出曝光不足和曝光過度這兩個情境下的圖片。以下分別說明兩種情境下的調整：

**模擬曝光不足**

(對圖片的亮度作反珈瑪校正後調暗圖片，再對圖片的亮度作珈瑪校正的公式) (對圖片的每個畫素加上標準差為 N，平均為 0 的高斯雜訊的公式)

**模擬曝光過度**

(對圖片的亮度作反珈瑪校正後調暗圖片，再對圖片的亮度作珈瑪校正的公式)

### 3.1.3 所使用的模型

在此架構中我們使用了 MSR-net 來作為正規器的架構。它受到了視網膜增強算法的啟發，是一個能夠對低亮度圖片做影像增強的模型，也就是說它能夠對圖片作局部亮度的調整；它在架構上的設計基於卷積神經網路，能夠接受不同大小的輸入資料；它和其他作影像增強的模型相比較為輕量，方便我們在進行預訓練後將其和偵測器接在一起作端對端的訓練優化。

## 3.2 主要訓練

在經過預訓練後，我們獲得了一個能夠消除光線對圖片影響的正規器。接下來的主要訓練會將正規器和人臉偵測器接在一起做端對端訓練 (如圖五)。

在主要訓練中，我們的目標是讓來自不同光線照射下的圖片在經過正規化處理後能夠有更多人臉成功被偵測出來。為了模擬不同光線照射下的情境，我們使用了上一節提到的演算法對訓練用的資料做了不同光線照射下的模擬，讓輸入資料包含曝光不足和曝光過度情境下的圖片。而在模型的挑選上，我們使用了 FaceBoxes 作為主要訓練中人臉偵測器的架構。它是一個基於卷積神經網路的人臉偵測器，透過對卷積層作優化來達到高效率高精確率人臉偵測。由於長遠來說我們希望能夠在車內進行高效率的人臉偵測，因此它是個很好的選擇。

# Chapter 4

# 實驗設定與結果

在這一章中，我會先介紹我們在實驗中所使用到的資料集，然後說明我們在訓練和測試時使用的設定與參數，接著展示我們做測試在數據與視覺上的結果，最後展示對我們的方法進行消熔實驗的結果和討論。

## 4.1 資料集

我們在實驗中所使用到的資料集主要分為訓練用和測試用的資料集。

訓練時我們使用了 WIDER FACE 作為主要的資料集，它是一個人臉資料集，其收集了來自各種不同情境下的人臉 (如圖六)，共計 32,203 張圖片、393,703 張人臉，在人臉偵測這個議題上是很經典的資料集。在我們的實驗中，我們用 WIDER FACE 分別做了曝光不足和曝光過度的模擬。其中曝光不足的模擬將圖片亮度隨機調為 0.03%、0.05%、0.07%，並將每張圖隨機分配 1 10 的數字 $N$，以 $\mu = 0$、$\sigma = N$ 的設定對每張圖做高斯雜訊；曝光過度的模擬則將圖片亮度隨機調為 100%、250%、400%。

測試時我們則使用了自己拍攝的車內影像。影像是將 Patriot F5 置於前擋風玻璃右上角進行拍攝的，圖片解析度為 $1280 \times 720$，幀率為 30 赫茲。影像中會有三個人，包含駕駛、副駕駛、後座的乘客。此資料集會有以下四個情境：白天進出隧道 900 張、夜間極暗 900 張、夜間等紅燈 900 張、夜間隧道內 1000 張 (如圖七)。白天進出隧道是四個情境中最亮的，用來測試進出隧道時的光線劇烈變化造成的影響；夜間極暗是一個環境光非常微弱的情境，用來測試缺乏環境光造成的

影響；夜間等紅燈是一個會受到紅燈直接照射的情境，用來測試資料有色差造成的影響；夜間隧道內是一個隧道光源較微弱的情境，用來測試環境光僅能照亮部分影像造成的影響。在後續的實驗我們也會將這四個情境分開測試。為求說明方便，後續提到資料集時會使用表 4.1 中對資料集的命名。

Table 4.1: 對不同資料集的命名

| 資料集名稱 | 說明 | 使用時機 |
|---|---|---|
| $D_{Original}$ | WIDER FACE 的原始資料集 | - |
| $D_{Dark}$ | 將 $D_{Original}$ 做曝光不足模擬後的資料集 | - |
| $D_{Bright}$ | 將 $D_{Original}$ 做曝光過度模擬後的資料集 | - |
| $D_{Triple}$ | 包含 $D_{Original}$、$D_{Dark}$、$D_{Bright}$ 三個資料集 | 預訓練 |
| $D_{Train}$ | 包含 $D_{Dark}$ 和 $D_{Bright}$ 兩個資料集 | 主要訓練 |
| $D_{Test}$ | 我們拍攝的車內影像 | 測試 |

## 4.2 實驗設定

訓練的流程包含預訓練和主要訓練。

在預訓練中，我們使用 $D_{Triple}$ 作為輸入資料集。首先我們將每張圖片調整為 $1024 \times 1024$ 的大小以配合後續訓練要求，然後將圖片切為 256 張 $64 \times 64$ 的小圖片餵進模型以 $\alpha = 0.05$ 進行訓練，經過 30 萬個時期 (Epoch) 後得到正規器的在主要訓練中的初始權重。接著我們進行主要訓練，在這個階段我們使用 $D_{Train}$ 作為輸入資料集，把預訓練中得到的正規器和人臉偵測器接在一起做 150 個時期的端對端訓練。測試時我們使用 $D_{Test}$ 作為輸入資料集，對其四個情境分別做測試和結果評估。

## 4.3 實驗結果

以下會展示用我們的方法做出來的結果和基線的結果在數據和視覺上的比較。基線是用將 FaceBoxes 這篇論文的架構以和論文中同樣的資料集 ($D_{Original}$) 訓練後得到的模型直接對 $D_{Test}$ 做測試，並已事先測試過該模型做在論文中提到的測試資料上結果和原先相似。數據上的結果比較如表 4.2，這裡的數據是我們用召回率 (Mean Average Precision) 對偵測結果計算得出的數字，並已經去除部分和本研究無關的結果影響。

Table 4.2: 基線和我們的方法之結果比較

| 模型名稱 | 白天進出隧道 | 夜間極暗 | 夜間等紅燈 | 夜間隧道內 |
|---|---|---|---|---|
| 基線 | 63.54% | 31.70% | 93.35% | 47.46% |
| 我們的方法 | 76.25% | 78.97% | 97.51% | 71.66% |

由數據結果可以發現我們的方法得出的結果在環境光較微弱的第 2、4 兩個情境下和基線相比有了顯著的進步，而在其他兩個情境則有小幅進步。從視覺上的結果 (見圖八) 也可看出我們的方法和基線相比能夠偵測出更多曝光不足的人臉，同時兼顧偵測受到正常光線照射下人臉的準確率。

## 4.4 消融實驗

在這一節，我們會進行各種不同訓練設定的測試，來確認我們的方法的各個步驟都是有效的。首先我們會先介紹在這一節中會用到的幾個訓練設定的細節，然後再展示數據間比較的結果並進行討論。

以下的設定包含了基線、雙倍資料、預訓練基線方法、預訓練基線端對端優化、原圖預訓練端對端優化、三圖預訓練、我們的方法這七個設定。基線方法和我們的方法在前面都已提過；雙倍資料是建立在基線之上，將訓練的資料集換成 $D_{Train}$ 的設定；預訓練基線方法是我們用 python 實作了 MSR-net 這篇論文的方法後，以經過視網膜增強算法處理 $D_{Dark}$ 後的圖片作為參考圖用 Ddark 訓練出一個正規器，固定正規器的權重後和人臉偵測器接在一起用 $D_{Train}$ 訓練，但僅訓練偵測器權重的設定；預訓練基線端對端優化便是預訓練和前個設定相同，但在主要訓練的階段進行端對端優化的設定；原圖預訓練端對端優化則在預訓練時以 Doriginal 作為 Dtrain 的參考圖訓練，主要訓練階段時以 $D_{Train}$ 進行端對端訓練的設定；三圖預訓練是在預訓練階段用三圖一組架構訓練，但主要訓練階段固定正規器的權重僅訓練偵測器權重的設定。詳細的數據比較如表 4.3。接下來我們分別對修改訓練資料集、在偵測前進行正規化處理、使用三圖一組架構進行預訓練、在預訓練後對正規器進行端對端訓練優化的成效進行比較與討論。

首先我們比較基線和雙倍資料這兩個設定的結果，這兩個設定只差在訓練所使用的資料集不同。我們發現使用較極端光線照射下的圖片作為輸入資料有助於讓偵測器認識更多在不同光線照射下的人臉，使其能夠偵測出更多人臉；數據上的

Table 4.3: 不同設定間之結果比較

| 模型名稱 | 白天進出隧道 | 夜間極暗 | 夜間等紅燈 | 夜間隧道內 |
|---|---|---|---|---|
| 基線 | 63.54% | 31.70% | 93.35% | 47.46% |
| 雙倍資料 | 70.97% | 63.16% | 90.75% | 62.73% |
| 預訓練基線 | 73.15% | 60.13% | 88.67% | 63.88% |
| 預訓練基線端對端優化 | 74.66% | 81.21% | 95.42% | 69.54% |
| 原圖預訓練端對端優化 | 73.56% | 76.02% | 96.59% | 68.73% |
| 三圖預訓練 | 22.61% | 41.13% | 83.31% | 30.35% |
| 我們的方法 | 76.25% | 78.97% | 97.51% | 71.66% |

結果在大部分情境下都有顯著進步，僅在等紅燈的情境下準確率略掉，但依然表現得很好。

接著我們比較雙倍資料、預訓練基線端對端優化、原圖預訓練端對端優化、我們的方法這四個設定，四個設定中的後三個設定採用了不同預訓練設定，但共通的是都有對圖片進行正規化處理，只有第一個設定沒有進行這樣的處理。我們發現對輸入資料進行正規化處理對人臉偵測很有幫助，而且在環境光較微弱的情境下更加明顯，這代表正規化處理有成功將圖片中的人臉調整到和其他情境下的人臉相似，使偵測更加順利。

然後我們比較原圖預訓練端對端優化和我們的方法。這兩個設定主要差別在預訓練的架構上，但使用的資料集總和一致。我們發現即使在預訓練和主要訓練這兩個設定使用的資料集總和一致，但我們的方法的結果在數據上卻比較好，這代表預訓練時三圖一組的架構能夠有效幫助正規器學習如何對圖片進行正規化處理。

最後我們分別比較兩對設定：預訓練基線和預訓練基線方法端對端、三圖預訓練和我們的方法，這兩對設定都是其中一個設定沒有對正規器進行端對端訓練優化，而另一個有。由數據結果我們可以發現單單對圖片進行正規化處理是不夠的，如果不進行後續的優化有可能結果反而會比不做正規化處理還要差。我們猜測這是由於正規器對圖片做的調整雖然拉近了不同光照情境下圖片間的距離，卻增加了一些本不應存在的雜訊，使得偵測器無法順利偵測出人臉。而端對端訓練的意義便是對正規器做的事情進行調整，使它在拉近圖片間距離的同時顧及偵測器對人臉偵測的需求並加以改進。

# Chapter 5

# 總結與未來目標

我們提出了三圖一組的訓練架構以訓練能將不同光線照射下的圖片正規化成相似結果的正規器，並將正規器和人臉偵測器接在一起做端對端訓練優化來提升人臉偵測的準確率。我們的方法證實對正規器的訓練有幫助，也提升了整體人臉偵測的準確率。

未來我們能夠改進的地方包含修改三圖一組架構中的損失函數使其能將圖片的正規化處理做得更好和修改生成曝光不足、曝光過度圖片的演算法，使其能夠更貼近真實圖片的樣貌以利訓練進行。

# Bibliography

[1] L. A. Rowe and R. Jain. Acm sigmm retreat report on future directions in multimedia research. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 1(1):3–13, 2005.