Abstract

挑戰：overlapping events, the presence of background noise, and the lack of benchmark data sets.

Introduction

YOLO：YOLO（You Only Look Once）是一種深度學習算法，主要用於物體檢測。它的主要特點是將物體檢測視為一個回歸問題，通過單次前向傳遞來同時預測多個物體的位置和類別。這種方法相較於傳統的物體檢測方法（通常需要多次處理圖像）更為高效。

1. 單一網絡結構：使用一個單一的卷積神經網絡（CNN）來進行整張圖像的處理，這樣可以在一次推理中同時識別多個物體。

2. 速度快：由於只需一次前向傳遞，能夠以實時的速度進行物體檢測，這使得它特別適合需要快速反應的應用場景，如自駕車和視頻監控。

3. 整體視野：YOLO 在處理整張圖像時考慮了上下文信息，能更好地捕捉物體之間的關係，減少了誤檢和漏檢的情況。

4.多樣性：YOLO 可以檢測多種不同類別的物體，並且能夠適應不同的場景和光照條件。

YOHO（You Only Hear Once）是一種專門用於音頻事件檢測的深度學習算法，類似於 YOLO 算法在計算機視覺的應用。YOHO 旨在識別音頻信號中的聲音事件，同時確定它們的開始和結束時間。

Methodology

Fourier Transform：Time domain →Frequency domain

Cepstral Features：將原來訊號的頻譜先轉成類似分貝的單位，再作逆傅立葉轉換，把它視為一種新的訊號做處理

Spectrogram：又稱聲譜圖（voicegram），是一種描述波動的各頻率成分如何隨時間變化的熱圖

Wavelet Transform：是指用有限長或快速衰減的「母小波」（mother wavelet）的振盪波形來表示訊號。該波形受縮放和平移以匹配輸入的訊號。

Statistical Features：可透過分析計算所得的數據特性

其他研究則使用常數 Q 倒譜係數(CQCC)和常數 Q 變換(CQT)來提取聲音事件的特徵，還有一些方法利用共同子空間學習(CSL)提取複雜環境中多個事件的語義信息，強調語義特徵提取的重要性。音頻事件檢測在健康監控和安全應用中非常重要。雖然手工設計的特徵可用於音頻事件檢測，但使用如譜圖、梅爾譜圖等非手工特徵的研究也顯示了更好的效果。

HMM

針對一個已知的輸出數列，我們調整函數，讓該輸出數列的出現機率最大，同時求出機率多寡。（理

論上應該要同時考慮所有輸出數列，讓整體的機率最大；但是這樣時間複雜度太高是指數時間，只好一次處理一個。）

針對一個新的輸出數列，我們找到可能性最高的輸入數列，同時求出機率多寡。

另外 HMM 可以用來分類數列。每一種類別，各自建立一個 HMM 。針對一個新的輸出數列，以機率多寡來判斷其分類。

GMM

KMeans： 透過最小化資料點與其各自群集質心之間的平方距離總和，將資料劃分為 K 個群集。

RNN-LTSM

LSTM 主要改善了以前 RNN 的一些 Memory 的問題，其由四個 unit 組成: Input Gate、Output Gate、Memory Cell 及 Forget Gate。

- ▶ Input Gate: 當資料輸入時，input gate 可以控制是否將這次的值輸入，並運算數值
- ▶ Memory Cell: 將運算出的數值記憶起來，以利下個 cell 運用
- ▶ Output Gate: 控制是否將這次計算出來的值 output，若無此次輸出則為 0
- ▶ Forget Gate: 控制是否將 Memory 清掉(format)

Deep learning classifiers

這篇文章回顧了音頻事件識別（AER）中使用的深度學習模型和技術

Stowell 等人使用深度學習技術自動識別音訊，並使用了基線分類器和現代分類器來提高性能。此外，殘差神經網絡（ResNets）因其能解決梯度消失問題而受到廣泛關注。前者轉換訊號樣本為 MFCC 表示法，且其分布透過 GMM 表示並模型化；後者主要區分各種音頻的種類。

殘差神經網絡（ResNets）是 CNN 模型的一個重要里程碑。實作上，只要將上層 input x 直接加入經 non-linear 轉換的輸出 F(x)，如此一來不但能得到輸出 F(x) + x，也不會增加額外的參數增加運算量，常見的深度學習框架也都可以簡單實現。

這些深度學習方法在 AER 任務中表現出色，未來需要探索新架構以提升在背景噪音下的即時應用性能。此外，開發即時音頻監控系統應考慮在多音源條件下識別異常聲音事件的挑戰。

Audio Surveillance

僅依賴視頻監控往往不足以準確識別事件。音頻分析不僅成本較低，且對於數據流和計算資源的需求較少。與攝影機相比，麥克風可以是單向或全向的，提供更廣的視野。此外，音頻波能夠穿透障

礙物，這是視頻處理的弱點。

音頻指紋技術

是一種通過提取音頻中的相關特徵來簡化和表示音頻流的最小特徵狀態。儘管該技術仍處於初期階段，但目前認為有效的音頻指紋框架應具備以下特性：

1、穩健性：在噪音環境下正常工作，並保持準確性

2、成對獨立性：避免重複

3、快速數據庫查詢：必須能夠快速且高效地查找

4、通用性：能夠對不同來源的音頻進行識別。

5、可靠性：語音識別系統應具備穩定和有效的性能

6、脆弱性：能夠檢測原始音頻信號的變更