EasyChair Preprint

№ 5347

Human Activity Recognition Based on Gramian Angular Field and Convolutional Neural Network

Wan-Yi Hsieh and Jui-Chung Hung

April 18, 2021

结合格拉姆角場與卷積神經網路 應用於人類動作辨識

謝宛頤 [1] 洪瑞鍾 [1,*]
[1] Department of Computer Science
University of Taipei, Taipei 100, Taiwan
juichung@gmail.com

摘要

人類動作辨識(Human Activity Recognition, HAR)是使用感測器資料來預測人類的動作，在物聯網的進步和微機電系統的發達下，HAR 在日常生活中的應用越來越普及，如手機、智慧手環都有內建能夠偵測身體的動作和狀態的感測器，可即時地預測使用者的活動。但因為感測器蒐集到的資料是具有時間序列性質的資料，其中的特徵值很難萃取，如果直接使用深度學習的方式索取特徵，這樣會無法保留資料中時間序列的特質，本研究中使用感測器偵測到加速度值作為訓練資料，將原本多個一維的時間序列資料利用格拉姆角場(Gramian Angular Field, GAF)做二維的圖像轉換，GAF 將原始感測器資料的直角座標轉換為極座標的方式來保有時間序列資料的相關性和連續性，並以三軸資料合併轉為二維的方式做為資料的輸入，分類器使用卷積神經網路(Convolutional Neural Network, CNN)，可以自動的從圖像資料中萃取特徵值。本研究使用 Actitracker 的資料集，比較我們所提出的方法與 CNN 模型相比正確率提升了 5.8%。

1 前言

智慧型手機在現代社會中已經是不可或缺的存在，由圖 1 可見近年來智慧型手機的普及率每年都在上升，全球手機普及率為智慧型手機使用者人數除以全球人口數，2019 年手機普及率達到 41.5%，2020 年手機普及率比 2019 年上升 5%。

依據調查，目前台灣有 88.2%的民眾平常會使用智慧型手機[1]，在智慧型手機普及之前，HAR 是在使用者的身體上穿戴數個感測器進行蒐集資料，資料的使用情境單一且蒐集的裝置繁瑣[2]，智慧型手機的普及再加上微機電系統(Microelectromechanical Systems, MEMS)的進步使智慧型手機有越來越多種類的感測器能偵測身體的動作狀態[3]。感測器數據使智慧型手機使用者受益在許多領域而且有廣泛的研究應用。

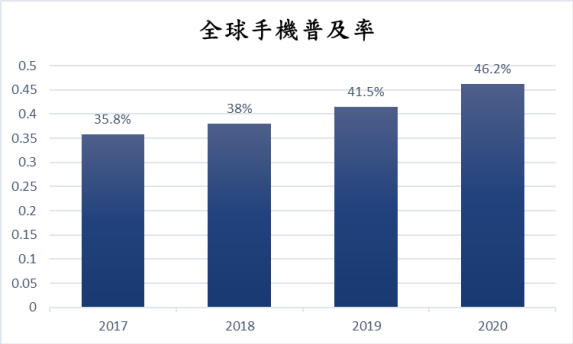

圖 1 全球手機普及率[4, 5]

HAR 使用感測器的數據來偵測身體的動作狀態，應用例如：老年人跌倒監控、工廠流程偵錯、舉重練習助理等等。HAR 現階段最常使用的感測器是隨身攜帶的智慧型手機或是智慧型手錶，人類行動時感測器蒐集的資料含有時間序列的特性。GAF 可以將時間序列資料轉換為有極座標資訊的圖像資料[6]，輸入到分類模型的資料就可保有時間序列資料的時序性。CNN 在圖形中有非常好的辨識力，所以處理資料的時候較好的方式是二維輸入，把蒐集到具有時間序列性質的 HAR 資料進行 GAF 的轉換， 本研究使用公開的 HAR 資料集 Actitracker 作為實驗資料，利用資料窗格的方式對資料進行前處理，對於長時間並且重複的活動，每一個特徵值裡面會有重複特性的變化，於是本研究選用 GAF 的方式將時間序列性質的資料經過二維化的轉換之後，得到的圖像會有兩個時間點之間的資訊，本研究因此結合 GAF 和有自動萃取特徵的 CNN 來建立分類器。

2 時間序列資料二維化

本研究採用美國紐約 Fordham 大學中無線感測器資料探勘實驗室(Wireless Sensor Data Mining Lab, WISDM)提供的 Actitracker 資料集[7]。該資料集紀錄 36 名使用者在褲子的前口袋放智慧型手機進行日常生活活動(Activity of daily living, ADL)的數據，是在受控的實驗室環境下收集的數據。智

慧型手機內建一個三軸加速度計，取樣頻率為 20 赫茲。資料集共包含 1,098,207 個取樣資料。使用者活動的類型共 6 種，分別為步行、慢跑、上樓、下樓、坐著、站著，36 位使用者做 6 種活動的資料筆數可見附錄一。由圖 2 可見，每種活動的數量相差很大，Actitracker 是一個不平衡的資料集。

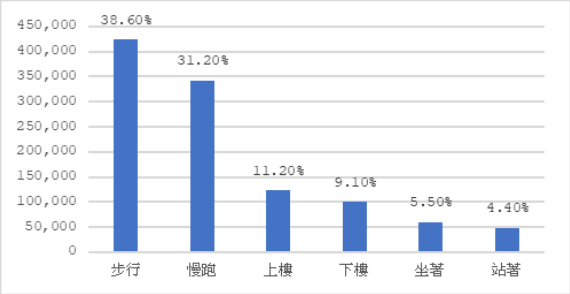

圖 2 Actitracker 資料集的活動取樣筆數

一般蒐集到的 HAR 資料，會以移動視窗(Sliding Window)的方式來取得每筆資料的物件，其中窗口的大小(Window Size)和位移(Stride)的數值必須固定[8]，視窗移動的秒數為 2 秒內被證明可以獲得精確的檢測結果[9]，位移為重疊(Overlap)50%，視窗的前半部分會包含前一個視窗後半部分的觀測資料[10]。圖 3 為窗格重疊 50%的示意圖。以第一位測試者的下樓梯活動為例，總共採樣 2,941 筆，設定重疊比例為 50%以及窗格秒數為 2 秒，則窗口大小為 40 筆採樣資料，並可依上述設定得到第一位測試者的下樓梯有 146 個資料窗格。

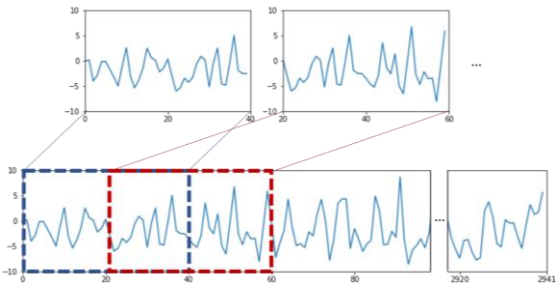

圖 3 窗格重疊

GAF 主要發想於格拉姆矩陣(Gramian matrix)，格拉姆矩陣是多個向量之間互相的內積所組成的對稱矩陣，所以可以獲得向量之間的關聯性，應用於影像的風格轉換[11]。但若 HAR 使用格拉姆矩陣效果則會有侷限，因為不能完整保留兩個值所給出的資訊。Wang 提出的一種時間序列的編碼方法 GAF[6]，是為類格拉姆矩陣(Quasi-Gramian Matrix)，將笛卡爾座標軸(Cartesian Coordinate)的一維時間序列轉為極座標軸(Polar Coordinate)的角度資訊，再經由三角函數轉換為二維的對稱矩陣，使結果可保有時間序列資料的相依性和連續性，使得模組更具有強健性，常用於時序性資料估測。GAF 矩陣的作法可分為通過資料正規化、坐標軸的變換和三角函數三個步驟，分述如下，第一步是資料正規化(Normalization)，將給定的時間序列 N 筆資料 $\boldsymbol{x}=\{x_{(1)}, x_{(2)}, \cdots, x_{(N)}\}$，經過正規標準化的處理後，使數值在[-1,1]區間，可表示：

$$\tilde{x}_{(i)} = \frac{(x_{(i)} - \max(\boldsymbol{x})) + (x_{(i)} - \min(\boldsymbol{x}))}{\max(\boldsymbol{x})\text{-}\min(\boldsymbol{x})} \tag{1}$$

其中 $\max(\boldsymbol{x})$代表序列裡面的最大值，$\min(\boldsymbol{x})$代表序列裡面的最小值，$x_{(i)}$為時間序列的第 i 筆資料，經過標準正規化之後表示為 $\tilde{x}_{(i)}$。

第二步將正規化後時間序列資料 $\tilde{x}_{(i)}$轉換為極坐標系統，式子可表示：

$$\begin{cases} \varphi_{(i)} = \arccos\left(\tilde{x}_{(i)}\right), -1 \le \tilde{x}_{(i)} \le 1 \\ r_{(i)} = \dfrac{i}{N}, i = 1, 2, ..., N \end{cases} \tag{2}$$

其中 $\varphi_{(i)}$為第 i 筆資料的極座標軸的角度、$r_{(i)}$為第 i 筆資料的極座標軸的半徑，$\varphi_{(i)}$的值在$[0, \pi]$的區間，且有單調遞減(Monotonic Decreasing)的特質。時間序列的值轉換成極座標的角度，而時間採樣則轉換為極座標的半徑，半徑 $r_{(i)}$會因時間進行而逐漸變長，相應的數值 $\tilde{x}_{(i)}$在圓的不同角度改變，極座標圖型像水波紋一樣呈螺旋狀，所以在此區間具有對射(Bijection)的性質，即為給定一個時間序列，其對射在極坐標系中只有產生一個結果，且逆向有唯一值。如圖 4 假設 N 為 6 筆的時間序列他所對應到的極座標的轉化為下圖所示：

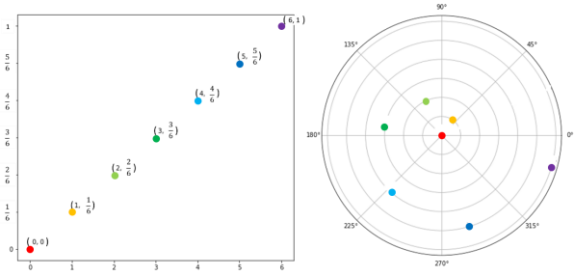

圖 4 直角座標轉換及座標

若直接將調整後的時間序列資料 $\tilde{x}_{(i)}$做格拉姆矩陣(Gramian matrix)將無法得到完整的資訊，因為直接將值兩兩做內積，其結果將偏向於時間長、半徑大的一個值，而且當內積到自己時，得到的結果也會喪失資訊。所以 Wang 提出了可以保有兩個值之間資訊的 GAF 方法[6]，優點是保有時間的依賴性(Temporal Dependency)。即當時間逐漸增加時，對應的時序資料是從矩陣的對角線由右下方至左上方進行存取，該矩陣形成的圖像代表了時間序列的角度資訊外在矩陣中仍可得到時間的順序。若使用兩角和的餘弦函數則得到 Gramian

Angular Summation Field (GASF)，若使用兩角差的正弦函數則得到 Gramian Difference Angular Field (GADF)，矩陣 GASF 和 GADF 具體表示為：

$$\mathrm{GASF}=\begin{bmatrix}\cos(\varphi_{(1)}+\varphi_{(1)}) & \cdots & \cos(\varphi_{(1)}+\varphi_{(N)})\\ \cos(\varphi_{(2)}+\varphi_{(1)}) & \cdots & \cos(\varphi_{(2)}+\varphi_{(N)})\\ \vdots & \ddots & \vdots\\ \cos(\varphi_{(N)}+\varphi_{(1)}) & \cdots & \cos(\varphi_{(N)}+\varphi_{(N)})\end{bmatrix} \quad (3)$$

$$\mathrm{GADF}=\begin{bmatrix}\sin(\varphi_{(1)}-\varphi_{(1)}) & \cdots & \sin(\varphi_{(1)}-\varphi_{(N)})\\ \sin(\varphi_{(2)}-\varphi_{(1)}) & \cdots & \sin(\varphi_{(2)}-\varphi_{(N)})\\ \vdots & \ddots & \vdots\\ \sin(\varphi_{(N)}-\varphi_{(1)}) & \cdots & \sin(\varphi_{(N)}-\varphi_{(N)})\end{bmatrix} \quad (4)$$

其中 $\cos(\varphi_{(i)}+\varphi_{(j)})$、$\sin(\varphi_{(i)}-\varphi_{(j)})$簡化後可得：

$$\begin{aligned}\cos(\varphi_{(i)}+\varphi_{(j)}) &= \cos(\varphi_{(i)})\cdot\cos(\varphi_{(j)})-\sin(\varphi_{(i)})\cdot\sin(\varphi_{(j)})\\ &= \cos(\arccos(\tilde{x}_{(i)}))\cdot\cos(\arccos(\tilde{x}_{(j)}))\\ &\quad -\sin(\arccos(\tilde{x}_{(i)}))\cdot\sin(\arccos(\tilde{x}_{(j)}))\\ &= \tilde{x}_{(i)}\cdot\tilde{x}_{(j)}+\sqrt{1-\tilde{x}_{(i)}^2}\cdot\sqrt{1-\tilde{x}_{(j)}^2}\end{aligned} \quad (5)$$

$$\begin{aligned}\sin(\varphi_{(i)}-\varphi_{(j)}) &= \sin(\varphi_{(i)})\cdot\cos(\varphi_{(j)})-\cos(\varphi_{(i)})\cdot\sin(\varphi_{(j)})\\ &= \sin(\arccos(\tilde{x}_{(i)}))\cdot\cos(\arccos(\tilde{x}_{(j)}))\\ &\quad -\cos(\arccos(\tilde{x}_{(i)}))\cdot\sin(\arccos(\tilde{x}_{(j)}))\\ &= +\sqrt{1-\tilde{x}_{(i)}^2}\cdot\tilde{x}_{(j)}-\tilde{x}_{(i)}\cdot\sqrt{1-\tilde{x}_{(j)}^2}\end{aligned} \quad (6)$$

3 分類器架構

CNN 能夠利用褶積過濾器的特質，可以從數據中自動學習到資料的特徵[12]。HAR 中的原始資料大多為一維的時間序列資料，Zeng 提出了 CNN 的方法來做 HAR 的辨識[13]，因為 CNN 的特質會具有局部依賴性(Capture Local Dependency)和規模不變性(Scale Invariance of a Signal)。局部依賴性指 Zeng 使用 CNN 提取輸入感測器資料的特徵，如果把運動得到的整筆資料都做為一張圖像輸入 CNN，卷積層會無法抽取和預測目標相關度較低的特徵，因為運動中鄰近秒數的感測器資料是有依賴性的，當這個依賴性在邊界的時候，抽取特徵的效能會降低，所以 Zeng 將運動資料分段後以重疊方式做為輸入資料。規模不變性是指當 CNN 在圖像識別時，訓練的圖像特徵可以有不同的尺寸，而在 HAR 時受試者可能以不同頻率的步伐行走或是進行不同強度的運動，CNN 就可以得到很好的效果。Hammerla 通過在一些公共 HAR 資料集上進行深度神經網路(Deep Neural Network, DNN)、遞迴神經網路(Recurrent Neural Network, RNN)和 CNN 的實驗，提出深度學習方法中主要有兩種方法適用於時間序列的資料，分別是 CNN 和 RNN。在持續時間短但具有時間順序的活動上，RNN 的性能會優於 CNN，因為遞迴方法有長時間觀察背景訊號的能力。但對於長時間或重複的活動，每一個特徵值裡面會有重複特性的變化，例如走路或跑步等，建議使用 CNN[14]，原因是可以透過多通道的卷積層將這些資訊取得相關的特徵。

CNN 的主要結構是由卷積層(Convolution Layer)和池化層(Pooling Layer)重複多次組成後，最後使用全連接層(Fully Connected Layer)建立分類器。卷積層的作用是針對某一個特性去做增強，池化層的功能是把特徵萃取出來，接著我們將依序說明卷積層、池化層和全連接層。

- 卷積層

本研究將探討二維的卷積，一般在卷積過程的式子可以表示為：

$$z(n_1,n_2)=\sum_{k_1=1}^{M_h}\sum_{k_2=1}^{M_w} f(k_1,k_2)\,y(n_1-k_1,n_2-k_2) \quad (7)$$

其中 $z(n_1, n_2)$為經過卷積之後第 n_1,列、第 n_2 行的輸出資料，$f(k_1, k_2)$代表過濾器第 k_1 列、第 k_2 行的過濾器值，M_h 和 M_w 是過濾器的長度和寬度，$y(n_1-k_1, n_2-k_2)$代表輸入訊號第 n_1-k_1 列、第 n_2-k_2 行的資料。

以圖 5 為例，採用 5x5 為輸入訊號，過濾器的長寬為 2x2，將過濾器放置在選定的像素上之後，再從過濾器中提取每個值，並將它們與輸入圖像中的相應值相乘。最後再將所有的乘積相加總，並將結果放在輸出位置上，經計算之後輸出特徵圖的尺寸為 4x4。

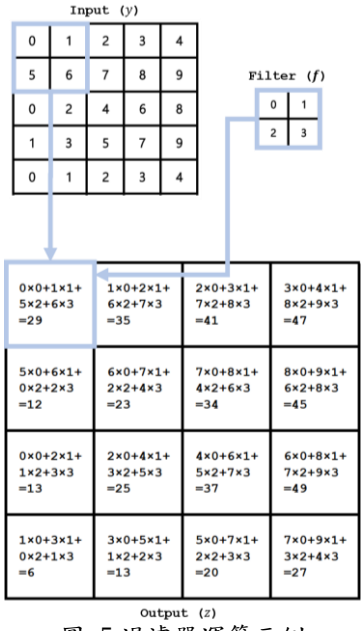

圖 5 過濾器運算示例

- 池化層

池化層要做卷積後的特徵萃取，一般進行最大或平均池化[15]。池化層能夠減少數據的大小、參數的數量、和所需的計算量，只保留特徵圖最重要的特性。池化層會擷取最重要的特徵所以會把卷積後的資料減少，作為下一層卷積層的使用，可以加快資料的訓練過程[16]。本研究採用最大池化層過程可以表示為：

$$c(d_1,d_2)=MAX\left\{\begin{matrix} z((d_1-1)\times S \\ +e_1,(d_2-1)\times S+e_2)) \end{matrix}\right\},$$

$$e_1=1,2,...,P_h,e_2=1,2,....,P_w \tag{8}$$

其中 $c(d_1, d_2)$為經過池化之後第 d_1 列、第 d_2 行的輸出資料，z 為池化層的輸入項，P_h和 P_w 分別代表池化過濾器的長和寬，S 為步幅大小。以圖 6 為例，步幅大小為 2，使得 2x2 池化過濾器將輸入矩陣分成 4 個區域。過濾器在前一個區域選取最大值後，會轉移到下一個 2x2 區域。首先，過濾器在{29,35,12,23}區域，取最大值作為輸出結果。再進行位移 2 步後到下一個區域{41,47,34,45}取最大值，直到結束。已知如果把卷積層看作是特定特徵的檢測器，那麼池化層處理的是將該特徵的最大值保留在池化矩形內。

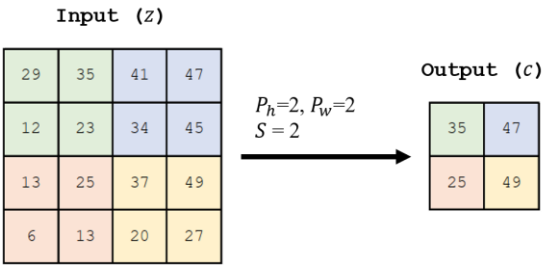

圖 6 池化層的處理程序

- 全連接層

全連接層會放在 CNN 中的末層，功能是最後的分類器。特徵圖矩陣經過扁平化後輸入至全連接層，從而計算權重與前一層之間的乘積，得到不同類型的正確概率，辨識輸入的圖像是屬於哪一個類別，全連接層的輸出值會被匯集成一個結果的輸出。本研究中，所使用的卷積神經網路架構以[13]為參考。如圖所示，包含兩層卷積層、兩層池化層和一層輸出層(Output layer)。而激勵函數使用 Softmax 函數。資料窗格共為 54634 筆，使用 75%資料集做為訓練資料 25%資料集作為驗證資料。

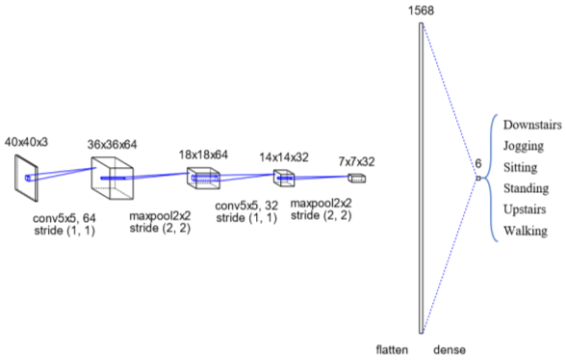

圖 7 CNN 架構圖

4 實驗結果

本研究採用 GAF 結合 CNN 的方法，得到準確率如圖 8 所示。分別使用統計方法(Statistical) [17]、受限玻爾茲曼機(Restricted Boltzmann Machine, RBM) [13]、主成分分析法(Principal Component Analysis, PCA) [13]、同架構的 CNN，以及 GAF 結合 CNN 的方法的準確率以圖 9 所示。

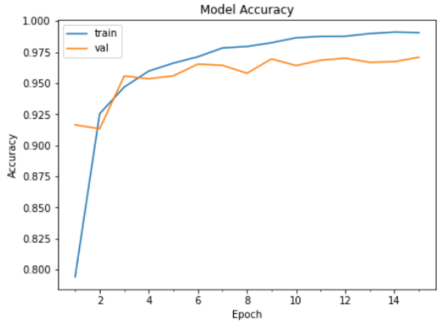

圖 8 實驗結果

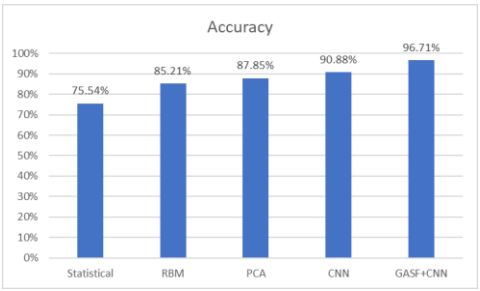

圖 9 結果統計

5 結論

本研究選用 GAF 的方式將時間序列性質的資料經過二維化的轉換之後，得到的圖像會有兩個時間點之間的資訊，並利用 CNN 自動萃取特徵能力，找出感測器蒐集到的人類活動資料中可能存在的規則。透過上述的研究方法，實證上顯示可以得到較佳的預測結果。

6 致謝

感謝臺北市立大學提供豐富的學習資源以及良好的環境，我才能學以致用並且完成學業。

參考資料

[1] 臺灣傳播調查資料庫."台灣民眾手機使用習慣調查." http://www.crctaiwan.nctu.edu.tw/epaper/%E7%AC%AC90%E6%9C%9F20190815.htm (accessed.

[2] U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher, "Activity recognition and monitoring using multiple sensors on different body positions," in *International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06)*, 2006: IEEE, pp. 4 pp.-116.

[3] 陳佳良, "基於人工智慧與手機感測的使用者行為辨識," 撰者, 2018.

[4] T. Gu. "Newzoo's 2020 Global Mobile Market Report." Newzoo. https://newzoo.com/key-numbers/ (accessed.

[5] U. N. U. N. D. o. E. a. S. A. UNDESA. "World population." United Nations. https://population.un.org/wpp/Download/Standard/Population/ (accessed.

[6] Z. Wang and T. Oates, "Encoding time series as images for visual inspection and classification using tiled convolutional neural networks," in *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, 2015, vol. 1.

[7] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74-82, 2011.

[8] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE communications surveys & tutorials*, vol. 15, no. 3, pp. 1192-1209, 2012.

[9] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors*, vol. 14, no. 4, pp. 6474-6499, 2014.

[10] J. O. Laguna, A. G. Olaya, and D. Borrajo, "A dynamic sliding window approach for activity recognition," in *International conference on user modeling, adaptation, and personalization*, 2011: Springer, pp. 219-230.

[11] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," *arXiv preprint arXiv:1705.08086*, 2017.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[13] M. Zeng *et al.*, "Convolutional neural networks for human activity recognition using mobile sensors," in *6th International Conference on Mobile Computing, Applications and Services*, 2014: IEEE, pp. 197-205.

[14] N. Y. Hammerla, S. Halloran, and T. Plötz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," *arXiv preprint arXiv:1604.08880*, 2016.

[15] S. Ha, J.-M. Yun, and S. Choi, "Multi-modal convolutional neural networks for activity recognition," in *2015 IEEE International conference on systems, man, and cybernetics*, 2015: IEEE, pp. 3017-3022.

[16] Y. Bengio, "Deep learning of representations: Looking forward," in *International Conference on Statistical Language and Speech Processing*, 2013: Springer, pp. 1-37.

[17] T. Plötz, N. Y. Hammerla, and P. L. Olivier, "Feature learning for activity recognition in ubiquitous computing," in *Twenty-second international joint conference on artificial intelligence*, 2011.