航測及遙測學刊　第三十卷　第 3 期　第 139- 151 頁　民國 114 年 9 月
Journal of Photogrammetry and Remote Sensing
Volume 30, No.3, 2025, pp. 139 – 151
DOI：10.6574/JPRS.202509_30(3).0001

應用深度特徵於多視角車載影像匹配之研究

張智安 [1*]　陳沛丞 [2]　陳亭霓 [3]　楊軒 [4]　林展慶 [4]　李冠毅 [4]
洪愷頡 [4]　呂振永 [4]

摘要

隨著移動測繪技術的進步，多視角車載影像逐漸成為道路觀測的重要資料來源，但傳統方法難以克服畸變與視角差異。為了提升影像匹配的精度和泛化能力，本研究探討基於深度學習的深度特徵匹配技術 (Deep Feature Matching, DFM)，利用 VGG19 預訓練模型與 CNN 卷積神經網路萃取特徵，並結合兩階段匹配策略與 RANSAC 演算法過濾錯誤點以確保可靠性，達成多視角車載影像的精確匹配和物空間三維定位。實驗採用三相機車載影像，以交通標誌作為定位目標。結果顯示，DFM 在尺度差異、畸變與遮蔽等情境下，匹配成功率與精度均優於 SIFT，特別在畸變與遮蔽下表現更佳。交通標誌定位成功率約 70%，平均誤差小於 0.5 m，證實 DFM 具備多視角三維定位的應用潛力。

關鍵詞：深度學習、深度特徵、影像匹配、多視角影像、交通標誌

1. 前言

1.1 研究背景與動機

隨著移動測繪載具技術的發展，以車載相機獲取多視角影像成為重要的道路觀測資料來源，這些車載影像可以提供大量觀測資料，且可應用於建構環景的道路環境，並可進一步從空間前方交會得到三度空間的坐標，因此多視角影像的影像匹配是一項重要的議題。

傳統的影像匹配方法通常依賴人工設計的影像特徵進行匹配，例如 SIFT (Scale-Invariant Feature Transform) (Lowe, 2004)、SURF (Speeded-up robust features (SURF) (Bay *et al.*, 2008)或 ORB (Oriented FAST and Rotated BRIEF) (Rublee, *et al.*, 2011) 等。然而，傳統方法在光線變化和視角較大變化的情境中，可能有所限制。為了提升影像特徵於匹配過程的泛化能力，從深度學習模型中萃取深度特徵 (Deep Features)，有助多視角影像匹配。深度特徵對光線變化和視角較大的場景擁有更強的抗變能力，可提高配對的精確度與突變性 (Dusmanu *et al.*, 2019)。

1.2 相關研究

傳統的影像匹配方法主要分為以區域為基礎 (Area-based Methods) 的方法和以特徵為基礎 (Feature-based Methods) 的方法。在以區域為基礎的方法中，這類方法最具代表性的是相關係數法，也就是 Normalized Cross-Correlation (NCC) (Zitová & Flusser, 2003)。NCC 依賴影像區域之間的相似性來轉換成相關係數值，以判斷影像視窗的匹配情況。在以特徵為基礎的方法中，特徵基礎的方法通過萃取影像中的特徵，如角點、或線段，其中，最為代表的是 SIFT (Scale-Invariant Feature Transform) (Lowe, 2004)，SIFT 是一種廣泛應用於影像匹配的方法，主要目標是從影像中提取出具有尺度不變性和旋轉不變性的關鍵點特徵符 (Keypoint Descriptor)，以便在不同視角、尺度變化、旋轉的情境下，依然能夠可靠地進行影像匹配。演算法核心流程為計算局部影像像素的灰階直方圖分佈產生

[1] 國立陽明交通大學土木工程學系 教授
[2] 國立陽明交通大學土木工程學系 博士候選人
[3] 國立陽明交通大學土木工程學系 碩士生
[4] 台灣世曦工程顧問股份有限公司地理空間資訊部 工程師
[*] 通訊作者, E-mail: tateo@nycu.edu.tw

收到日期：民國 114 年 01 月 10 日
修改日期：民國 114 年 04 月 23 日
接受日期：民國 114 年 05 月 09 日

影像梯度，並將此梯度分佈組合成關鍵點特徵符。此類人工設計的特徵萃取方式後續催生了例如 SURF 等演算法，廣泛應用於攝影測量領域。

近年來深度學習廣泛應用在影像處理領域(He *et al.*, 2016)，在使用深度學習進行影像匹配的方法中，影像匹配通常需要依賴大量的資料集來訓練模型。此類方法有兩種主要策略：一個重要方向是使用 Siamese Network 來完成影像匹配。Siamese Network 的核心是將兩張影像分別擺入兩個一樣的模型結構中，通過其中的層層轉換，將影像的特徵向量轉換成一個更具代表的特徵表示。之後，將兩個特徵向量作差異或雙輪模型來判斷影像之間的相似度。如果兩張影像為相似影像，那麼其差異就較小，通過此分類器來判斷影像是否相似，進而達成影像匹配之目的。使用 Siamese Network 的重要優勢是可以直接利用影像之間的相對關聯來進行判識 (Melekhov *et al.*, 2016)，這類深度學習的影像匹配方法，對於與資料特性相似的影像具有良好的表現能力，適用於與訓練資料特性相似的影像，但如果預測影像特性與訓練資料集有明顯差異，模型的表現能力可能會下降。使用深度學習進行影像匹配的另一個策略是使用預訓練模型計算深度特徵 (Efe *et al.*, 2021、Zagoruyko & Komodakis, 2015)，與需要重新訓練深度學習模型的方式不同，該方法利用預訓練模型計算深度特徵，充分應用遷移學習 (Transfer Learning) 的概念。在這種策略中，常使用的預設模型包括 VGG16、ResNet 和 Inception 等預訓練模型。這些預訓練模型使用如 ImageNet 這類的大規模訓練資料集上進行訓練，得到的特徵表示能力，對於多種遷移學習的任務均有良好的表現能力。從預訓練模型得到深度特徵後，經由特徵比對，可達成影像匹配之目的。這類方法的另一大優點是免去重新訓練模型，因不需要重新訓練，故有較佳的使用彈性。

比較 SIFT 演算法及深度學習匹配方法，SIFT 使用人工設計的關鍵點特徵符，而深度學習使用 CNN 架構透過大量訓練資料進行資料導向 (Date Driven) 建立深度特徵萃取方式，SIFT 適用在相同感測器的資料，而深度學習影像匹配方法經由模型訓練與學習，可應用在異質感測器的匹配中，例如 Hughes *et al.*(2018)，使用 Pseudo-Siamese CNN 深度學習架構，克服光學影像與雷達影像的差異性，達成異質資料間的匹配。

近年來深度學習技術的快速發展，大幅提升了影像匹配的準確性，成功應對如視角變化、光照差異與旋轉等挑戰。傳統上，局部特徵匹配方法可分類為基於偵測器 (Detector-based) 與非偵測器 (Detector-free) 兩大類 (Xu *et al.*, 2024)。兩者差異在於是否直接使用特徵萃取偵測器提供之特徵，基於偵測器之特徵點匹配是先進行特徵萃取，再對兩影像之特徵進行匹配，例如 Superpoint (DeTone *et al.*, 2018) 使用單一特徵點萃取 CNN 網路架構進行匹配，Merkle *et al.* (2018) 則是利用條件式生成對抗網路 (cGANs) 匹配光學影像與雷達影像。而非偵測器則是直接比對兩圖片特徵。近年來的趨勢逐漸有許多基於影像特徵匹配的技術發展完善如 NCNet (Rocco *et al.*, 2018) 基於 CNN 特徵並使用共識網路進行匹配、LoFTR (Sun *et al.* 2021) 將 CNN 特徵以 Transformer 技術進行編碼後匹配。

針對遙測影像匹配，MU-Net 提出一種無監督多尺度架構，可有效處理多模態影像對間的大幅視角變形 (Ye *et al.*, 2022)。在旋轉不變性的挑戰上，SE2-LoFTR-4* 相較於傳統 LoFTR 展現更佳的旋轉影像匹配效能 (Bökman & Kahl, 2022)。Neighbourhood Consensus Network 採用端對端可訓練架構，透過半區域約束與弱監督策略，強化影像間的密集對應能力 (Rocco *et al.*, 2018)。Deep-Image-Matching 為一開源工具箱，整合傳統與深度學習方法，針對高解析度多視角影像匹配提供不同的匹配方案 (Morelli *et al.*, 2024)。綜合上述技術的發展，導入深度學習技術可提升了影像匹配在多樣化場景下的精度與表現。

1.3 研究目的

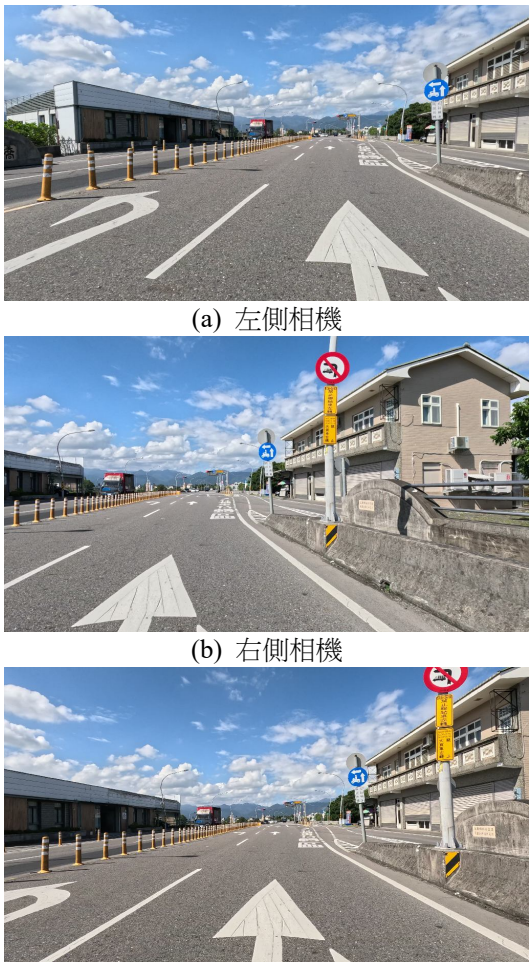
本研究之目的為探討深度特徵在多視角車載影像匹配的應用，由於車載相機常使用較廣角的相機，影像間有較大的變形，因此，本研究選擇基於深度學習的影像匹配技術，通過深度學習預訓練模

型提取的深度特徵。影像的深度特徵為 CNN 演算法經大量複雜資料集 (MS COCO 1K)訓練而來，其含有 100 多萬張影像 1000 個類別，在訓練過程中會自行學習最佳影像特徵萃取卷積核，對於不同光線和視角的變化擁有更強的泛化能力，因此可提高多視角車載影像匹配成功率及可靠性。由於車載前視相機的影像中，道路兩側的交通標誌會出現在靠近影像邊緣的兩側區域內，該區域通常有比較大的畸變，SIFT 的尺度不變特性不易吸收畸變造成的差異。又者本研究鎖定路側的交通標誌為匹配目標，影像尺寸約為 500 × 500 像素以內，SIFT 演算法在交通標誌的小影像匹配是可行的 (Ren *et al.*, 2009)，以深度特徵進行影像匹配大多數研究著重於整幅完整影像匹配，而非針對物件框選後局部影像間之匹配，應進一步實做驗證此方法的效益。

2. 研究資料

本研究實驗範圍位於宜蘭縣羅東鎮北城橋，測區道路長度約為 750 m。研究中使用三台 GoPro11 相機立多視角車載影像 (車頭三相機)，左右相機基線長度約 0.88 m，取樣資料格式為 5K 畫質的影片，5K 影片像幅大小為 5312 × 2988 pixels，取樣頻率每秒 29.97 幅，後處理以車載軌跡間隔 1 m 為相片取樣間距，每台相機取得 1479 張相片，三台相機共取得 4437 張相片，多視角影像間具有高重疊率，有利後續方位求解作業，三重疊之相片如圖 1 所示。

收集車載往返軌跡長度約為 1.5 km，資料獲取時，同時使用 eGNSS 記錄汽車載體的軌跡，以軌跡提供相機初始值進行方位求解 (Teo, 2015)，每間隔 200 m 至少佈設 1 個控制點，測試區共佈設 9 個控制點及 3 個檢核點，檢核點的平面及高程之均方根誤差分別 0.514 m 及 0.195 m，可滿足公尺(m) 等級的定位要求，解算後的相機位置、控制點及檢核點如圖 2 所示。

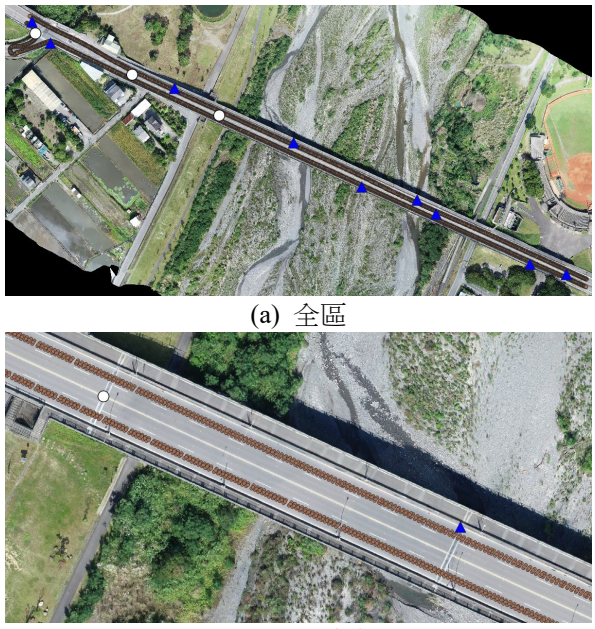


(a) 左側相機

(b) 右側相機

(c) 中間相機

圖 1 三重疊相片展示



(a) 全區

(b) 局部放大

圖 2 研究區域：相機位置 (紅色圓形)、控制點 (三角形) 及檢核點 (白色圓形)

3. 研究方法

本研究深度特徵影像匹配方法採用 Deep Feature Matching (DFM) (Efe *et al.*, 2021) 演算法，取得立體像對之間的共軛點後，接著，使用影像的內、外方位參數及共軛點坐標進行空間前方交會，以獲得影像點對應的三維坐標。

3.1 萃取深度特徵

由於行車影像採用的 GoPro 相機具有較大的畸變，影像存在較大的畸變差，使用傳統影像特徵萃取及匹配方式可能會造成匹配錯誤。因此本研究使用基於深度學習的影像匹配 DFM 技術，以提高匹配的準確性和可靠性。輸入兩張影像分別做為目標影像與待匹配影像，經過 DFM 演算法計算並後處理，得到兩張影像的共軛點影像座標。

DFM (Efe *et al.*, 2021) 演算法使用 VGG19 (Simonyan & Zisserman, 2014) 預訓練模型，VGG19 是一種深度卷積神經網路 (Very Deep Convolutional Networks)，該模型使用 ImageNet 訓練資料集訓練而得，含有 1.44 億個參數，VGG19 的可訓練層 (卷積層和全連接層) 總數為 19 層，經由多個卷積層的堆疊使網路能夠學習到更豐富的影像特徵，因此 VGG19 的卷積層可作為其他深度學習或影像處理的特徵萃取器。由於 VGG19 已從 ImageNet 1400 萬張影像學習影像特徵萃取，故 DFM 直接採用 VGG19 預訓練模型萃取影像深度特徵，優點是不需要對輸入影像進行訓練，同時，VGG19 架構可萃取不同尺度下的影像特徵，以利由粗到細 (Coarse-to-fine) 的影像匹配架構。

3.2 深度特徵匹配

在萃取出深度特徵後，DFM 演算法使用 DNNS (Dense Nearest Neighbor Search) 進行特徵匹配，DNNS 在目標影像及待匹配影像的特徵圖中，以相互最近鄰搜尋法進行特徵匹配，透過計算向量之間的 L2-norm 歐幾里得距離 (Euclidean distance) 來衡量相似性，若兩特徵圖中的某一對匹配點滿足給定的 L2-norm 距離門檻且為雙向滿足條件，則視為成功匹配。

由於僅依賴最近鄰可能會導致錯誤匹配，為了提高匹配的準確性，進一步同時使用比率檢測 (ratio test) (Lowe, 2004) 篩選匹配對。比率檢測的步驟如下：

(1) 對於目標影像中的每個深度特徵，找到待匹配影像中最近的鄰近(d_1)和次近的鄰近(d_2)距離。

(2) 檢查(d_1/d_2)是否小於設定的門檻值 t (例如 t=0.75)。

(3) 只有當(d_1/d_2) 小於該閾值時，才認為匹配是可靠的。

由於匹配的點中仍可能包含錯誤點(Outliers)，需要進一步使用隨機採樣一致性算法 (Random Sample Consensus, RANSAC) 過濾錯誤點，研究中以 Homography matrix 建立初始共軛點間的轉換，通過隨機選取一部分匹配點，估計這個轉換矩陣；再檢查其餘的匹配點是否滿足該轉換矩陣，計算殘差；若殘差大於門檻值 (如 3 pixels)，不符合轉換的點被視為錯誤點，經由迭代計算統計可能的錯誤點，以找到最優解。

3.3 兩階段匹配策略

DFM 為兩階段式 (Stage-0 及 Stage-1) 深度學習式匹配技術，此架構中先以 VGG19 預訓練模型萃取不同尺度的影像深度特徵，再以由粗到細 (Coarse-to-Fine) 策略進行精密影像匹配。在第一階段 (Stage-0) 中，首先會進行一次深度特徵萃取，然後以 DNNS 進行特徵匹配。匹配結果將用於粗略估計待匹配影像及目標影像間的幾何轉換，如此可將待匹配影像進行初步幾何轉換，以提高兩張輸入影像的相似度，以利後續精密匹配，如圖 3 所示。

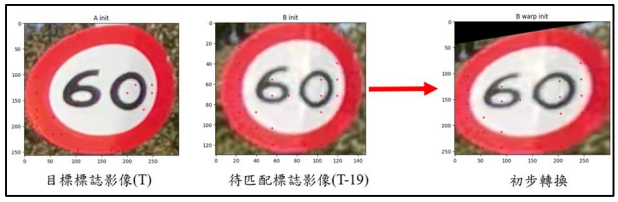


圖 3 Stage-0 初步匹配轉換

在第二階段 (Stage-1) 中，模型將分別萃取四個不同尺度 (Levels 3,2,1,0) 的深度特徵 (圖 4) 的深度特徵圖 (Dense Feature Maps)，並由粗至細逐層進行 DNNS 匹配，由粗至細縮小搜尋範圍，精化匹配成果。四個尺度下的匹配成果如圖 5 所示，從 Level 3 至 Level 0 的匹配成果可以發現，隨著階層的提升匹配的數量減少，但其精密度較佳。

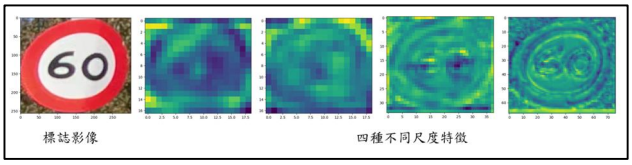


圖 4 Stage-1 初步匹配轉換

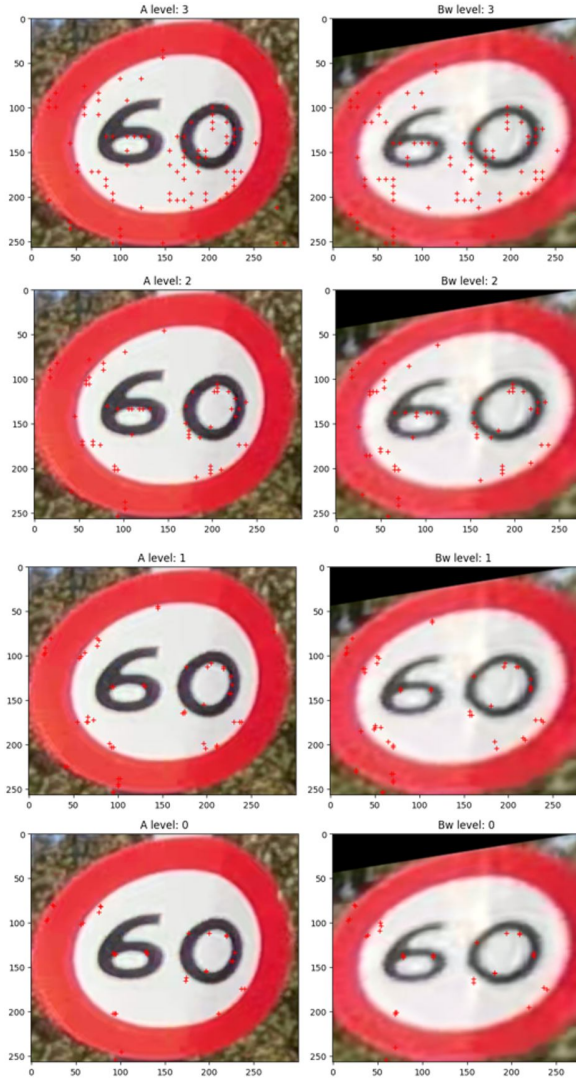
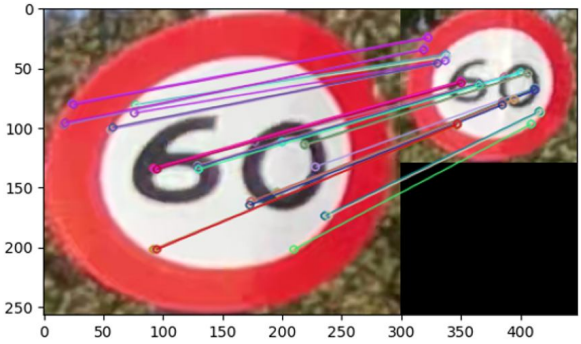
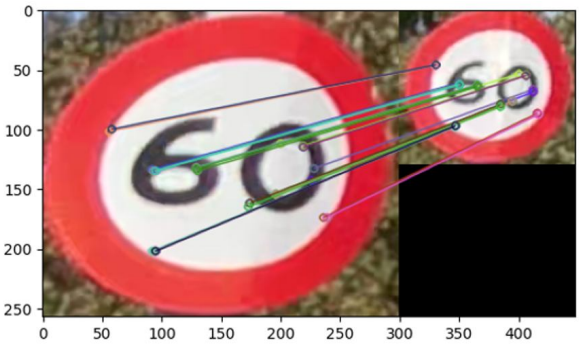


圖 5 由上到下展示四種尺度下特徵匹配成果，左圖為匹配目標影像，右圖為待匹配影像經過 Stage-0 初步轉換後的影像，紅點為兩幅影像的匹配點位置

最後，在完成 DFM 匹配完成後，再利用隨機抽樣一致 (Random Sample Consensus，RANSAC) 演算法，以幾何約制過濾錯誤匹配點。整合四種尺度下的匹配成果後，可得到圖 6(a) 的匹配，此時的匹配多數正確，但為了進一步提高精確度，使用 RANSAC 進行過濾幾何一致性較差的點位，成果如圖 6(b) 所示。



(a) 幾何約制前共有 33 組匹配成功圖



(b) 幾何約制後共有 22 組匹配成功圖

圖 6 比較 RANSAC 過濾錯誤匹配點

4. 實驗成果

本研究應用深度特徵於多視角車載影像匹配作業，實驗分析包含兩個部份，第一部份比較 DFM 深度特徵與 SIFT、SURF、ORB 特徵於影像匹配的表現；第二部份則針對研究區域的交通標誌進行匹配及定位，以標誌的定位成功率及誤差進行分析。

4.1 匹配方法比較

為了探討深度特徵影像匹配的成效，本研究設計三種多視角車載影像差異情境以比較 DFM、SIFT、SURF 及 ORB 方法在不同影像條件下的匹配效能。為了確保模型匹配對比的一致性，四種方法的匹配

參數和幾何約制參數均設置相同，唯影像匹配的特徵不同。影像匹配結果應用於建立六參數轉換，再用於產生套合影像。為了量化評估影像匹配的精度，採用人工量測的真實共軛點分析六參數轉換的精度。這三種影像差異情境如下：

Case 1 (Scale)：影像間存在覆蓋縮放差異，不同影像之間的覆蓋縮放比例差異對影像匹配的對應能力進行分析。

Case 2 (Distortion)：影像間存在形變差異，分析當影像在形變時的匹配效果。

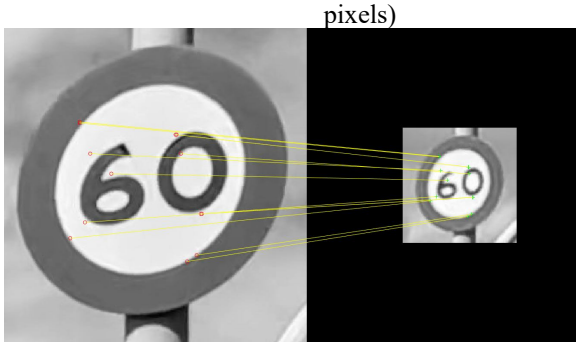
Case 3 (Occlusion)：影像間存在部分遮蔽，分析在一張影像被遮蔽一部分的情境下，不同匹配能力的表現。

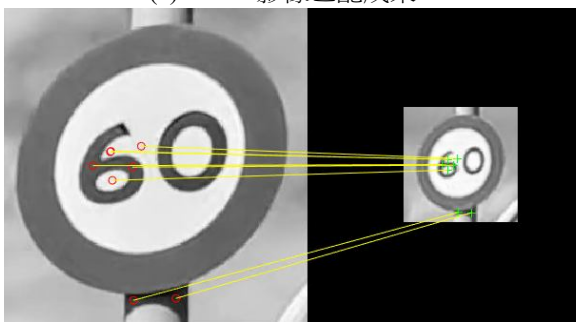
Case 1 尺度差異的影像匹配成果比較如圖 7 所示，影像 1 的空間解析度較高，而影像 2 的空間解析度僅有影像 1 的 1/3，故實驗中兩張輸入影像的尺度差異大約 3 倍。圖 7(c) 是使用 DFM 深度特徵的匹配成果，使用 DFM 的匹配點對影像 2 進行幾何校正得到套合後影像，將影像 1 及套合影像 2 疊合並分別使用紅綠波段進行展示，圖 7(g) 呈現以深度特徵進行匹配的影像套合成果具有高一致性。SIFT、SURF 及 ORB 的影像匹配成果如圖 7(d)~(f) 所示。與 DFM 影像匹配成果相比，DFM 的匹配成功點數較 SIFT 及 SURF 多。SIFT、SURF 及 ORB 幾何校正之套合後影像如圖 7(h)~(j) 所示，四種特徵匹方法都能克服尺度的差異，但由於 DFM 的匹配成功的可靠點數較多，有較多可靠的套合點可應用在建立轉換參數，故 DFM 呈現的匹配及套合成果較其他三種方法佳。

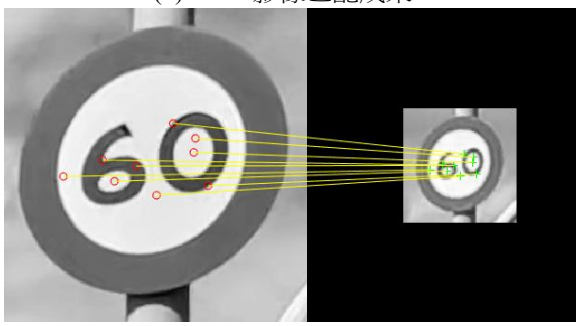
Case 1 尺度差異的影像匹配量化成果整理如表 1 所示，使用匹配點的六參數轉換均方根誤差在 1.1 pixels 左右，使用人工量測 5 個共軛點進行精度評估，DFM 的均方根誤差較低，代表精度略高於 SIFT 演算法。比較 DFM 及 SURF、ORB 的精度，SURF 及 ORB 檢核點的均方根誤差在 6.4~6.6 pixels，在尺度差異的情況下，SURF 及 ORB 有較佳的表現，表示在尺度差異的情況下，四種方法有良好的表現。

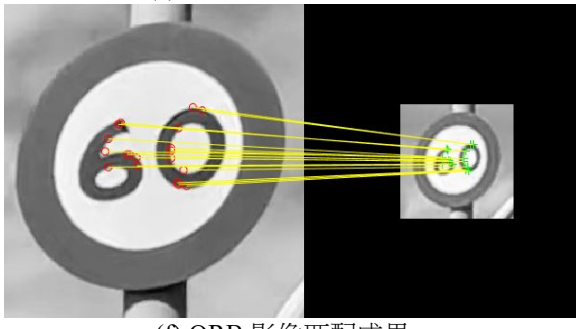

(a) 影像1 (ID 3876FR2, 300×289 pixels)

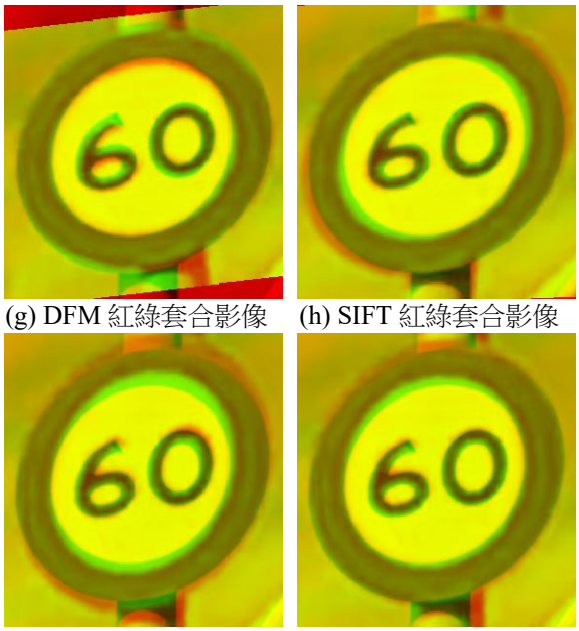

(b) 影像 2 (ID 3833MLA, 109 × 108 pixels)


(c) DFM 影像匹配成果


(d) SIFT 影像匹配成果


(e) SURF 影像匹配成果


(f) ORB 影像匹配成果



(g) DFM 紅綠套合影像 (h) SIFT 紅綠套合影像

(i) SURF 紅綠套合影像 (j) ORB 紅綠套合影像

圖 7 不同匹配方法之比較 (Case 1)

表 1 精度評估比較

		DFM	SIFT	SURF	ORB
Case1 尺度差異	匹配成功點	10	7	9	19
	匹配點 MeanError(pixel)	1.041	1.022	2.366	2.216
	匹配點 RMSE(pixel)	1.127	1.104	2.624	2.878
	人工檢核點	5	5	5	5
	檢核點 MeanError(pixel)	6.691	7.660	5.724	5.573
	檢核點 RMSE(pixel)	7.620	8.116	6.695	6.475
Case2 形變差異	匹配成功點	281	9	6	12
	匹配點 MeanError(pixel)	3.333	1.743	0.578	0.609
	匹配點 RMSE(pixel)	3.992	2.005	0.684	0.676
	人工檢核點	5	5	5	5
	檢核點 MeanError(pixel)	2.231	3.488	2.312	3.556
	檢核點 RMSE(pixel)	2.392	3.882	2.655	4.563
Case3 遮蔽差異	匹配成功點	67	5	13	8
	匹配點 MeanError(pixel)	4.338	1.693	3.916	0.834
	匹配點 RMSE(pixel)	4.885	2.370	4.643	0.958
	人工檢核點	4	4	4	4
	檢核點 MeanError(pixel)	5.304	9.163	3.344	2.632
	檢核點 RMSE(pixel)	5.919	11.249	4.120	2.942

Case 2 形變差異的影像匹配成果比較如圖 8 所示，此形變主要是由影像視角及畸變差造成，影像內容為一圓形交通標誌，影像 1 的形變較低仍呈現圓形，而影像 2 中的圓形變形為橢圓形，故實驗中兩張輸入影像存在相對較大的形變。圖 8(c)~(f)分別是四種方法的匹配成果，實驗結果顯示，DFM 提取的深度特徵能成功匹配形變差異較大的影像，而其他三種方法萃取的特徵在匹配的成效較差，DFM 匹配成功的點數遠多於其他三種方法。四種方法個別進行幾何轉換，得到的套合影像如圖 8(g)~(j) 所示，DFM 則能達成較佳的成果，其他三種方法之套合影像的偏差量明顯較大。經由此實驗可知，雖然其他三種方法能應對尺度差異，但在面對形變差異時具有局限性。針對車載廣角影像的交通標誌的局部特性，使用深度特徵可得到更佳的匹配成果。

Case 2 形變差異的影像匹配量化成果整理如表 1 所示，此案例，DFM 的匹配成功點數達高於其他三種方法，有利於前方交會產生三維點雲，使用匹配點的六參數轉換均方根誤差在 4 pixels 左右；SURF 及 ORF 的匹配點均方根誤差小於 1 pixel，代表匹配成果分佈較為集中，不利後續套合轉換。使用人工量測 5 的個共軛點進分精度評估，DFM 的均方根誤差約為 2.4 pixels，較其他方法有更佳的幾何轉換精度，表示 DFM 的量化精度較佳。

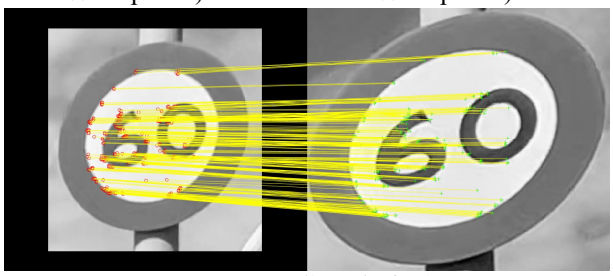
Case 3 使用具有部份遮蔽的影像對進行匹配，成果比較如圖 9 所示，影像 1 呈現一完整的圓形交通標誌，而影像 2 中的交通標誌有一半被切除，完整性較低。圖 9(c)~(f)是四種方法的匹配成果，此實驗成果顯示，DFM 深度特徵在完整性較低的情況下仍能達成成功匹配，而其他方法萃取的特徵在匹配的成效較差，DFM 匹配成功的點數遠多於其他方法。可能的原因是 SIFT 需要計算影像金字塔的 DoG，在影像維度較小時，鄰近邊界區域的特徵較少，故無法匹配。四種方法個別進行幾何轉換，得到的套合影像如圖 9(h)~(j)所示，DFM 則能達成較佳的成果，其他方法受套合點數數量之影響，套合影像有明顯偏差。經由此實驗可知，匹配目標完整性較低的區域，深度特徵有更好的匹配表現。

Case 3 形變差異的影像匹配量化成果整理如表 1 所示，此案例中，DFM 的匹配成功點數比較多，基於匹配點的六參數轉換，均方根誤差約為 4 pixels 左右，使用人工量測 4 的個共軛點進分精度評估，DFM 的均方根誤差約為 5.9 pixels，而 SIFT

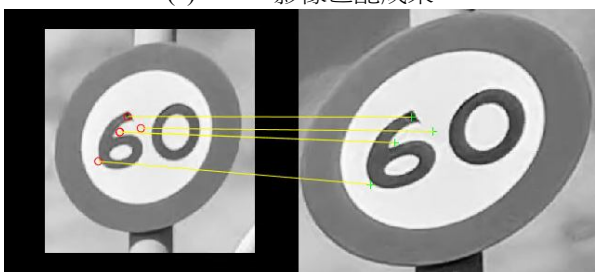
均方根誤差為 11 pixels 較大，表示 DFM 的量化精度較 SIFT 佳。SURF 及 ORB 有較低的均方根誤差，主要原因是人工可量測的 4 個共軛點是坐落在交通標誌的數字 6 的角點上，SURF 及 ORB 僅有成功匹配到數字 6 的匹配點，故轉換誤差較小，但套合影像成果有明顯偏差。

(a) 影像1 (ID 3876FR2, 300×289 pixels)

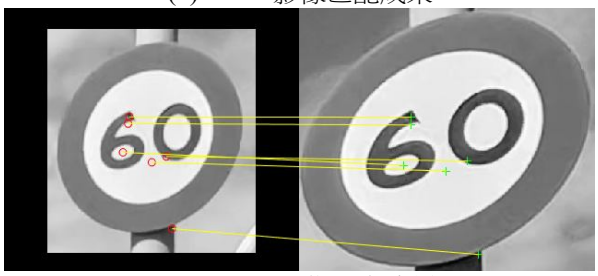
(b) 影像2 (ID 3879MLA, 357×411 pixels)



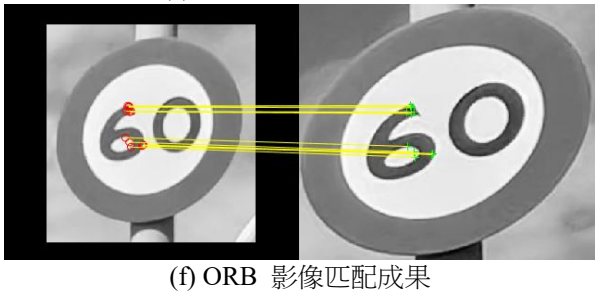
(c) DFM 影像匹配成果



(d) SIFT 影像匹配成果



(e) SURF 影像匹配成果



(f) ORB 影像匹配成果

(g) DFM 紅綠套合影像

(h) SIFT 紅綠套合影像

(i) SURF 紅綠套合影像

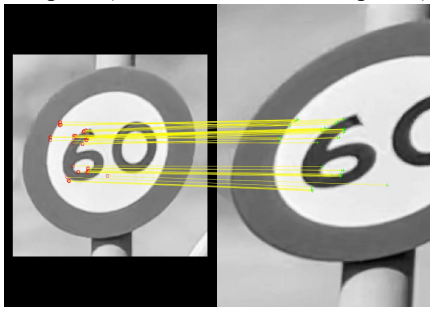
(j) ORB 紅綠套合影像

圖 8 不同匹配方法之比較 (Case 2)

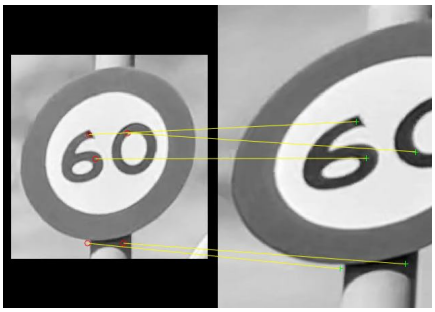
此三組案例中使用 SIFT、SURF、ORB 尋找到的匹配共軛點數量與 DFM 相比差距甚大並且有分佈不均勻的現象。在影像匹配任務中若只是需要判定兩張影像存在同一物件，四種演算法都能做到；若是利用匹配共軛點進行幾何轉換甚至進行前方交會定位則會須要考慮點位分佈與數量，DFM 使用深度特徵進行匹配具有優勢。

(a) 影像1 (ID 3876FR2, 300×289 pixels)

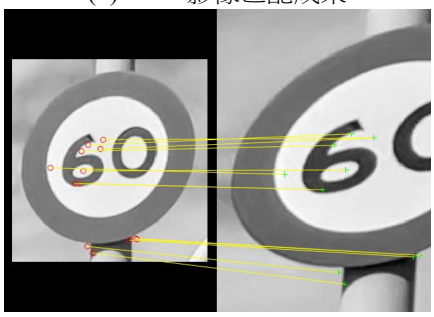
(b) 影像2(ID 3876FL1, 453×318 pixels)



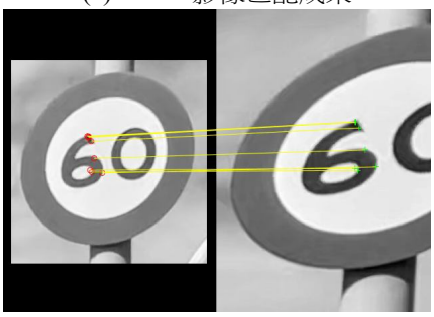
(c) DFM 影像匹配成果



(d) SIFT 影像匹配成果



(e) SURF 影像匹配成果



(f) ORB 影像匹配成果

(g) DFM 紅綠套合影像

(h)SIFT 紅綠套合影像

(i)SURF 紅綠套合影像

(j) ORB 紅綠套合影像

圖 9 不同匹配方法之比較 (Case 3)

4.2 定位成果分析

使用車載多視角影像進行交通標誌的挑戰是多視角影像的短基線僅有 0.88 m，影像匹配的誤差，加上短基線會造成三維定位在軌跡行徑方向有弱交會幾何的問題，需要仰賴大量的影像重複觀測交通標誌，經由統計分析得到交通標誌的位置，因此可經由交通標誌定位分析影像匹配之效益。

本項實驗鎖定交通標誌進行定位分析，研究中使用 Mapillary Vistas Dataset (Neuhold *et al.*, 2017) 交通標誌資料集訓練完成之 YoLo v4 深度學習模型 (Wang *et al.*, 2020) 進行交通標誌偵測 (Chen & Teo, 2019)，僅針對交通標誌區域進行深度特徵的匹配，取得共軛點後，配合影像內外方位參數進行前方交會取得物空間三維坐標點雲，最後以 DBScan 對物空間相鄰的點雲進行聚類分析，以聚類成果的中心點代表交通標誌三維坐標位置。

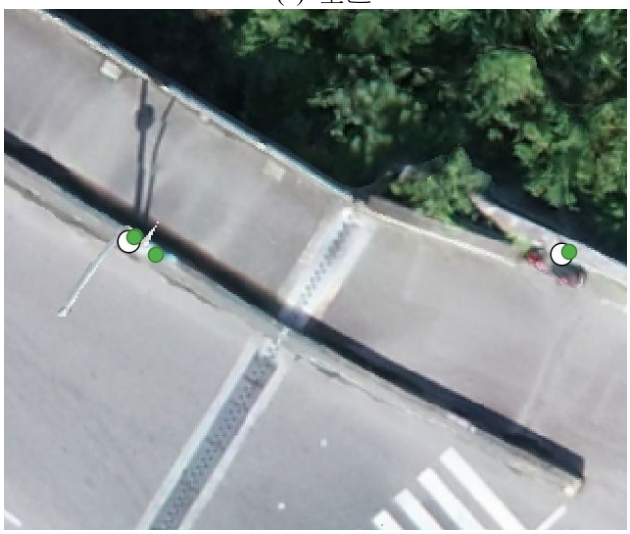
精度分析使用的參考資料為人工量測地面光達點雲的交通標誌坐標，此路段人工量測得到 26 個可視的交通標誌。交通標誌匹配及定位共得到 72 個交通標誌位置，由於匹配及前方交會帶有誤差，造成真實世界的一個交通標誌可能被重複定位，因此匹配及定位的交通標誌數量會比人工量測的數量多。

精度分析以地真資料為基準，以地真位置為圓心，半徑 1 m 內是否有成功被偵測的點位，圖 10(a) 套疊人工量測交通標誌位置及自動化定位點，其中，白色是人工量測點，綠色是偵測及定位正確，紅色是定位錯誤。圖 10(b) 局部放大以說明重複定位的問題，左側一個標誌(白色)被重複定位兩次(綠點)。實驗顯示，在 1 m 範圍內有成功偵測資料的點數量為 49 點，顯示使用車載影像進行標誌定位成功率約 68%(＝49/72)；其餘 23 個點中，有 17 點為錯誤偵測及 6 點為錯誤定位，錯誤偵測是指所得到的定位點屬是條紋色護欄或三角錐，而錯誤定位是指有偵測成功，但影像匹配及交會之標誌定位失敗。

最後針對正確定位的點位進行定位誤差分析，統計這 49 個點與人工量測點的誤差，得到距離平均誤差為 0.395 m，顯示正確偵測與定位的點位可達到 0.5 m 內的定位精度。圖 11 展示交通標誌定位案例，影像匹配與定位的點雲與地面光達點雲具有高的一致性；圖 12 展示兩個錯誤案例，分別是紋色

護欄及三角錐被錯誤偵測為交通標誌。

(a) 全區



(b) 局部放大：重複定位

圖 10 比較人工數化及自動化定位點位：人工數化點 (白色圓形)、正確定位 (綠色圓形)、錯誤定位點 (紅色圓形)

5. 結論與未來工作

本研究採用卷積神經網路 (CNN) 的預訓練模型來提取深度特徵，並利用這些特徵進行影像匹配。有別於傳統影像匹配採用人工設計的 handcrafted features 進行匹配作業，深度學習採用卷積神經網路自行萃取深度特徵，這些深度特徵具有更好的泛用性，能有效應對立體影像間的尺度和形變差異，達成可靠的影像匹配。

研究中針對多視角車載影像進行實驗分析，由於車載廣角影像有較大的形變，採用深度特徵進行匹配可獲取物空間三維點雲，以達成物空間三維定位之目的。實驗中分別針對交通標誌的尺度差異、形變差異及完整性進行比較分析，實驗證明，與傳統演算法相比，DFM 演算法具有更多的成功匹配點。此外，並針對橋樑路段之三重疊影像進行交通標誌定位，交通標誌定位成功率可接近 70%，且成功定位點之絕對誤差小於 0.5 m。

(a) 交通標誌

(b) 影像定位成果套疊光達點雲

圖 11 交通標誌定位成功偵測案例

(a) 錯誤偵測 (條紋上色護欄)

(b) 錯誤偵測 (三角錐)

圖 12 交通標誌偵測錯誤

未來研究將進一步探索影像特徵的選擇，探索適合特定場景的多尺度深度特徵提取方法，並優化匹配策略以進一步提高匹配成功率。此外，考慮在匹配過程中使用對比學習 (Contrastive Learning) 和強化學習 (Reinforcement Learning) 技術，以自動優化匹配過程中的關鍵參數，減少手動調參的成本，並進一步提升影像匹配的效率和準確性。為精進交通標誌的偵測及定位，在影像匹配過程中融入交通標誌類別屬性，以提升交通標誌定位的成功率及精度。

參考文獻

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L., 2008. Speeded-up robust features (SURF), Computer Vision and Image Understanding, 110(3): 346-359, DOI: 10.1016/j.cviu.2007.09.014.

Bökman, G., and Kahl, F., 2022. A case for using rotation invariant features in state of the art feature matchers, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, pp.5110–5119, DOI: 10.1109/CVPRW56347.2022.00559.

Chen, P.C., and Teo, T.A., 2019. Mapping traffic signboards from mobile mapping systems using a deep learning approach, in Proceedings of the 40th Asian Conference on Remote Sensing, October 14–18, Daejeon, Korea.

DeTone, D., Malisiewicz, T., and Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPRW), Salt Lake City, UT, USA, pp.337-349, DOI: 10.1109/CVPRW.2018.00060.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., and Sattler, T., 2019. D2-net: A trainable cnn for joint description and detection of local features, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, pp.8084-8093, DOI: 10.1109/CVPR.2019.00828.

Efe, U., Ince, K.G., and Alatan, A., 2021. DFM: A performance baseline for deep feature matching, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 4279–4288, DOI: 10.1109/CVPRW53098.2021.00484.

He, K., Zhang, X., Ren, S., and Sun, J., 2016. Deep residual learning for image recognition, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp.770–778, DOI: 10.1109/CVPR.2016.90.

Hughes, L.H., Schmitt, M., Mou, L., Wang, Y., and Zhu, X.X., 2018. Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN, IEEE Geoscience and Remote Sensing Letters, 15(5): 784–788, DOI: 10.1109/LGRS.2018.2799232.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, 60(2): 91–110, DOI: 10.1023/B:VISI.0000029664.99615.94.

Melekhov, I., Kannala, J., and Rahtu, E., 2016. Siamese network features for image matching, in Proceedings of the 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, pp.378–383, IEEE, DOI: 10.1109/ICPR.2016.7899663.

Merkle, N., Auer, S., Mueller, R., and Reinartz, P., 2018. Exploring the potential of conditional adversarial networks for optical and SAR image matching, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(6) : 1811-1820, DOI: 10.1109/JSTARS.2018.2803212.

Morelli, L., Ioli, F., Maiwald, F., Mazzacca, G., Menna, F., and Remondino, F., 2024. Deep-image-matching: A toolbox for multiview image matching of complex scenarios, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 48(2/W4): 309–316, DOI: 10.5194/isprs-archives-XLVIII-2-W4-2024-309-2024.

Neuhold, G., Ollmann, T., Rota Bulo, S., and Kontschieder, P., 2017. The Mapillary vistas dataset for semantic understanding of street scenes, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp.5000–5009, DOI: 10.1109/ICCV.2017.534.

Ren, F., Huang, J., Jiang, R., and Klette, R., 2009. General traffic sign recognition by feature matching, in Proceedings of the 24th International Conference Image and Vision Computing New Zealand, Wellington, New Zealand, pp. 409-414, DOI: 10.1109/IVCNZ.2009.5378370.

Rocco, I., Cimpoi, M., Arandjelović, R., Torii, A., Pajdla, T., and Sivic, J., 2018. Neighbourhood consensus networks, in Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18), Montréal Canada, Curran Associates Inc., Red Hook, NY, USA, pp.1658–1669.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G., 2011. ORB: An efficient alternative to SIFT or SURF, in Proceedings of the International Conference on Computer Vision (ICCV), Barcelona, Spain, pp.2564–2571, DOI: 10.1109/ICCV.2011.6126544.

Simonyan, K., and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, DOI: 10.48550/arXiv.1409.1556.

Sun, J., Shen, Z., Wang, Y., Bao, H., and Zhou, X., 2021. LoFTR: Detector-free local feature matching with transformers, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp.8918-8927, DOI: 10.1109/CVPR46437.2021.00881.

Teo, T.A., 2015. Video-based point cloud generation using multiple action cameras, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 40(4/W5): 55–60, DOI: 10.5194/isprsarchives-XL-4-W5-55-2015.

Wang, C.Y., Bochkovskiy, A., and Liao, H.Y.M., 2020. Scaled-YOLOv4: Scaling cross-stage partial network, arXiv Preprint, arXiv:2011.08036, DOI: 10.48550/arXiv.2011.08036.

Xu, S., Chen, S., Xu, R., Wang, C., Lu, P., and Guo, L., 2024. Local feature matching using deep learning: A survey, Information Fusion, 107: 102344, DOI: 10.1016/j.inffus.2024.102344.

Ye, Y., Tang, T., Zhu, B., Yang, C., Li, B., and Hao, S., 2022. A multiscale framework with unsupervised learning for remote sensing image registration, IEEE Transactions on Geoscience and Remote Sensing, 60: 5622215, DOI: 10.1109/TGRS.2022.3167644.

Zagoruyko, S., and Komodakis, N., 2015. Learning to compare image patches via convolutional neural networks, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp.4353–4361, DOI: 10.1109/CVPR.2015.7299064.

Zitová, B., and Flusser, J., 2003. Image registration methods: A survey, Image and Vision Computing, 21(11): 977–1000, DOI: 10.1016/S0262-8856(03)00137-9.

Deep Feature Matching for Multi-View Images in Mobile Mapping Systems

Tee-Ann Teo [1*] Pei-Cheng Chen [2] Ting-Ni Chen [3] Hsuan Yang [4]
Zhan-Qing Lin [4] Kuan-Yi Lee [4] Kai-Chieh Hung [4] Chen-Yung Lu [4]

Abstract

With the advancement of mobile mapping technology, multi-view imagery has become an important source for road observation. However, traditional matching methods struggle to overcome the challenges posed by distortions and variations in viewing angles between images. To enhance the precision and reliability of image matching, this study investigates a deep feature matching technique based on deep learning. By utilizing convolutional neural networks (CNN) to extract deep features, accurate matching and 3D spatial positioning of multi-view images can be achieved. The study employs Deep Feature Matching (DFM) technology, which is based on the pre-trained VGG19 model. Through a two-stage matching strategy and the RANSAC (Random Sample Consensus) algorithm, erroneous matching points are filtered out to ensure the reliability of the matching results. The experimental data consists of multi-view images, with traffic signs serving as the target objects for image matching and positioning. The research results reveal that, compared with the traditional SIFT method, DFM demonstrates a higher success rate and improved positioning accuracy in various image scenarios, including scale differences, shape distortions, and occlusion conditions. Notably, DFM achieves significantly more matching points than SIFT in distortion and occlusion scenarios. Furthermore, the analysis of traffic sign positioning indicates that the success rate of traffic sign positioning reaches 70%, with an average error of less than 0.5 m for successfully located points. This finding highlights the practical application potential of DFM in 3D positioning using multi-view images in complex scenarios and confirms that it achieves higher success rates and accuracy.

Keywords: Deep Learning, Deep Features, Image Matching, Multi-View Images, Traffic Signboard

[1] Professor, Department of Civil Engineering, National Yang Ming Chiao Tung University
[2] Ph.D Candidate, Department of Civil Engineering, National Yang Ming Chiao Tung University
[3] Master Student, Department of Civil Engineering, National Yang Ming Chiao Tung University
[4] Engineer, Department of Geomatics, CECI Engineering Consultants, Inc., Taiwan
[*] Corresponding Author, E-mail: tateo@nycu.edu.tw

Received Date: Jan. 10, 2025
Revised Date: Apr. 23, 2025
Accepted Date: May 09, 2025