Journal of Sensor Science and Technology
Vol. 33, No. 2 (2024) pp. 78-85
http://dx.doi.org/10.46670/JSST.2024.33.2.78
pISSN 1225-5475/eISSN 2093-7563

영상 기반 위치 인식을 위한 대규모 언어-이미지 모델 기반의 Bag-of-Objects 표현

정승운[1] · 박병재[1,+]

Large-scale Language-image Model-based Bag-of-Objects Extraction for Visual Place Recognition

Seung Won Jung[1] and Byungjae Park[1,+]

Abstract

We proposed a method for visual place recognition that represents images using objects as visual words. Visual words represent the various objects present in urban environments. To detect various objects within the images, we implemented and used a zero-shot detector based on a large-scale image language model. This zero-shot detector enables the detection of various objects in urban environments without additional training. In the process of creating histograms using the proposed method, frequency-based weighting was applied to consider the importance of each object. Through experiments with open datasets, the potential of the proposed method was demonstrated by comparing it with another method, even in situations involving environmental or viewpoint changes.

Keywords: Visual place recognition, Bag-of-objects, TF-IDF, Multi modal AI, Visual localization

1. 서 론

VPR (Visual Place Recognition)은 센서를 통해 획득된 현재 이미지와 이전에 저장된 이미지 matching을 통해 현재의 위치를 이전에 관찰된 위치와 일치하는지를 판단하는 작업이다. VPR은 로봇의 SLAM (Simultaneous Localization and Mapping)에서 loop closure detection이나 pose estimation을 위한 후보 이미지 matching과 같은 분야에서 중요한 요소로 작용한다.

전통적인 VPR 기법은 이미지 내에서 고유하게 식별될 수 있는 keypoint 주변의 맥락 정보를 수치적으로 표현하는 descriptor에 주로 기반하고 있다. 이러한 descriptor를 서로 다른 이미지에서 비교함으로써 유사도를 평가하며, 이를 통해 위치 인식 작업을 수행한다. 기본적으로 descriptor의 종류는 local descriptor와 global descriptor로 나뉘게 된다. Local descriptor는 특정 영역의 주요 특징을 추출함으로써 다양한 환경과 조건에서 이미지의 높은 차원의 특징을 안정적으로 검출할 수 있는 알고리즘이다. 이러한 local descriptor는 특정 지역의 특징만을 고려하기 때문에 전체적인 이미지를 파악하기 어려울 수 있고, 특정 패턴이나 객체에 대한 일반화가 제한적일 수 있다는 한계가 존재한다.

Global descriptor는 이미지나 영상의 전체적인 특징을 하나의 고차원 벡터로 효율적으로 표현하는 기법이다. 하지만 global descriptor는 전체 이미지 정보를 하나의 벡터로 압축하기 때문에 큰 dataset에서의 matching 성능이 떨어질 수 있고, 다양한 객체나 배경에 대한 세부 정보 손실이 발생할 수 있다는 한계가 존재한다.

이러한 한계를 극복하고자 두 descriptor들의 중간 성격을 가진 이미지에서 중요하고 식별력 있는 이미지 조각 단위인 패치들을 기준으로 VPR를 수행하는 방법들이 제안되었다 [1].

본 논문은 RGB 이미지 센서로부터 들어오는 이미지 내의 추출된 물체들을 중심으로 통계적 처리 방법을 도입하여 물체 각각에 가중치를 할당함으로써 VPR의 효율성을 높이는 가능성을 제시하고 대표적인 VPR 모델인 VLAD와 정성적 비교 평가를 수행한다.

2. 관련 연구

2.1 Visual Place Recognition

VPR 작업은 입력 영상인 query 이미지와 같은 위치에서 찍은 이미지를 추출하는 것을 목표로 한다. 이러한 VPR 작업은

[1]한국기술교육대학교 기계공학부 (School of Mechanical Engineering, Korea University of Technology and Education)
1600, Chungjeol-ro, Byeongcheon-myeon, Dongnam-gu, Cheonan-si, Chungcheongnam-do, 31253, Korea
[+]Corresponding author: bjp@koreatech.ac.kr
(Received: Feb. 16, 2024, Revised: Mar. 4, 2024, Accepted: Mar. 12, 2024)

컴퓨터 비전의 이미지 유사도 검색(image retrieval) 작업과 굉장히 비슷하지만 이미지 유사도 검색 작업은 단순히 query 이미지와 유사한 이미지를 추출하는 것에 목적을 두는 것에 비해 VPR은 특정 장소와 위치를 인식하는 목적으로 작업을 수행하는데 차이가 존재한다. 앞서 언급대로 VPR 작업에 있어 특징을 추출하는데 크게 두 가지 방법이 존재한다. 첫 번째로는 SIFT (Scale-Invariant Feature Transforms) [2], SURF (Speeded-Up Robust Features) [3]와 같은 local descriptor를 추출하는 방식이다. Local descriptor를 계산하기 위해서는 이미지의 keypoint를 추출하는 작업이 선행되어야 한다. 이러한 local descriptor는 이미지의 회전, 크기 변화, 일반적인 변형 등에 대한 강인하다는 것이 장점으로 꼽힌다. 하지만 local descriptor는 한 이미지에서 최소 몇 백 개의 keypoint를 추출하기 때문에 바로 이미지 matching을 진행하는 것은 컴퓨팅 자원과 효율성이 떨어질 수 있다. 이러한 단점을 극복하고자 이미지를 하나의 고정된 벡터로 표현하는 global descriptor인 Bag-of-Words(BoW) [4]와 같은 모델이 소개되었다. BoW는 keypoint를 추출하고 추출된 keypoint들에 대해 clustering을 수행하여 cluster들의 중심인 visual word를 찾아낸다. 이렇게 찾은 visual word들을 histogram으로 표현하여 visual dictionary를 생성한다. 따라서 이미지 하나 당 하나의 histogram이 생성되며 visual word의 출현 빈도를 기반으로 이미지 유사도를 비교하는 작업을 수행할 수 있다. 이렇게 표현된 visual dictionary는 이미지의 복잡한 구조를 상대적으로 간단한 형태로 표현할 수 있어 local descriptor들의 유사도를 직접 비교하는 것 보다 계산 효율성이 높다. 이 밖에도 밝기 histogram, 텍스처 특징으로부터 같은 global descriptor를 추출하거나 local descriptor를 clustering하여 global descriptor로 만드는 VLAD [5]나 VLAD의 기법을 CNN 네트워크로 대체해 global descriptor로 추출하는 NetVLAD [6]와 같은 방법들이 제안되었다.

2.2 Bag-of-Objects

많은 연구에서 BoW의 계산 효율성을 활용한 VPR 방법이 제안되었다 [7,8]. 하지만 BoW은 visual word 그리고 visual dictionary가 이미지 촬영 시 조도나 밝기와 같은 환경에 대해서 강건하지 않다는 한계가 있다. 이를 극복하고자 BoW의 변형 모델인 Bag-of-Objects(BoO)와 같은 모델들이 위치 추정 방식에 적용되고 있다 [9]. BoO 모델은 이미지의 픽셀 단위가 아닌 패치 단위로 특징이 되는 물체들을 식별하여 위치 추정 방식에 적용하는 기법이다. BoO 모델은 BoW의 visual word와 같은 역할인 visual object를 각 이미지에서 추출하여 장소 인식에 사용한다. Visual object를 추출하기 위해서 딥러닝 모델을 통해 각 이미지를 대표할 수 있는 물체 탐지 작업을 실행한다. 추출된 물체들을 histogram화하여 이미지를 표현하고 표현된 이미지를 벡터화하여 dataset 안에 있는 이미지와 cosine similarity나 euclidean distance 방식을 통해 query 이미지와 동일한 장소의 이미지를 추출해 낸다. 이러한 BoO 방식은 기존의 BoW 보다 조명이나 시점 차이 등에 더 강건하다는 장점이 있다.

Fig. 1. BoO example

2.3 Language-Image Model

컴퓨터 비전 분야에서는 전통적으로 이미지만을 데이터로 입력 받아 어떻게 모델을 구성하면 더 좋은 표현을 학습하는지에 대해 연구되어왔다. CNN(Convolutional Neural Network) 모델을 기반으로 한 Inception [9], Resnet [10] 모델들이 등장했고 Attention 모듈을 적용하는 SENet [11], BAM [12], CBAM [13]과 같은 모델들이 등장했다. 이와는 다르게 자연어 처리 분야는 2017년 긴 문장을 효과적으로 처리할 수 있는 Transformer [14]모델의 등장으로 초거대 언어 모델인 LLM(Large Language Model)을 향한 비약적인 발전을 이뤘으며, 비전 분야보다 한 발 앞서 나아가는 형태로 발전해왔다. 이러한 Trasformer의 등장으로 2021년 이후에도 비전 분야의 트렌드도 Transformer의 구조를 적용하는 것이었고, 이러한 트렌드에 맞추어 ImageGPT [15], Vision Transformer [16] 모델들이 등장하였다. 기존의 컴퓨터 비전 모델들은 labeling된 이미지 데이터만을 이용해 사전 학습(pre-training)을 진행하고 자신의 특정 작업에 맞는 dataset에 대해서 파인 튜닝(fine-tuning)하는 방식이였다. 하지만 이러한 방식은 학습에 사용되지 않은 이상치의 데이터가 들어왔을 때 예측률이 현저히 떨어지는 문제점이 존재해 일반화가 어렵다는 한계가 존재하고 무엇보다 labeling된 dataset의 용량에 대해서 한계가 존재했다. 이러한 한계를 극복하고자 컴퓨터 비전에서도 LLM과 같은 큰 모델과 큰 데이터를 사용한 CLIP [17]과 같은 연구들이 등장하였다. CLIP은 기존의 labeling된 dataset을 사용하지 않고 웹에서 존재하는 이미지와 이미지마다 달려있는 자연어 문장을 그대로 dataset으로 활용하여 4억장의 dataset을 구축하였다. 하지만 단일 유형의 데이터만 다루는 기존 모델과는 다르게 CLIP은 이미지와 자연어를 동시에 구축한 dataset을 사용하기 때문에 기존 모델과는 서로 다른 두 개의 데이터를 학습하는 인코더로 구성되고 이러한 모델을 멀티모달(Multi Modal) 모델이라 정의한다. 이러한 멀티모달 모델은 다양한 유형의 데이터를 동시에 처리하여 복잡하고 더욱 좋은 성능을

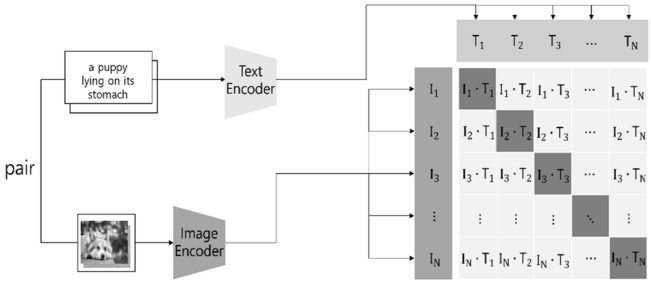

Fig. 2. CLIP structure visualization

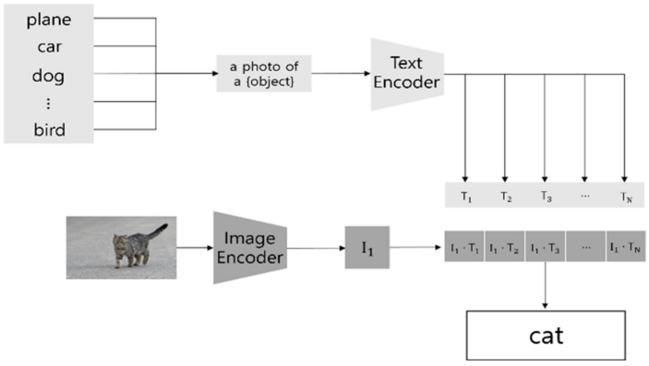

Fig. 3. Zero Shot Prediction

낼 수 있으며 다양한 유형의 데이터들을 함께 사용함으로써 상호보완적인 정보를 제공할 수 있으며, 전체 시스템의 성능을 향상시킬 수 있다. CLIP은 기존 labeling된 dataset에 대해서 학습하는 것이 아닌 이미지 마다 매칭되어 있는 문장으로 학습하기 때문에 기존 softmax로 구분하는 학습 방식은 적합하지 않다. 따라서 CLIP은 이미지와 텍스트 데이터를 대조하는 방식을 사용하여 학습을 진행하였고 대조 학습에서는 배치 단위로 이루어진 N개의 이미지와 텍스트를 각각 인코더에 통과시켜 각각의 특징 벡터 I_i와 T_i를 산출한다. 학습을 위한 정답 레이블로는 자기 자신의 레이블을 제외한 이미지 또는 텍스트는 서로 다르다는 관계, 즉 N 개의 쌍에서 서로 정답인 쌍(positive pair)은 N 개, 정답이 아닌 쌍(negative pair)은 N^2-N 개로 정의하여 진행한다. 이렇게 정답인 레이블 쌍에 대해서는 cosine similarity인 I_i * T_i가 최대가 되도록, 나머지 쌍에 대해서는 최소가 되도록 하는 방향으로 학습시킨다 [17].

학습된 CLIP은 기존의 ImageNet으로 학습된 모델에서는 기대하기 어려운 처음 보는 dataset, 즉 학습되지 않은 문제를 맞추는 Zero Shot Prediction(ZSP)에서 큰 강점을 보인다. ZSP를 수행하려면 먼저 이미지의 특징을 모델의 이미지 인코더를 통해 추출하게 된다. 그런 다음 이미지에 해당되는 클래스를 예측하기 위해서 사람이 직접 클래스 개수와 종류를 설정하고 이러한 텍스트 클래스에 대해서도 텍스트 인코더로 특징을 추출한다. 마지막으로 추출된 이미지 특징과 텍스트 특징들 간의 cosine similarity를 측정한 다음, 가장 높은 cosine similarity를 갖는 값을 해당 클래스로 예측하게 된다.

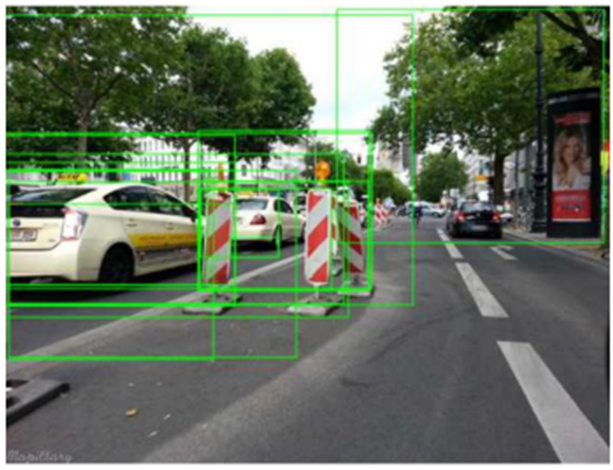

Fig. 4. Edge box example

3. 물체 탐지 및 선정

CLIP을 통해 이미지 내 물체 탐지 작업을 수행하기 위해서는 두 가지의 사전 작업이 필요하다. 첫 번째는 기존의 CLIP의 저자들이 진행했던 분류 문제와는 다르게 물체 탐지 작업을 진행하기 위해서는 이미지 내에서 물체가 있을만한 영역을 CLIP의 이미지 인코더에 넣어줘야 한다. 이러한 region proposal 방법은 이미 학습된 Edge box [18] 모델을 통하여 진행하였다. Edge box는 물체 탐지 작업에 있어 이미지 내에서 존재하는 물체의 위치를 나타내는 사각형인 bounding box를 생성하는데 사용되는 방법 중 하나이다. Edge box는 Structure edge detection을 사용하여 이미지의 구조적인 특징을 학습하고 edge를 검출한다. 이렇게 검출된 edge들의 집합인 edge group의 유사성을 고려하여 하나의 객체나 구조를 나타내는 정도를 파악하여 객체가 존재할 만한 영역을 제안한다.

두 번째로는 CLIP의 텍스트 인코더에 들어갈 부분으로 label을 지정해주는 작업이다. CLIP을 물체 탐지 모델로 선정한 것은 처음 보는 dataset에 대해서도 성능이 높기 때문이다. 하지만 위에서도 언급하였듯이 처음 보는 dataset에 대해서도 기준을 사람이 정해주는 작업이 선행되어야 한다. 본 연구에서는 labeling 작업을 prompt engineering을 통해 진행하였다. Prompt engineering이란 생성형 모델에 특정 작업에 대한 명확하고 효과적인 지시를 제공하여 원하는 결과를 얻기 위한 입력 문장을 고안하는 과정이다. 이러한 prompt engineering을 OpenAI사의 ChatGPT를 통하여 “여러 딥러닝 모델 dataset 중 도시 거리를 기반으로 한 label에 대해서 정의해줘”와 같은 질문으로 도시 환경에서 탐지될 수 있는 40가지의 label을 추출하였다.

Table 1. DF-IDF Value

Label	DF	IDF	Label	DF	IDF
Concrete Barriers	90	0.891818	Street lights	26	2.106841
Pedestrian Crosswalks	98	0.807558	Subway Entrances	17	2.512306
Traffic Cameras	58	1.325140	Roadside Mirrors	63	1.243794
Traffic Calming Device	168	0.272779	Traffic Cones	14	2.694627
Construction Barrier	36	1.791759	Graffiti	11	2.917771
Bus Shelters	125	0.566395	Parking Meters	16	2.569464
Emergency Call Boxes	29	2.001480	Bike Racks	21	2.311653
Pedestrian Guardrails	73	1.098612	Gas Stations	3	4.016383
Pedestrian Bridge	29	2.001480	Sewer Grates	2	4.304065
Street Signs	33	1.876317	Trash Bins	5	3.610918
Rail Barriers	83	0.971861	ATM Machine	12	2.837728
Bus Stops	57	1.342234	Speed Bumps	20	2.358155
Traffic Island	85	0.948330	Roadside Vegetation	22	2.267183
Curb Ramps	122	0.590493	Roadside Vegetation	2	4.304065
Median Strips	134	0.497403	Manhole Cover	5	3.610918
Utility Poles	87	0.925341	Benches	1	4.709530
Traffic Lights	15	2.630089	Pedestrian Tunnel	2	4.304065
Road Bollard	46	1.552530	Fire Hydrants	0	5.402677
Crosswalk Signals	17	2.512306	Roundabouts	3	4.016383
Billboards	42	1.641477	Roadside Vegetation	2	4.304065

4. Bag-of-Objects

각각의 이미지 내에서 어떠한 물체들이 존재하는지에 대한 검출 작업이 끝났다면 검출된 물체의 빈도수를 기준으로 이미지를 하나의 벡터로 나타낼 수 있는 벡터화 작업이 가능하다. 아래의 식처럼 하나의 이미지를 대표하는OBJ 벡터가 N개의 클래스의 빈도수에 대해서 obj_n와 같이 표현이 가능하여 최종적으로$1\times N$ 크기의 벡터를 얻게 될 수 있다.

$$\mathrm{OBJ} = [obj_1, obj_2, obj_3, \ldots, obj_N] \tag{1}$$

하지만 이렇게 표현된 벡터에 대해서 바로 VPR 작업을 수행하기에는 일반화나 성능 측면에서 떨어질 가능성이 높다. 왜냐하면 도심 환경 속에서 자주 출현되는 가로수나 신호등 같은 경우는 빈번히 출현될 가능성이 높기 때문에 VPR 작업에 있어서 성능을 저하시킬 수 있다. 또는 버스 정류장이나 조형물 같은 빈번히 출현되지 않고 작업을 수행하기에 가치가 높은 물체들은 빈도 수가 낮게 나오기 때문에 똑같은 가중치를 주는 것은 불공평하다. 따라서 본 연구에서는 추출된 물체들에 대해서 자연어 처리에서 사용되는 통계적 처리 방법인 TF-IDF(Term Frequency Inverse Document Frequency)를 사용하여 가중치 작업을 조정하였다.

TF(Term Frequency)는 특정 물체가 전체 dataset에서 얼마나 자주 등장하는지를 나타내는 값으로 각 이미지에서 출현된 각 물체를 모두 더한 값으로 나타낼 수 있다.

$$TF(obj_i, i) = \frac{\text{특정 물체 } obj_i \text{가 이미지 } i \text{ 내에서 등장하는 횟수}}{\text{이미지 } i \text{ 내의 모든 물체의 총 등장 횟수}} \tag{2}$$

DF(Document Frequency)는 특정 물체가 등장한 이미지의 수의 값이다. IDF(Inverse Document Frequency)는 DF의 역수로 특정 물체가 이미지 군에서 사용이 많다고 판단되면 값을 작아지게 사용이 적다고 판단되면 값을 크게 하는 가중치 역할을 하는 값이다.

$$IDF(obj_i, I) = \log\left(\frac{\text{전체 이미지의 수}}{\text{물체 } obj_i \text{를 포함한 이미지의 수} + 1}\right) \tag{3}$$

$$\mathrm{TF-IDF}(obj_i, i, I) = \mathrm{TF}(obj_i, i) \cdot \mathrm{IDF}(obj_i, I) \tag{4}$$

이렇게 구해진 TF와 IDF값을 곱한 TF*IDF를 통해서 최종적으로 VPR 작업을 수행한다.

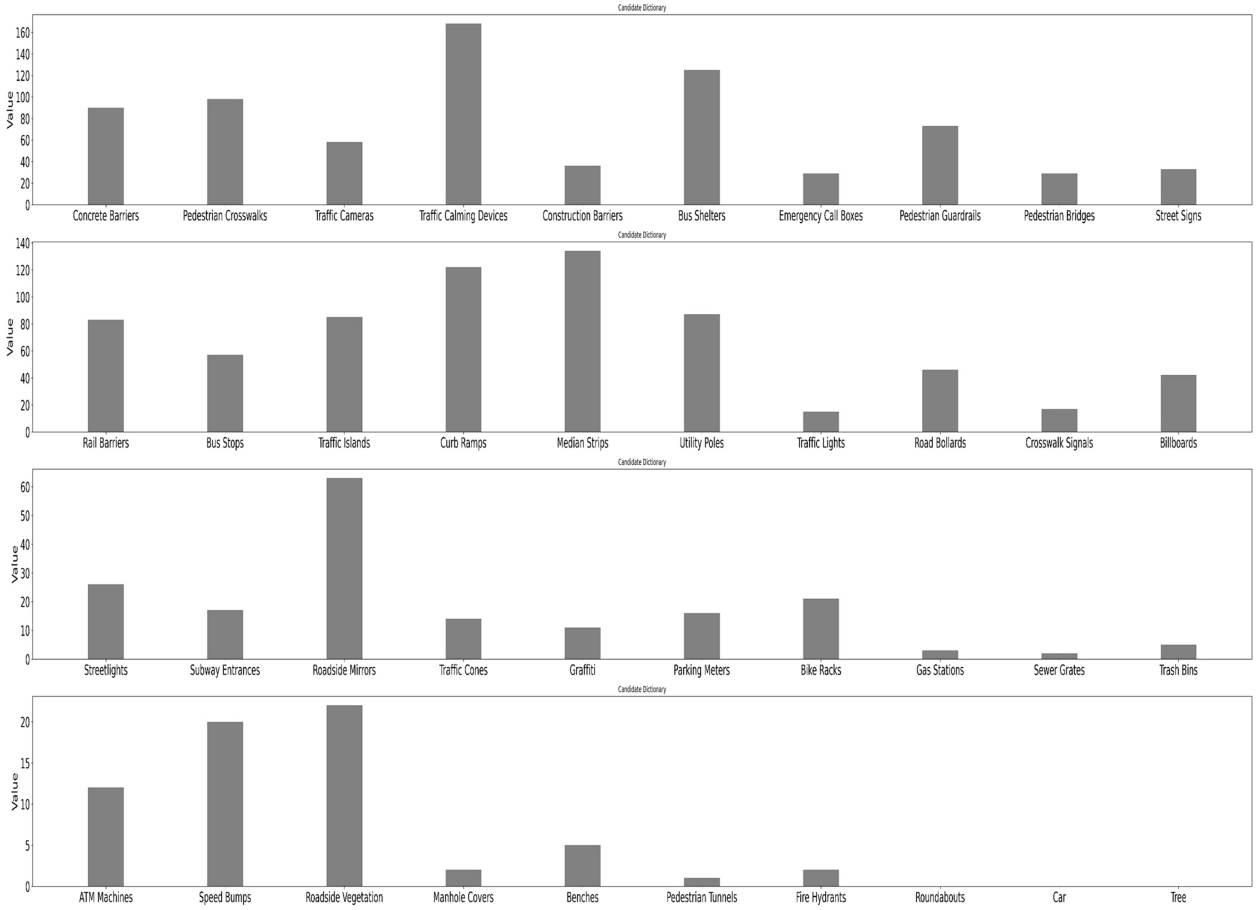

Fig. 5. TF histogram

5. 실 험

5.1. Dataset

도심 환경 속에 대한 데이터를 사용하기 위해 dataset 공유 웹 사이트인 Mapillary에서 Berlin Kurfürstendamm dataset를 사용하였다. 연구의 순서도는 물체 탐지 모델로 물체를 탐지한 후 TF-IDF를 적용해 이미지 벡터화를 진행했고 이렇게 추출된 벡터를 통하여 cosine similarity로 VPR을 수행하여 query 이미지에 대한 최상위 유사도 이미지 5장, 최하위 유사도 이미지 5장을 뽑아 대표적인 VPR 모델인 VLAD와의 정성적 평가를 실시하여 실험을 마무리하였다. 또한 조도에 대한 실험을 위해 Chicago Night City dataset [19]에 대해서도 동일한 방법을 이용하여 모델의 조명 변화에 대한 강건성을 대략적으로 파악하였다.

5.1.1 Berlin Kudamm

Berlin Kudamm dataset은 총 3.4 km의 거리로 222장으로 구성되어 있으며 각 이미지 시퀀스는 10~15 m 사이의 거리 차이를 두고 있다.

Fig. 6. Traffic calming device examples.

5.1.2 Chicago Night City

Chicago Night City dataset은 총 195장의 이미지로 구성되어

	Concrete Barriers	Construction Barriers	Emergency Call Boxes	Traffic Calming Devices	Rail Barriers	...	Fire Hydrants	Manhole Covers	Roundabouts	Pedestrian Tunnels	Car Washes
1	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	1.0	1.0	1.0	...	0.0	0.0	0.0	0.0	0.0
3	1.0	0.0	0.0	3.0	0.0	...	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	6.0	1.0	...	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	4.0	0.0	...	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...
218	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
219	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0
220	0.0	3.0	0.0	3.0	1.0	...	0.0	0.0	0.0	0.0	0.0
221	0.0	1.0	3.0	5.0	0.0	...	0.0	0.0	0.0	0.0	0.0
222	1.0	2.0	2.0	6.0	1.0	...	0.0	0.0	0.0	0.0	0.0

Fig. 7. TF dataframe

있으며 각 이미지 시퀀스는 5~10 m 사이의 거리 차이를 두고 있다

5.2. Bag-of-Objects 표현

Dataset의 각각의 이미지를 BoO 표현을 위해 CLIP을 통해서 물체 탐지를 완료한 후 각 물체의 DF와 IDF를 적용한 결과는 Table 1과 같고 이러한 물체를 기준으로 BoO형식으로 표현한 그래프는 Fig. 5와 같다.

Fig. 7은 각 이미지에서 나온 물체들을 pandas 라이브러리의 dataframe 형식으로 표현한 것이다. 모든 222장의 각 이미지에 대해서 각 물체의 빈도수를 나타낸 것으로 하나의 이미지를 1×40 크기의 하나의 벡터로 표현할 수 있다.

추출된 DF를 살펴보면 가장 많이 출현된 물체로는 Fig. 6의 빨간색 상자와 같이 속도를 제한하는 교통 장치인 traffic calming device였다. 제일 출현 빈도가 낮은 물체로는 회전 교차로인 roundabouts였다.

이렇게 추출 빈도수와 IDF를 고려하여 VPR의 작업에 있어 가중치를 부여하는 작업이 가능하다.

IDF를 살펴보면 DF에서 낮은 값을 얻은 회전교차로나 소화전 같은 물체들이 빈도수가 현저히 낮아 가중치가 높게 부여되었다. 이렇게 가중치 작업이 완료된 벡터를 통해 query 이미지에 대한 가장 유사한 이미지를 dataset에서 추출할 수 있다. 또한 실험에서 모든 이미지에 대해서 빈번히 출현되는 차나 나무와 같은 물체들은 추출이 잘되는 물체에 속하지만 VPR 작업을 수행하기에는 적합하지 않아 벡터화 작업에서는 배제하였다.

5.3. Visual Place Recognition

최종적으로 query 이미지에 대해서 가장 유사한 이미지를 찾기 위한 작업들이 마무리되었고 Fig. 9 왼쪽과 같이 query 이미지에 대해서 최상위 유사도 이미지 5개를 추출하였다.

Fig. 8. Bottom-K visual place recognition

Query 이미지는 dataset의 201번째 이미지이며 최상위 유사도 이미지로는 차례대로 202, 208, 193, 206, 205번째 이미지가 추출되었다.

시퀀스에서도 알 수 있듯이 첫 번째 후보 이미지부터 다섯 번째 이미지까지 query 이미지와 동일한 장소와 멀리 떨어지지 않은 장소인 것을 확인할 수 있다.

Fig. 9. Ours and VLAD top ranked candidate scenes

Fig. 9의 오른쪽과 같이 VPR의 대표적인 모델인 VLAD에 대해서도 같은 Berlin Kudamm dataset을 이용하여 실험하였다. VLAD의 경우, 첫 번째와 두 번째 최상위 유사도 이미지가 query 이미지에서 찍은 이미지와 가까운 거리에서 찍은 이미지를 얻을 수 있었으나 나머지 후보 이미지들에 대해서는 좋지 못한 결과를 얻으므로 본 연구의 VPR에 대한 가능성과 경쟁력을 확인하였다.

최하위 유사도는 Fig. 8과 같이 query 이미지 201번째에 대해서 차례대로 87, 82, 121, 143, 139번째 이미지를 얻은 것으로 query 이미지와 상관없는 위치 장소인 것을 확인할 수 있다.

Fig. 10. Top k visual place recognition (Chicago night city)

마지막으로 VPR의 조명 변화에 대한 실험 결과는 Fig. 10과 같다. Chicago Night City 46 번째 이미지를 query 이미지로 해서 차례대로 94, 47, 96, 64, 44 번째 이미지를 얻을 수 있었다. 상위 두 번째 이미지(47), 다섯 번째 이미지(44)와 같이 좋은 결과를 얻을 수도 있었지만 나머지 3개의 후보 이미지를 살펴보면 query 이미지에 대해서 전혀 상관없는 이미지인 것을 확인할 수 있다. 하지만 두 번째와 다섯 번째 이미지에 대해서 좋은 후보 이미지를 얻은 점을 고려하면 충분히 상대적으로 밤 거리와 같은 어두운 환경에 대해서도 가능성을 보여줬다.

6. 결 론

기존의 BoO 형태의 global descriptor와 처음 보는 dataset에서 강인한 대규모 언어 이미지 모델인 CLIP을 통해 VPR 작업을 수행했다. 본 연구는 local descriptor와 global descriptor 중간 성격의 descriptor를 추출하는 방법을 제시하였고 이러한 descriptor들의 특징으로는 직관적이며 위 descriptor들의 단점을 충분히 보완할 수 있었다.

또한 자연어 처리에서 시작된 통계 방법인 TF-IDF를 이미지에 적용하여 단순히 물체를 탐지하는 것이 아닌 가중치를 조절하여 모델을 차별화하였고 실험을 통해서 가능성을 직관적으로 확인하였다. 하지만 기존 VPR의 고질적 한계점인 조명 변화에 대해서 완벽하지 않다는 한계를 보여주었다. 따라서 향후 연구

를 통하여 조명 변화에 대한 강인성을 확보하고 loop closure검출 성능을 다른 방법과 정량적으로 비교해야 한다.

감사의 글

이 논문은 2023년도 정부(과학기술정통부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임 (No. 2021R1F1A1057949).

REFERENCES

[1] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches", *Proc. of Computer Vision–ECCV 2012: 12th European Conf. Computer Vision*, pp. 73-86, Florence, Italy, 2012.

[2] D. G. Lowe, "Object recognition from local scale-invariant features", *Proc. of the seventh IEEE international conf. computer vision*, pp. 1150-1157, Kerkyra, Greece, 1999.

[3] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features", *Proc. of Computer Vision–ECCV 2006: 9th European Conf. Computer Vision*, pp. 404-417, Graz, Austria, 2006.

[4] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints", *Proc. of Workshop on statistical learning in computer vision*, Vol. 1. No. 1-22, pp. 1-6, 2004.

[5] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation", *Proc. of* 2010 *IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, pp. 3304-3311, San Francisco, USA, 2010.

[6] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 40, No. 6, pp. 1437-1451, 2018.

[7] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping", *Proc. of 2007 IEEE International Conf. Robotics and Automation*, pp. 1-7, Rome, Italy, 2007.

[8] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0", *Int. J. Rob. Res.*, Vol. 30, No. 9, pp. 1100-1123, 2011.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", *Proc. of the IEEE conf. computer vision and pattern recognition*, pp. 73-86, Boston, USA, 2012.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", *Proc. of the IEEE conf. computer vision and pattern recognition*, pp. 770-778, Las Vegas, USA, 2016.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks", *Proc. of the IEEE conf. computer vision and pattern recognition*, pp.7132-7141, Salt Lake City, USA, 2018.

[12] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module", *arXiv preprint arXiv:1807.06514*, pp. 1-14, 2018.

[13] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module", *Proc. of the European conf. computer vision (ECCV)*, pp. 3-19, Munich, Germany, 2018.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need", *Proc. of 31st Annual Conf. Neural Information Processing Systems (NIPS 2017)*, pp. 1-11, California, USA, 2017.

[15] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels", *Proc. of In International conf. machine learning (ICML)*, pp. 1691-1703, 2020.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale", *arXiv preprint arXiv:2010.11929*, pp. 1-22, 2020.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision", *Proc. of International conf. machine learning (ICML)*, pp. 8748-8763, 2020.

[18] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges", *Proc. of Computer Vision–ECCV 2014: 13th European Conf.*, pp. 391-405, Zurich, Switzerland, 2014.

[19] X. Tan, K. Xu, Y. Cao, Y. Zhang, L. Ma, and R. W. H. Lau, "Night-time scene parsing with a large real dataset". *IEEE Trans. Image Process.*, Vol. 30, pp. 9085-9098, 2021.