Visual Place Recognition 기술 동향

세종대학교 한현덕

Abstract – Visual Place Recognition (VPR) 기술은 자율주행, 위치 추정, 3D 복원 등 많은 분야에서 사용될 수 있는 기술이다. VPR 기술은 이미지의 특징을 검출해 descriptor를 만들고 이를 database에 저장한다. 구축된 database를 바탕으로 query image가 입력으로 들어오면 입력된 이미지의 descriptor를 만들고 database에 저장된 descriptor들 중 가장 유사한 descriptor를 찾는다. 이러한 관점에서 VPR 기술에는 두 가지의 이슈가 있다. 첫번째는 descriptor가 이미지의 특징을 얼마나 잘 표현하는 것에 대한 이슈다. 두번째 이슈는 database에 저장할 수 있는 메모리는 한정적이고 많은 이미지의 descriptor를 저장해야 하므로 descriptor의 크기가 너무 커도 안 된다. 이 문서는 위의 이슈들을 극복하기 위해 제안된 여러 방법들을 소개하고 최근 연구 동향인 딥러닝 기반 방법들도 소개한다.

1. 서론

Visual Place Recognition (VPR) 기술은 자율주행, 위치 추정, 3D 복원과 같은 많은 분야에서 사용될 수 있는 기술로 컴퓨터 비전 분야에서 많은 연구가 진행되어 왔다. VPR 문제는 보통 image retrieval task [1], [2], [3], [4] 로 여겨지는데 이는 query image가 주어졌을 때 가장 비슷한 이미지를 반환하는 것이다. 좀 더 구체적으로 보면, database에 학습 데이터 이미지들의 descriptor를 저장해두고 query image가 입력으로 주어지면 query image에 대한 descriptor를 만들어 database에 존재하는 것들과 비교한 후 가장 유사한 descriptor를 갖는 이미지를 database에서 찾아 반환한다.

VPR에는 크게 두 가지의 이슈가 있다. 우선 descriptor가 이미지의 특징을 잘 표현할 수 있어야 한다. 이미지의 특징을 잘 표현한다는 것은 같은 장면에 대한 이미지에 대해서 비슷한 descriptor를 만들어야 한다는 것과 같다. 하지만 이는 쉽지 않은 문제이다. 같은 장면을 담고 있는 이미지여도 낮에 촬영된 이미지와 밤에 촬영된 이미지는 매우 다른 특성을 갖는다. 또한 지나다니는 사람이나 자동차와 같은 물체의 존재도 descriptor가 이미지의 특징을 잘 표현하기 힘들도록 한다.

Descriptor의 성능 관련 이슈 뿐 아니라 database에 저장할 수 있는 메모리의 양이 제한적이라는 점도 고려해야 한다. 이는 descriptor의 성능을 높이기 위해 descriptor의 차원을 무한히 늘릴 수 없다는 것을 의미한다. Descriptor의 차원이 높으면 database의 메모리에도 부담을 주고 query image가 들어왔을 때 이와 유사한 이미지를 탐색하는데 걸리는 시간에도 부담을 준다. 따라서 VPR 문제에서는 위 두 가지 이슈들을 잘 고려해야 한다.

2. VPR 기술 동향

2-1. Bag-of-Words

VPR을 위한 descriptor를 만들기 위해 aggregation-based 방법들이 많이 제안되어 왔다. 초기의 방법 중 하나인 Bag-of-Words (BoW) [5], [6] 방법은 우선 SIFT , SURF와 같은 방법을 이용하여 hand-crafted local feature를 추출한다 [7], [8]. SIFT, SURF와 같은 방법으로 추출된 feature descriptor는 VPR에서 사용하는 descriptor와 구분하기 위해 앞으로 local descriptor라는 표현을 사용한다. 이미지에서 추출된 local descriptor들을 바탕으로 K-means clustering을 수행한다. 이때 clustering된 집합은 $\mathcal{C} = \{c_1, \dots, c_k, \dots, c_K\}$로 정의하고 이때 c_k의 차원은 이미지에서 추출된 local descriptor의 차원과 동일하다. 보통 이미지의 특징 추출에 SIFT를 이용하는데 이 경우 추출된 local descriptor와 c_k의 차원은 모두 128차원이다.

만약 어떤 이미지에서 N개의 D차원의 local descriptor $\{\boldsymbol{x}_i\}$가 존재하고 K개의 cluster centers $\{\boldsymbol{c}_k\}$가 존재한다면, BoW의 방법으로 이미지를 표현하는 descriptor V는 $K \times D$ 행렬로 나타낼 수 있다. 구체적으로 행렬 V의 (j,k) 위치의 원소는 다음과 같이 계산될 수 있다.

$$V(j,k) = \sum_{i=1}^{N} a_k(\boldsymbol{x}_i) \tag{1}$$

식 (1)에서 $a_k(\boldsymbol{x}_i)$는 local descriptor $\boldsymbol{x}_i$가 k^{th} cluster center $\boldsymbol{c}_k$와 가장 가까운 경우에만 1의 값을 갖고 나머진 0의 값을 갖는다. 즉, BoW 방법은 K개의 cluster centers를 만들고 각각의 군집 안에 local descriptor가 포함된 수를 이용하여 descriptor를 만드는 방법이다.

2-2. Vector of Locally Aggregated Descriptor (VLAD)

Vector of Locally Aggregated Descriptor (VLAD) [9], [10] 방법은 BoW 방법을 개선한 방법이다. VLAD 방법에서도 BoW와 마찬가지로 SIFT, SURF와 같은 방법으로 local descriptor를 추출한 후 이를 바탕으로 K-means clustering을 수행한다. BoW에서와 마찬가지로 어떤 이미지에서 N개의 D차원의 local descriptor $\{\boldsymbol{x}_i\}$가 존재하고 K개의 cluster centers $\{\boldsymbol{c}_k\}$가 존재한다면, BoW의 방법으로 이미지를 표현하는 descriptor V는 $K \times D$ 행렬로 나타낼 수 있다. 구체적으로 행렬 V의 (j,k) 위치의 원소는 다음과 같이 계산될 수 있다.

$$V(j,k) = \sum_{i=1}^{N} a_k(\boldsymbol{x}_i)\,(x_i(j) - c_k(j)) \tag{2}$$

식 (2)에서 $x_i(j)$와 $c_k(j)$는 각각 i^{th} descriptor와 k^{th} cluster center의 j차원을 의미한다.

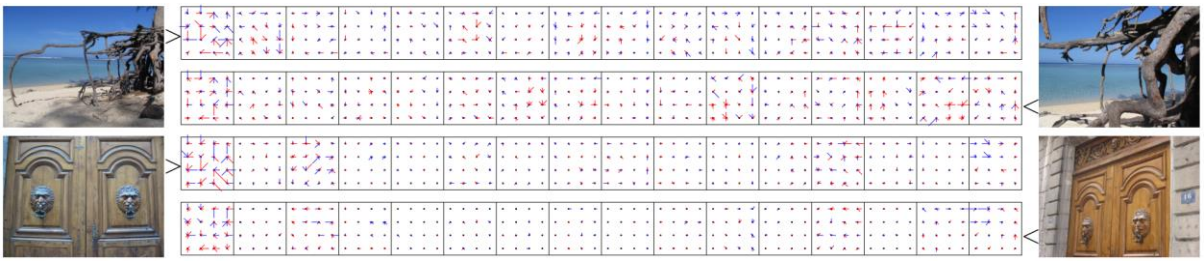

그림 1. VLAD를 SIFT처럼 표현한 예시 그림

[출처] H. Jégou, M. Douze, C. Schmid, and P. Pérez. "Aggregating local descriptors into a compact image representation." *In Proc. CVPR*, 2010.

그림 1은 $\boldsymbol{x}_i$가 128차원이고 $K = 16$일 때, 128차원을 SIFT처럼 표현한 그림이다. 식 (2)에 의하면 각 차원의 값은 음수가 나올 수 있으므로 음수인 경우엔 빨간색으로 표시한 모습이다.

2-3. NetVLAD

NetVLAD [11] 방법은 VLAD 방법을 기반으로 학습 가능하도록 개선시킨 deep-learning based 방법이다. 식 (2)에서 $a_k(\boldsymbol{x}_i)$는 불연속함수로 미분을 힘들게 만드는 요인이고 이는 backpropagation 방법으로 학습할 수 없다는 것을 의미한다. 이를 해결하기 위해 NetVLAD 저자들은 $a_k(\boldsymbol{x}_i)$를 다음과 같이 $\bar{a}_k(\boldsymbol{x}_i)$로 대체한다.

$$\bar{a}_k(\boldsymbol{x}_i) = \frac{e^{-\alpha\|x_i - c_k\|^2}}{\sum_{k'} e^{-\alpha\|x_i - c_{k'}\|^2}} \tag{3}$$

식 (3)에서 α는 hyperparameter로 무한대의 값을 갖는다면 기존의 VLAD 방법과 동일해진다. 식 (3)은 식 (4)로 정리될 수 있고 NetVLAD에서 사용되는 descriptor V는 식 (5)와 같다.

$$\bar{a}_k(\boldsymbol{x}_i) = \frac{e^{-w_k^T x_i + b_k}}{\sum_{k'} e^{-w_{k'}^T x_i + b_{k'}}} \quad where \quad \boldsymbol{w}_k = 2\alpha \boldsymbol{c}_k,\ b_k = -\alpha\|\boldsymbol{c}_k\|^2 \tag{4}$$

$$V(j,k) = \sum_{i=1}^{N} \left(\frac{e^{-w_k^T x_i + b_k}}{\sum_{k'} e^{-w_{k'}^T x_i + b_{k'}}} \left(x_i(j) - c_k(j)\right) \right) \tag{5}$$

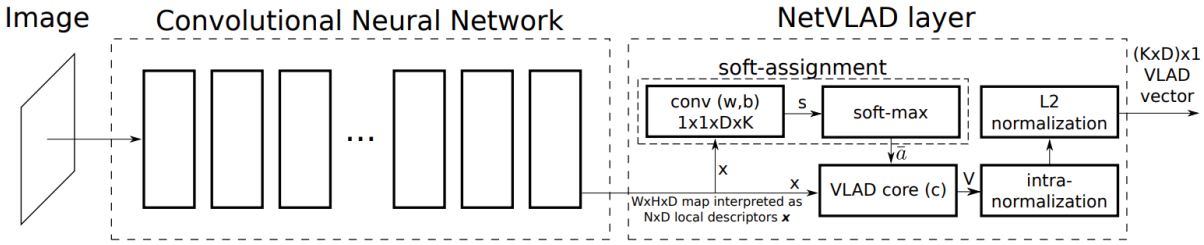

그림 2. Flowchart of NetVLAD

[출처] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. "Netvlad: Cnn architecture for weakly supervised place recognition." *IEEE International Conference on Computer Vision*, 2015.

그림 2는 NetVLAD의 flowchart를 나타낸다. NetVLAD의 flowchart는 크게 두 부분으로 구성된다. 이 중 식 (5)에 해당하는 부분은 NetVLAD layer이다. Convolutional Neural Network는 이미지의 local descriptor를 추출하는 과정으로 NetVLAD 저자들은 AlexNet [12] 과 VGG-16 [13] 을 사용하여 실험했고 두 경우 모두 마지막 convolutional layer까지만 crop하여 사용했다.

NetVLAD 저자들은 test data로 Google Street View Time Machine을 사용했다. 이 dataset의 특징

은 GPS 정보를 이용하여 대략적인 위치는 파악할 수 있지만 대략적인 위치가 비슷하더라도 서로 다른 물체를 담고 있는 이미지일 수 있다는 점이다. NetVLAD에서는 이러한 dataset에서 학습할 수 있도록 potential positives 그룹과 definite negatives 그룹으로 나눈 후 Triplet ranking loss를 정의하여 학습에 사용했다.

2-4. Semantic Reinforced Attention Learning (SRALNet)

SRALNet [14] 방법은 BoW, VLAD, NetVLAD방법에서 모든 local descriptor들을 사용하여 descriptor를 구성했는데 모든 local descriptor가 어떤 task에 연관되어 있지는 않다고 주장한다. 즉, 추출된 local descriptor 중 현재 task에 필요한 local descriptor만을 이용하여 descriptor를 만들면 더 효과적일 수 있다는 것이다. SRALNet 방법은 사전에 학습된 DeepLabV3 [15] 모델의 semantic 정보를 VPR task에서 사용할 수 있도록 하는 방법을 제안했다.

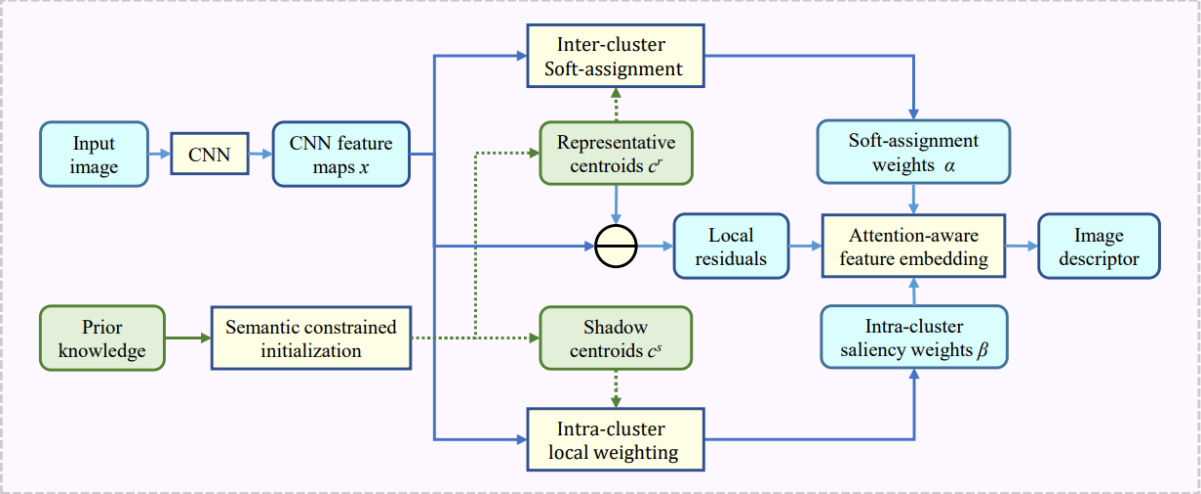

그림 3. Flowchart of SRALNet

[출처] G Peng, Y Yue, J Zhang, Z Wu, X Tang, and D Wang. "Semantic reinforced attention learning for visual place recognition." *In 2021 IEEE International Conference on Robotics and Automation (ICRA),* 2021.

그림 3은 SRALNet의 전체적인 흐름도를 보여준다. Inter-Cluster Soft-assignment는 NetVLAD의 식 (3)과 같고($\boldsymbol{c}_k^r = \boldsymbol{c}_k$) 이는 local feature $\boldsymbol{x}_i$가 k^{th} cluster $\boldsymbol{c}_k^r$에 포함될 확률을 나타낸다. 그림 3의 Intra-cluster는 Inter-cluster로 나누어진 cluster 내에서 Voronoi cell로 다시 분할하여 $\boldsymbol{c}_k$가 포함된 분할은 informative area I라고 정의하고 나머지 분할 영역들은 ambiguous area s_l로 간주한다. 여기서 ambiguous area에 속한 local descriptor들을 이용하여 shadow centroid $\boldsymbol{c}_{kl}^s$를 정의한다. Sub-cluster들의 분포가 uniform하고 각각 동일한 공분산행렬을 지닌 가우시안 분포와 같다고 가정하면, Intra-cluster local weight $\beta_k(\boldsymbol{x}_i)$는 식 (6)으로 구할 수 있다. 이때 $\beta_k(\boldsymbol{x}_i)$의 의미는 local feature $\boldsymbol{x}_i$가 informative area I에 위치할 확률이다.

$$\beta_k(\boldsymbol{x}_i) = P(I|\boldsymbol{x}_i, \boldsymbol{c}_k) = \frac{P(\boldsymbol{x}_i|I,\boldsymbol{c}_k)P(I|\boldsymbol{c}_k)}{\sum_{l=1}^{S} P(\boldsymbol{x}_i|s_l,\boldsymbol{c}_k)P(s_l|\boldsymbol{c}_k)+P(\boldsymbol{x}_i|I,\boldsymbol{c}_k)P(I|\boldsymbol{c}_k)} = \frac{e^{-\alpha\|x_i-c_k^r\|^2}}{\sum_{l=1}^{S} e^{-\alpha\left\|x_i-c_{kl}^s\right\|^2}+e^{-\alpha\|x_i-c_k^r\|^2}} \quad (6)$$

최종적으로 SRALNet에서 사용되는 descriptor V는 식 (7)과 같으며 $\bar{a}_k(\boldsymbol{x}_i)$는 식 (3)에서 사용한 것과 같다.

$$V(j,k) = \sum_{i=1}^{N} \left(\bar{a}_k(\boldsymbol{x}_i) \beta_k(\boldsymbol{x}_i) \left(x_i(j) - c_k(j) \right) \right) \tag{7}$$

3. VPR 성능 평가

표 1은 VLAD 방법과 BoW 방법의 비교를 나타낸다. 해당 표에서 BoF는 Bag-of-Feature를 나타내며 BoW와 같은 의미다.

Method	bytes	UKB	Holidays
BOF, k=20,000 (from [9])	10.364	2.92	0.446
miniBOF [9]	20	2.07	0.255
	80	2.72	0.403
	160	2.83	0.426
VLAD, k=16, ADC 16×8	16	2.88	0.460
VLAD, k=64, ADC 32×10	40	3.10	0.495

표 1. VLAD와 BoF의 mAP (mean average precision) 비교

[출처] H. Jégou, M. Douze, C. Schmid, and P. Pérez. "Aggregating local descriptors into a compact image representation." *In Proc. CVPR*, 2010.

표 1은 VLAD 방법이 BoW 방법에 비해 더 적은 양의 메모리로 더 높거나 비슷한 성능을 낼 수 있음을 보인다. 그림 4는 같은 메모리에서 VLAD 방법이 BoW 방법보다 월등히 좋은 성능을 나타냄을 보이는 그래프다.

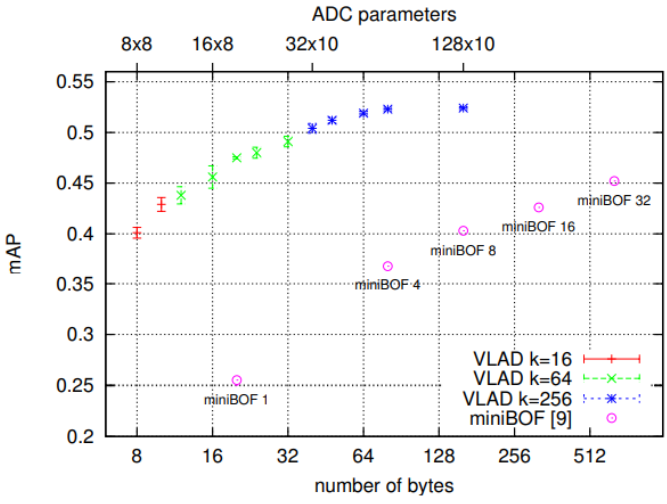

그림 4. 같은 메모리에서 VLAD와 BoF의 mAP (mean average precision) 비교

[출처] H. Jégou, M. Douze, C. Schmid, and P. Pérez. "Aggregating local descriptors into a compact image representation." *In Proc. CVPR*, 2010.

그림 5는 NetVLAD와 VLAD 방법의 성능 비교 그래프다. 두 방법 모두 local descriptor를 추출하

는 방법은 SIFT, SURF, AlexNet, VGG와 같은 방법을 사용할 수 있다. 해당 그림은 다양한 local descriptor를 추출하는 방법을 사용했을 때의 성능도 함께 보이고 있다. NetVLAD 방법이 빨간색 계열의 그래프이며 파란색 계열은 VLAD방법에서 이용되는 local descriptor를 deep learning based 방법으로 추출한 경우다. 검정색은 deep learning 방법이 아닌 SIFT와 같은 방법으로 local descriptor를 추출하고 VLAD 방법을 사용했을 때의 성능을 보인다. 그래프를 보면 알 수 있다시피 여러 데이터셋에서 NetVLAD의 성능이 VLAD의 성능보다 좋은 것으로 확인된다.

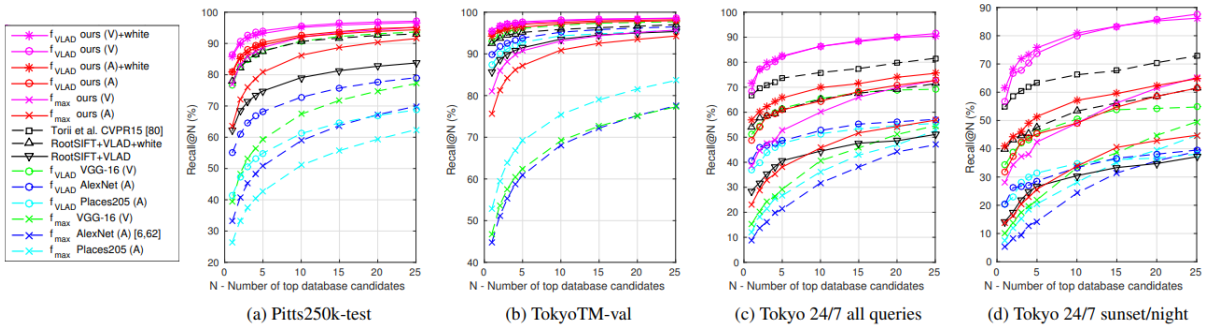

그림 5. Recall comparison between NetVLAD and VLAD with methods of extracting local descriptor

[출처] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. "Netvlad: Cnn architecture for weakly supervised place recognition." *IEEE International Conference on Computer Vision*, 2015.

마지막으로 그림 6은 SRALNet과 NetVLAD의 성능을 비교한다. 그림 6에서 (··) 모양의 그래프는 4096 차원의 local descriptor를 이용했을 때의 결과이고 (– · –) 모양의 그래프는 512차원의 local descriptor를 이용했을 때의 결과를 나타낸다. 해당 그래프에서 SRALNet 방법은 빨간색의 그래프이고 나머지 그래프들은 NetVLAD 방법을 기반으로 하는 여러 방법들을 나타낸다. 그림 6을 보면 local descriptor의 차원이 클 때와 작을 때 모두 SRALNet 방법이 NetVLAD 방법보다 좋은 성능을 보임을 확인할 수 있다.

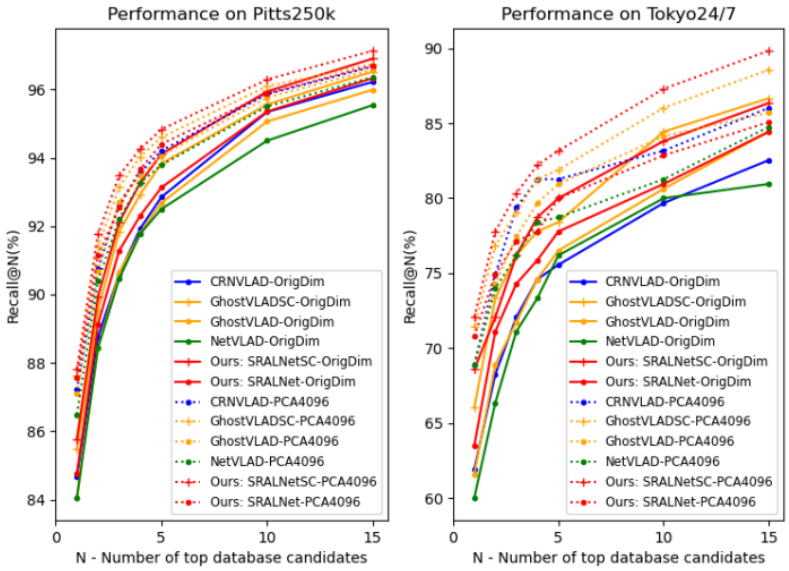

그림 6. Evaluation between SRALNet and methods based on NetVLAD

[출처] G Peng, Y Yue, J Zhang, Z Wu, X Tang, and D Wang. "Semantic reinforced attention learning for visual place recognition." *In 2021 IEEE International Conference on Robotics and Automation (ICRA),* 2021.

4. 결론

지금까지 VPR을 목적으로 한 다양한 논문을 살펴봤다. BoW는 가장 먼저 제안된 방법으로 이후의 방법들보다는 성능이 약간 떨어진다. VLAD 방법은 BoW를 개선한 방법으로 더 적은 메모리로 BoW 방법보다 좋은 성능을 보인다. 이후 VLAD 방법을 기반으로 한 많은 연구가 진행되었으며 이를 학습 가능한 모델로 만든 NetVLAD 방법이 제안되기까지 이른다. 이후에도 NetVLAD 방법을 기반으로 한 많은 연구가 있었고 최근에는 task에 적합한 local descriptor만을 이용하여 정확성을 높이려는 SRALNet과 같은 모델도 제안되고 있다.

- References

[1] R. Arandjelovic and A. Zisserman. "DisLocation: Scalable descriptor distinctiveness for location recognition." *In Proc. ACCV,* 2014.

[2] D. M. Chen, G. Baatz, K. Koeser, S. S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. "City-scale landmark identification on mobile devices." *In Proc. CVPR*, 2011.

[3] J. Knopp, J. Sivic, and T. Pajdla. "Avoiding confusing features in place recognition." *In Proc. ECCV,* 2010.

[4] G. Schindler, M. Brown, and R. Szeliski. "City-scale location recognition." *In Proc. CVPR*, 2007.

[5] G. Csurka, C. Bray, C. Dance, and L. Fan. "Visual categorization with bags of keypoints." *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[6] J. Sivic and A. Zisserman. "Video Google: A text retrieval approach to object matching in videos." *In Proc. ICCV*, volume 2, pages 1470–1477, 2003.

[7] D. Lowe. "Distinctive image features from scale-invariant keypoints." *IJCV*, 60(2):91–110, 2004.

[8] T. Tuytelaars and K. Mikolajczyk. "Local invariant feature detectors: A survey." *Foundations and Trends R in Computer Graphics and Vision,* 3(3):177–280, 2008.

[9] R. Arandjelovic and A. Zisserman. "All about VLAD." *In Proc. CVPR*, 2013.

[10] H. Jegou, M. Douze, C. Schmid, and P. Perez. "Aggregating local descriptors into a compact image representation." *In Proc. CVPR*, 2010.

[11] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. "Netvlad: Cnn architecture for weakly supervised place recognition." *In CVPR,* 2016.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks." *In NIPS,* pages 1106–1114, 2012.

[13] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition." *In Proc. ICLR,* 2015.

[14] G Peng, Y Yue, J Zhang, Z Wu, X Tang, and D Wang. "Semantic reinforced attention learning for visual place recognition." *In 2021 IEEE International Conference on Robotics and Automation (ICRA),* 2021.