CNN을 이용한 SIFT-NonSIFT 영상 패치 분류

*김종영 **정지수 ***원치선

동국대학교 전자전기공학부

*endofmovie108@naver.com **whtnek@gmail.com ***cswon@dongguk.edu

SIFT-NonSIFT Classification of Image Patches using CNN

*Jong-young Kim **Gi-su Chung ***Chee Sun Won

Dept. of Electronic and Electrical Eng., Dongguk University

요약

본 연구는 주어진 영상 패치가 SIFT 특성(즉 SIFT 특징점)을 포함하는지 아니면 SIFT 특성을 나타내지 않은지(즉, NonSIFT)를 구분하기 위해 CNN 구조를 활용한다. 우선 CNN을 훈련시키기 위해 기존의 SIFT 알고리즘으로 SIFT 특징점을 검출하고 이 점들을 중심으로 SIFT 영상 패치를 생성하며, 반대로 SIFT 특징점을 포함하지 않는 영상 패치를 같은 수만큼 생성하여 데이터 베이스를 구축한다. 이를 바탕으로 3개의 컨벌루션 레이어를 갖는 CNN구조를 훈련시켰다. 훈련된 CNN에 대해 오픈소스 라이브러리 VL_FEAT의 SIFT 알고리즘을 수행한 결과를 그라운드 트루스 (ground truth)로 정하고 제안된 방법의 결과와 비교한 결과 90.8%의 분류 정확도를 얻었으며 CNN을 이용한 SIFT 특징점 검출의 가능성을 확인하였다.

1. 서론

영상에 존재하는 특징점은 카메라 캘리브레이션, 객체 추출, 영상 스티칭(stitching), 및 검색 등 여러 분야에 활용된다. 이미지의 회전, 스케일 등 기하학적 변화에 강인한 성능을 보여주는 SIFT는 특징점 추출 및 매칭에서 우수한 성능을 갖는 Hand-crafted 방법의 특징점 추출 및 매칭의 대표적인 방법이다[1]. 한편, CNN이 다양한 영상문제에서 우수한 성능을 보이면서 CNN을 통한 영상 특징 벡터의 활용에 대한 기대가 커지고 있다. SIFT와의 비교결과에서도 대부분의 영상 검색 문제에서 CNN이 우수한 결과를 보이고 있다. 그러나 회색조 영상과 같이 특정 영상에 대해 아직도 SIFT의 우수성이 확인되고 있다[2]. 따라서 CNN과 SIFT의 장점을 결합하여 성능 향상을 기대할 수 있다[3]. 본 논문에서는 주어진 영상 패치가 SIFT의 특성을 보유하고 있는지(즉, SIFT 패치), 아니면 SIFT의 특징점을 포함하고 있지 않은 영상 패치인지(NonSIFT 패치)를 구분하는 CNN 구조를 제안한다.

2. 본론

2.1 SIFT Training dataset 확보

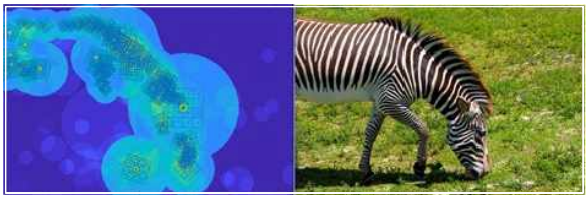
SIFT 및 NonSIFT 영상 패치를 구분하기 위한 CNN을 구현하기 위해 우선 training dataset의 확보가 필수적이다. 이 논문에서는 같은 영상 내에서 SIFT 영역과 NonSIFT 영역을 추출해 내는 방법을 채택하였다. 이를 위해 VL_FEAT[4]에서 제공하는 MATLAB 오픈소스 라이브러리를 사용하였으며, SIFT 알고리즘을 수행하는 VL_SIFT 함수를 이용한다. SIFT의 특징점으로 인정되는 최저의 contrast양을 조절하는 파라메터인 peak_thresh값은 15를 부여하였는데 이는 영상의 특징점으로서 큰 역할을 하는 영역을 추출하기 위함이다.

영상 패치는 descriptor의 크기(스케일)에 해당하는 영역을 추출하였고 32x32으로 크기 변경 (resize) 하였고 orientation은 0으로 고정하였다. 원본 영상은 범용적으로 쓰이는 MS COCOdataset[5]을 이용하였다. SIFT 특징점은 추출된 영상 패치의 중앙에 위치하게 하여 특징점의 가외자리 위치에 따른 분류에 있어서의 오차를 없게 하였다. 이를 통해 SIFT 영상 패치 10만장을 확보하였다. 이때 영상에서 SIFT 패치의 주변은 NonSIFT 패치의 주변에 영향을 줄 수 있기 때문에 정확한 분류작업이 필요하다.

2.2 NonSIFT Training dataset 확보

NonSIFT 영상 패치를 같은 영상에서 추출하기 위해 SIFT 영상 패치의 특징점 좌표를 추출하고 해당 특징점의 주변에 descriptor의 대각선길이를 반지름으로 하는 원을 생성하여 가중치를 더하는 방식으로 <그림 1>과 같이 SIFT 주변영역을 나타내는 saliency map을 제작하였으며, 이를 통해 SIFT 주변영역으로부터 완전히 격리된 영역에서 순수한 NonSIFT 영역을 확보할 수 있도록 하였다. SIFT의 peak_thresh 파라메터를 0으로 설정하였을때의 training dataset 또한 확보하였는데, peak_thresh의 파라메터에 따라 추출되어지는 training dataset의 특성이 달라질 수 있으므로 peak_thresh 파라메터에 따라 서로 다른 training dataset으로 학습시킨 별개의 CNN의 테스트 score를 비교할 수 있다.

<그림 1>의 saliency map을 바탕으로 영상 내에 NonSIFT 영역에 해당하는 곳이 어딘지 확인할 수 있다. 이 saliency map의 SIFT 가중치가 0 인 영역 즉, 지도상 진한 파랑색 영역에 해당하는 곳을 떼어 내어 32x32 크기의 NonSIFT 영상 패치 3만장을 만들었다. <그림 2>는 추출한 SIFT 및 NonSIFT 영상 패치의 예를 보여주고 있다.

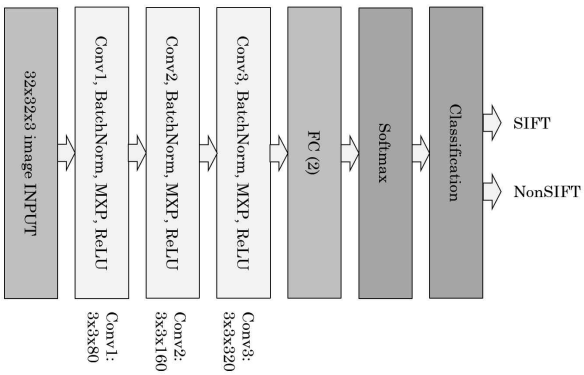


<그림 1> SIFT saliency map (왼쪽), 원영상 (오른쪽)

<그림 2> 32x32 SIFT 및 NonSIFT 영상 패치

2.3 CNN 구조

본 논문에서는 <그림 3>과 같이 convolution(Conv), max pooling(MXP), ReLU, 그리고 batch normalization을 갖춘 3개의 layer을 사용하였다. 즉, 32x32x3 RGB 영상 패치를 input data로 갖는 input layer와 3x3 크기의 convolution 커널 3개를 구성하고 fully-connected layer를 구성하였다. convolution layer는 stride 속성을 부여하지 않았으며, max pooling layer로 이미지 사이즈를 줄였다. 구축한 각 50,000개 총 100,000개의 영상 패치 중에 20%는 validation 이미지, 80%는 training 이미지로 사용하였다.

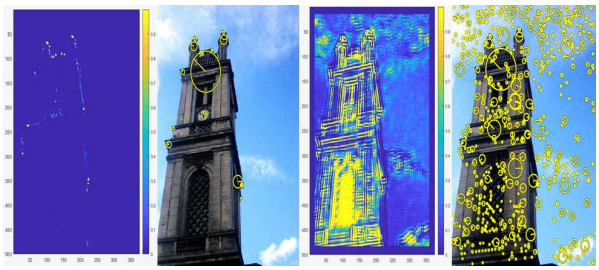


<그림 3> CNN 구조

2.4 CNN 테스트 score의 시각화

CNN의 성능을 시각적으로 확인하기 위해 다음과 같이 샘플 이미지를 32x32의 window로 이동시키며 잘라내어 CNN의 input data로 넣는 window-sliding 방식으로 해당 영상 패치의 중심 pixel 좌표에 CNN의 테스트 score를 넣어 <그림 4>와 같이 시각화하였다. <그림 4>에서 왼쪽의 영상은 위의 방식으로 제작한 CNN의 테스트 score 영상이며 노란색 부분은 높은 SIFT 점수를 가지며 강한 특징점을 나타내고, 파란색 부분은 약한 SIFT 점수로 특징점으로서의 의미가 적은 부분이다.

<그림 4> window-sliding을 통한 CNN score 시각화



<그림 5> peak_thresh 파라메터 변경에 따른 서로 다른 CNN모델의 테스트 score 이미지의 변화

<그림 5>의 왼쪽 두 개의 영상은 각각 SIFT training dataset을 VL_SIFT의 peak_thresh 파라메터값 15를 부여하여 제작한 것으로 학습시킨 CNN모델로 시각화한 CNN의 테스트 score 영상과 VL_SIFT함수를 사용하여 기존의 SIFT feature를 시각화한 영상이며, 오른쪽 두 개의 영상은 각각 SIFT training dataset을 peatk_thresh 파라메터값 0을 부여하여 제작한 것으로 학습시킨 CNN모델로 시각화한 CNN의 테스트 score 영상과 왼쪽의 두 번째 영상과 같은 방법으로 SIFT feature를 시각화한 영상이다. 기존의 SIFT알고리즘이 각각의 파라메터에 해당하는 특징점을 다수 추출한 영역에서 설계한 CNN의 테스트 score 또한 동일하게 높은 점수를 획득함을 알 수 있다. VL_SIFT함수가 peak_thresh 파라메터에 따라 검출해내는 SIFT 특징점의 수의 경향과 설계한 CNN이 비슷한 경향을 보이는 것을 확인할 수 있다.

2.5 Gray saliency map을 활용한 ROI제시

이미지 검색 기술에서 이미지 내에 중요한 정보를 포함하는 ROI (Region Of Interest)를 추출해내는 기술에 대한 요구가 높아지고 있다. Zhen Liang의 논문[6]에서는 Canny edge 검출 알고리즘을 이용하여 ROI를 제시하며 이를 바탕으로 이미지의 검색에 필요한 Local descriptor의 추출을 제시한다. 하지만 많은 이미지들은 edge가 아닌

부분에서 특별한 의미를 갖는 경우가 많으며 우리가 설계한 CNN으로 제작한 gray saliency map은 이미지내의 ROI를 제시하는 또 다른 방법이 될 수 있다.

2.6 제안한 CNN과 VL_SIFT와의 비교

설계한 CNN의 VL_SIFT와의 영상 패치 내 특징점 포함 비교 테스트를 위해 SIFT-NonSIFT에 대해 각각 10,000개의 32x32 영상 패치를 test dataset으로 사용하였다. SIFT Ground Truth 테스트 패치는, 패치 내에 VL_SIFT 함수로 부터 특징점이 하나라도 검출이 되면 SIFT로 라벨링하여 구축하였으며 반대로 하나도 검출이 되지 않으면 NonSIFT로 라벨링 하였다. CNN testing 분류 결과 정확성은 다음 표와 같다.

<표 1> SIFT/Non-SIFT testing 분류 결과

label CNN predict	SIFT	Non-SIFT
SIFT	**99.34%**	0.66%
Non-SIFT	17.69%	**82.31%**

<표 1>과 같이 SIFT 영상 패치를 SIFT로 예측하는 경우는 99.34%로 매우 높은 정확도를 보였다. NonSIFT 영상 패치를 NonSIFT로 예측하는 경우는 82.31%의 정확도를 나타내었다. 따라서 전체 평균 인식률은 90.8%를 얻었다. SIFT 영상 패치에 비해 NonSIFT 영상 패치의 testing 분류 정확도가 상대적으로 낮은 경향을 보이는데, 이는 SIFT 특징점은 그 특성상 임의의 영상에 대해 공간적 위치상으로 군집하는 경향을 보이므로 SIFT 영상 패치 내에 다수의 SIFT 특징점이 존재할 가능성이 높아 우리의 CNN이 높은 분류 정확도를 보인다. 반면에 영상 패치의 중심부근에 특징점이 위치하지 않는 NonSIFT 영상 패치의 경우, 중심에서 벗어난 영상 패치의 가장자리에서 SIFT 특징점의 경향을 보이는 성분들이 포함될 수 있으므로 NonSIFT의 영상 패치 이지만 SIFT 영상 패치로 분류되는 오류가 발생할 수 있다.

3. 결론

본 논문은 기존의 SIFT 알고리즘을 대체할 수 있을 정도의 성능을 가진 CNN을 통해서 영상의 특징점을 갖는 영상 패치를 구분할 수 있는 가능성을 보였다. CNN을 위한 training data의 확보가 필수적이며 정확한 SIFT, NonSIFT data를 확보하는 것이 관건이라고 할 수 있다. 본 논문에서 제안한 CNN 구조를 이용하여 SIFT 및 NonSIFT의 구분 정확도를 90.8%정도 확보하였다. 다양한 SIFT 및 NonSIFT 영상 패치 dataset을 training dataset에 추가함으로써 정확도를 더욱 높일 수 있을 것으로 기대된다. 후속 연구로는 VL_SIFT의 peak_thresh 및 edge_thresh 파라메터에 따라 다른 training data를 확보하도록 함으로써 다양한 파라메터에 대응할 수 있는 CNN의 설계로 사용자가 파라메터를 임의로 부여하여 원하는 feature를 추출할 수 있도록 할 수 있을 것이다. 또한, SIFT-NonSIFT를 분류하는데 있어서 필요한 training 영상에 data augmentation기법을 활용하여 training dataset을 더욱 풍부하게 보완할 수 있을 것이며 나아가 특징점의 orientation까지 정확히 예측하는 새로운 CNN을 설계하여 SIFT 기법을 대체할 수준으로 성능을 향상시킬 것이다.

4. 감사의 글

본 연구는 한국연구재단(교육부) 기본연구지원사업 (NRF-2015R1D1A1A01057269)의 지원을 받음.

5. 참고문헌

[1] D. G. Lowe, “Distinctive image features from scale-invariant keypoints”. International journal of computer vision, vol. 60, no. 2, Pages 91-110, 2004.

[2] Liang Zheng, Yi Yang, and Qi Tian, “SIFT Meets CNN: A Decade Survey of Instance Retrieval”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 5, Pages 1224-1244, 2018.

[3] Ke Tan, Yaowei Wang, Dawei Liang, Tiejun Huang, and Yonghong Tian, “CNN vs. SIFT for Image Retrieval: Alternative or Complementary?”, Proceedings of the 2016 ACM on Multimedia Conference, Pages 407-411, 2016.

[4] http://www.vlfeat.org/

[5] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Persona, D. Ramanan, P. Dollar, and C. L. Zitnick, “Microsoft COCO: Common objects in context”, in ECCV, ser. Lecture Notes in Computer Science, vol. 8693, Pages 740 - 755, 2014.

[6] Liang Z., Fu H., Chi Z., Feng D. “Salient-SIFT for Image Retrieval.” In: Blanc-Talon J., Bone D., Philips W., Popescu D., Scheunders P. (eds) Advanced Concepts for Intelligent Vision Systems. ACIVS 2010. Lecture Notes in Computer Science, vol 6474. Springer, Berlin, Heidelberg, 2010.