다중 입력 영상과 Cross-Input Neighborhood Differences를 이용한 사람 재인식 기법

김현우*, 김형준*, 임동혁**, 황인준*
*고려대학교 전기전자공학과
**한국전자통신연구원
e-mail : *{guihon12, hyungjun89, ehwang04}@korea.ac.kr,
**iammoni@etri.re.kr

A Person Re-identification Scheme Using Multiple Input images and Cross-Input Neighborhood Differences

Hyeonwoo Kim*, Hyungjoon Kim*, Dong-Hyuck Im**, Eenjun Hwang*
School of Electrical Engineering, Korea University

요 약

최근 CCTV 사용이 보편화되면서 방범 목적으로 서비스 시설이나 공공시설에 설치되는 CCTV의 수가 급격하게 증가하고 있다. 그에 따라 CCTV를 감시하는 노동력이 부족해지는 문제가 발생하여 이를 대체하기 위해 카메라 영상을 통하여 한번 인식한 사람을 다른 시간이나 장소에서 촬영된 영상에서 다시 인식하는 사람 재인식 기술이 주목받고 있다. 또한, 이러한 사람 재인식 기술은 보안 분야뿐만 아니라 영화나 드라마와 같은 영상 컨텐츠에 적용되어 불법 복제물을 찾는 일에 사용될 수도 있다. 기존의 사람 재인식에는 이미지의 유사도를 계산하는 방법이 사용되었지만, 조명이나 카메라 각도가 달라지면 성능이 급격하게 떨어지는 문제가 있었다. 최근에는 딥러닝 기술이 발달하면서 전반적인 영상처리 분야의 성능이 향상되었고, 사람 재인식 분야 역시 딥러닝을 활용하면서 성능이 향상되었다. 하지만 딥러닝을 활용한 방법의 경우 보통 두 개의 이미지를 입력으로 사용하여 같은지 다른지를 판단하게 되므로 각 이미지의 공통점이나 차이점을 동시에 고려하기는 어려운 점이 있다. 본 논문에서는 이러한 점을 해결하기 위해 세 개의 사람 이미지를 입력으로 사용하여 특징을 추출하고, 특징 맵을 재구성하여 각 이미지의 차이점과 공통점을 동시에 고려하며 학습할 수 있는 모델을 제안한다.

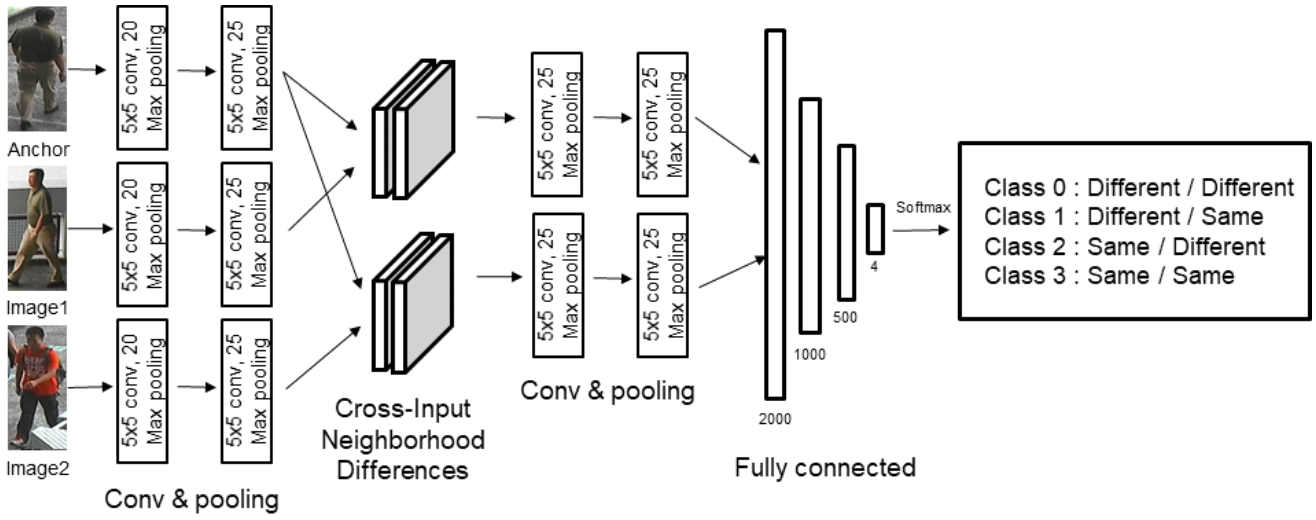
1. 서 론

최근 CCTV 사용이 보편화되면서 방범 목적으로 설치되는 CCTV 수가 급격하게 증가하고 있다. 이를 통해 실종된 사람을 찾거나 범죄자를 추적하는 것이 가능해졌지만, 24시간 동안 수많은 CCTV를 감시할 수 있는 노동력은 한정되어 있다는 문제가 있다. 이러한 문제를 해결하기 위해 카메라를 통해 한번 인식한 사람을 다른 시간 혹은 장소에서 촬영된 영상을 통해 다시 인식할 수 있는 사람 재인식 기술이 주목받고 있다. 이러한 사람 재인식 기술은 CCTV 뿐만 아니라 영화나 드라마와 같은 영상 컨텐츠에 적용될 수 있다. 예를 들어, 사람과 의상을 같이 인식함으로써 다른 영상에서 같은 사람을 재인식하여 불법 복제된 영상을 검출할 수 있다.

하지만 사람 재인식 기술에는 영상에 있는 객체의 급격한 움직임이나 밝기 변화 등으로 인해 성능이 급격하게 떨어진다는 문제가 있다. 최근에는 딥러닝의 출현으로 영상 분류, 객체 인식과 같은 영상처리 분야의 난제들이 해결되고 있으며, 사람 재인식 연구들도 딥러닝을 사용하면서 성능이 크게 향상되고 있다[1, 2].

대부분의 딥러닝 기반의 사람 재인식 모델은 각 이미지의 특징을 추출하기 위해 ResNet[3], VGG[4], Inception[5]과 같은 이미지 분류에서 사용되는 네트워크를 사용한다. 이러한 이미지 분류 네트워크들은 서로 다른 종류의 객체를 분류하기 위해 고안된 모델로, 같은 종류의 객체 사이에서 이 객체가 동일한 지 다른지를 구분하지 못한다. 이를 해결하기 위해, 이러한 모델들의 경우 추출된 특징에서 얼굴, 상의, 하의, 신발 등과 같이 사람을 나타내는 영역끼리 특징 맵을 재구성하여 학습하는 방법을 사용하였다[6, 7].

이와 다르게 IDLA[10]에서는 가벼운 네트워크를 사용하여 두 개 이미지에서 특징 맵을 추출하고, 추출된 특징 맵의 Cross-Input Neighborhood Differences를 계산하여 두 이미지 사이의 차이에 대한 특징 맵을 새롭게 구성하였다. 이렇게 만들어진 특징 맵은 다시 네트워크를 통과하게 되고 최종적으로 Softmax를 통해 두 이미지가 같은지 다른지 판단한다.



(그림 1) 네트워크 구조

Triplet loss[8]를 적용한 연구 사례도 있다[9]. 기존의 방법들에서는 두 개의 이미지를 사용하여 같은지 다른지를 판단하기 때문에 차이점에 대한 부분만을 학습하게 되어 차이점과 공통점을 같이 고려하지 못한다는 문제가 있었다. 이 연구에서는 같은 이미지 쌍(Positive pair)과 다른 이미지 쌍(Negative pair)을 동시에 고려하는 Triplet loss를 적용하여 서로 다른 사람일 경우 거리를 늘리고 서로 같은 사람일 경우 거리를 좁히도록 학습시켰다. 또한, 다양한 데이터 셋으로 모델을 평가하여 성능이 크게 향상된 것을 확인하였다. 이러한 점들로 미루어 보아, 네트워크를 학습시킬 때 같은 이미지 쌍(Positive pair)과 다른 이미지 쌍(Negative pair)을 동시에 사용하여 네트워크를 학습시키면 이미지가 같은지 다른지를 판단할 때 차이점과 공통점을 동시에 고려할 수 있다고 볼 수 있다. 따라서 본 논문에서는 세 개의 이미지에서 특징을 추출하고, 추출된 특징들을 두 쌍의 특징 맵으로 재구성하여 학습에 사용하는 모델을 제안한다.

본 논문의 나머지 구성은 다음과 같다. 먼저 2장에서는 사람 재인식 모델에 대한 전반적인 내용에 대하여 설명한다. 3장에서는 데이터 셋에 대해 설명하고 4장에서는 기존 방법과 성능 비교를 한 결과에 대하여 설명한다. 마지막으로 5장에서는 결론 및 향후 연구 방향으로 본 논문의 끝을 맺는다.

2. 사람 재인식 모델

위에서 언급했듯이, 본 논문에서는 세 개의 이미지를 사용하여 특징을 추출하고, 추출된 특징들 사이의 관계를 두 쌍의 특징 맵으로 재구성하여 학습하는 방법을 제안한다. 이를 위해, IDLA[10]의 네트워크를 그림 1과 같이 확장하여 새로 구성하였다.

확장된 네트워크에서는 세 개의 이미지를 사용하므로 기존보다 한 쌍의 특징 맵이 추가로 만들어지게 된다. 이러한 특징 맵들은 마지막에 있는 Fully connected 레이어를 통해 정보를 공유하게 되고, 최종 결론을 내리게 된다. 하지만 IDLA의 경우 두 개의 Fully connected 레이어로만 구성되어 있기 때문에 늘어난 특징 맵들의 정보를 모두 반영하기 어려울 수 있다는 문제가 있다. 이러한 문제를 해결하기 위해, 확장 모델에서는 Fully connected 레이어를 추가하여 모든 특징 맵들을 충분히 반영하여 결론을 내릴 수 있도록 만들었다.

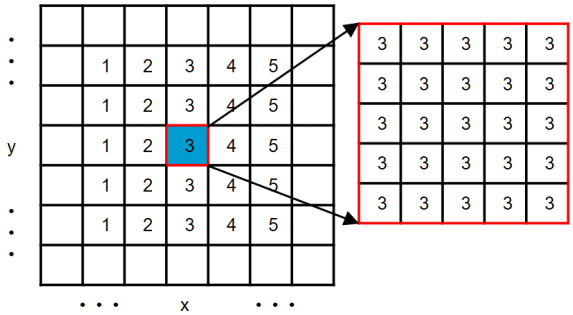
그림 1을 보면 세 개의 이미지에서 추출된 특징 맵들이 Cross-Input Neighborhood Differences 레이어를 통해 재구성되는 것을 볼 수 있다. 식 (1)은 두 특징 맵 사이의 관계를 계산하는 방법을 나타낸 것이다. f_i는 컨볼루션 레이어를 통해 Anchor 이미지에서 추출된 특징 맵이고 $g_{i,1}$과 $g_{i,2}$는 비교되는 이미지(Image1/Image2)에서 추출된 특징 맵을 의미한다. 그리고 $\mathbb{1}(5,5)$는 1로 구성된 5×5 행렬을 뜻하고, $N[g_i(x,y)]$는 g_i행렬에서 (x,y)가 중심인 5×5 행렬을 의미한다. 그림 2에서는 이해를 돕기 위해 식 (1)을 실제 특징 맵에 적용했을 때의 예시를 보여주고 있다.

$$K_i(x,y) = f_i(x,y)\mathbb{1}(5,5) - N[g_{i,n}(x,y)] \quad (1)$$

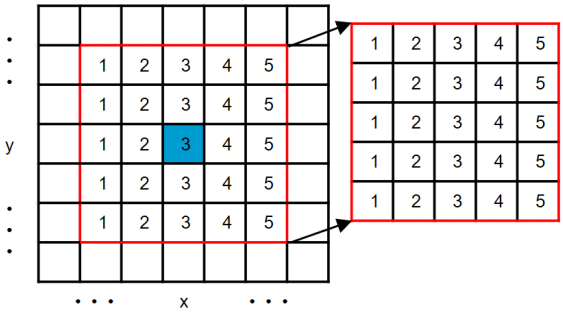
$$K_i^{'}(x,y) = g_{i,n}(x,y)\mathbb{1}(5,5) - N[f_i(x,y)] \quad (2)$$

Cross-Input Neighborhood Differences 레이어를 통해 특징 맵 K_i와 식 (1)을 반전시켜서 만들어진 특징 맵 $K_i^{'}$가 만들어지는데, $K_i^{'}$의 식을 $N[g_i(x,y)] - f_i(x,y)\mathbb{1}(5,5)$와 같이 구성하게 되면 K_i와 절댓값이 같은 특징 맵을 결과로 얻게 된다. 그렇게 되면 두 개의 식을 사용하여 특징

맵을 구성하는 것이 의미가 없어지므로, 식 (2)와 같이 식을 구성하여 특징 맵 $K_i^{'}$를 구성하였다.



(1) $f_i(x,y)\mathbb{1}(5,5)$ 예시



(2) $N[g_i(x,y)]$ 예시

(그림 2) Cross-Input Neighborhod Differences

본 논문에서는 세 개의 이미지를 사용하므로 그림 1과 같이 두 쌍의 특징 맵이 만들어진다. 이 특징 맵들은 식 (1)과 식 (2)를 통해 Cross-Input Neighborhood Differences 레이어에서 재구성된다. 이렇게 만들어진 특징 맵은 다시 컨볼루션 레이어를 통과하면서 특징을 추출하게 되고 최종적으로 Fully connected 레이어와 Softmax를 통해 네 개의 클래스로 분류된다.

3. 데이터셋

본 논문에서는 CUHK03 데이터 셋[11]을 사용하여 네트워크를 학습시키고 평가한다. CUHK03 데이터 셋은 6개의 서로 다른 카메라에서 촬영된 1,360명의 데이터로, 총 13,164개의 이미지로 구성되어있다. 또한, 이 데이터 셋에서는 보행자 인식기를 통해 자동으로 얻어진 이미지 셋과 수작업을 통해 사람의 영역을 지정하여 얻은 이미지 셋 2가지를 제공한다.

본 논문에서는 수작업을 통해 만들어진 이미지 셋을 네트워크 학습에 사용하였으며, 학습 데이터에는 1,260명으로 구성된 12,221장의 이미지가 사용되었으며 평가 데이터로는 100명의 943장의 이미지를 사용하였다. 학습 데이터는 같은 사람의 이미지 쌍과 서로 다른 사람 이미지 쌍을 같은 비율로 만들어 사용하였다.

4. 실험 결과

제안하는 모델이 사람 재인식을 얼마나 잘하는지 평가하기 위해 CUHK03 데이터를 이용하여 [10] 모델과 우리가 제안하는 모델을 학습하고 성능을 비교하였다. 본 실험은 Intel (R) Core (TM) i7-8700 CPU, 32G DDR4 memory, NVIDIA Geforce GTX 1080ti, Python 3.5 환경에서 진행되었다.

성능 평가지표로는 Rank-1, Rank-5, Rank-8 정확도를 사용하였으며, 여기서 Rank란 그림 3과 같이 쿼리 이미지 셋과 갤러리 이미지 셋을 구성하였을 때, 쿼리 이미지와 같다고 인식된 갤러리 셋 이미지들의 점수 순위를 말한다. Rank-1은 갤러리 셋에서 쿼리 이미지와 같다고 인식한 점수가 가장 높은 이미지가 실제 정답인 경우를 뜻하며, 이와 마찬가지로 Rank-5는 갤러리 셋에서 쿼리 이미지와 같다고 인식한 점수가 높은 다섯 개의 이미지 안에 정답이 있는 경우를 의미한다.

(그림 3) 평가 데이터 구성

아래의 표 1에서는 IDLA, Fully connected 레이어를 추가하기 이전 모델, 우리가 제안하는 모델 세 개 모델의 Rank-1, Rank-5, Rank-8 정확도를 보여주고 있다. 표를 보면 IDLA에서 입력 이미지를 추가하여 확장한 모델의 경우 Rank-1 정확도가 기존 모델보다 약 10% 향상되었으며, 여기에 Fully connected 레이어를 추가한 모델의 경우에는 Rank-1 정확도를 포함하여 성능이 전체적으로 향상된 것을 확인할 수 있다.

<표 1> 성능 평가 결과

	IDLA[10]	Model 1	Model 2
Rank-1	66.6	77.2	**85.2**
Rank-5	89.4	93.5	**96.3**
Rank-8	96.5	98.5	**99.1**

Fully connected 레이어를 추가하기 이전 모델 - Model 1
우리가 제안하는 모델 - Model 2

5. 결 론

본 논문에서는 사람 재인식 모델이 두 이미지 사이의 차이점과 공통점을 반영하면서 학습되도록 만들기 위해 세 개의 사람 이미지를 입력으로 사용하여 두 쌍의 특징 맵으

로 재구성하는 방법을 제안한다. 기존의 사람 재인식 모델을 확장하고 CUHK03 데이터 셋을 사용하여 모델을 학습하였다. 그리고 실험을 통해 Rank-1, Rank-5, Rank-8 정확도를 측정하였으며, 제안하는 모델이 사람 재인식 모델의 성능을 향상시킨다는 것을 입증하였다..

이번 연구에서는 기존의 모델과의 성능 비교를 위해 IDLA와 같은 가벼운 네트워크를 사용하여 이미지의 특징을 추출하여 실험을 진행하였지만, 향후에는 ResNet, VGG, Inception과 같이 검증된 네트워크를 통해 이미지의 특징을 추출하여 제안하는 모델의 학습에 사용하는 방법에 대하여 연구할 예정이다.

Acknowledgement

본 연구는 문화체육관광부 및 한국저작권위원회의 2019년도 저작권기술개발사업의 연구결과로 수행되었음. [2018- micro-9500, 음악 및 동영상 모니터링을 위한 지능형 마이크로 식별 기술 개발]

참고문헌

[1] Geng, Mengyue, et al. "Deep transfer learning for person re-identification." arXiv preprint arXiv:1611.05244 (2016).

[2] Bai, Xiang, et al. "Deep-person: Learning discriminative deep features for person re-identification." arXiv preprint arXiv:1711.10658 (2017).

[3] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[4] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[5] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[6] Quan, Ruijie, et al. "Auto-ReID: Searching for a Part-aware ConvNet for Person Re-Identification." arXiv preprint arXiv:1903.09776 (2019).

[7] Lin, Yutian, et al. "Improving person re-identification by attribute and identity learning." Pattern Recognition (2019).

[8] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[9] Hermans, Alexander, Lucas Beyer, and Bastian Leibe. "In defense of the triplet loss for person re-identification." arXiv preprint arXiv:1703.07737 (2017).

[10] Ahmed, Ejaz, Michael Jones, and Tim K. Marks. "An improved deep learning architecture for person re-identification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[11] Li, Wei, et al. "Deepreid: Deep filter pairing neural network for person re-identification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.