Pyramid pooling을 이용한 CNN 기반의 Human Parsing 기법

*최인규 **고민수 **송 혁

전자부품연구원

*cig2982@keti.re.kr **kmsqwet@keti.re.kr **hsong@keti.re.kr

CNN-based Human Parsing Technique Using Pyramid Pooling

*Choi, Inkyu **Ko, min-soo **Song, hyok

Korea Electronics Technology Institute

요약

최근 딥러닝 기술의 발전으로 영상 분류 및 영상 내 객체 검출뿐만 아니라 CNN 기반의 segmentation 기술도 개발되어 다른 요소까지 포함한 직사각형 영역의 검출 영역이 아닌 경계까지 고려한 분리가 가능하게 되었다. 더불어 사람 영역을 신체 부위나 의류 부분과 같은 세부 영역으로 나누어 분리하는 human parsing 기술까지 연구되고 있다. Human parsing은 의류 스타일 분석 및 검색, 사람의 행동 인식 및 추적과 같은 분야에도 응용될 수 있다. 본 논문에서는 Spatial pyramid pooling layer를 이용하여 영상 전체에 대한 공간적 분포 및 특성 정보를 고려한 human parsing 기법을 제안한다. Look into person(LIP) dataset을 이용하여 기존의 다른 segmentation 및 human parsing 기법과 제안하는 기법을 비교하여 제안하는 기법의 human parsing 결과가 보다 정교한 분리가 가능한 것을 확인하였다.

1. 서론

딥러닝 기술의 발전으로 단순한 목적의 영상 처리를 넘어 영상분석 및 이해와 같은 문제도 해결이 가능해짐에 따라 자율주행 자동차, 능동형 CCTV와 같은 최첨단 정보통신기술에도 활용이 가능하게 되었다. 카메라로 획득한 비디오에 대해서 YOLO, SSD[1, 2]와 같은 CNN 기반의 객체검출 기술이 개발되어 실시간 검출이 가능해졌으며 검출 영역에 해당하는 대상에 대해 추적 알고리즘을 적용하여 다수의 객체에 대한 객체 추적 또한 가능하다. 그리고 CNN 기반의 semantic segmentation[3] 기술이 개발되어 pixel 단위의 클래스 분리가 가능해지면서 보다 정교한 객체분리와 이를 통한 영상의 context 정보 또한 추출이 가능하다.

Human parsing은 segmentation의 일환으로 사람 영역을 상의, 하의, 얼굴 등의 여러 분야로 분리하는 기술이다. 이를 이용하여 의류 인식 및 데이터 검색, 행동 인식 그리고 사람 식별과 같은 고수준의 기술에 응용할 수 있다.

본 논문에서는 convolutional neural network와 spatial pyramid pooling layer를 이용한 human parsing 기법에 대해 서술한다. 고수준의 특징을 추출하는 CNN과 영상 전체의 공간적 특성을 고려하기 위한 pyramid pooling layer를 접목하여 pixel 단위의 정교한 human parsing 결과를 획득한다. LIP(Look Into Person) dataset을 이용하여 기존의 segmentation과 human parsing 기법과 제안하는 기법의 결과를 비교한 결과 보다 정교한 parsing 성능을 보여줌을 확인하였다.

2. 본론

2.1 Look into person(LIP) dataset

인간의 부분적 의미 해석을 위해 만들어진 dataset으로 19개의 부분 클래스로 나뉜 50,000장의 영상으로 구성되어 있다. 머리카락, 얼굴, 팔, 다리 등 인간의 신체 부분과 상의, 코트, 바지, 드레스 등 의류 부분으로 나뉘어져 있다. 그림 1과 같이 각 클래스에 대해 픽셀 단위로 정교하게 구분되어 있다.

Figure 1. LIP dataset examples

2.2 Spatial pyramid pooling layer

Human parsing은 영상의 전체의 공간적 분포 및 특성이 중요하기 때문에 부분 receptive field를 고려한 convolution 연산으로는 정확한 결과를 기대하기 어렵다. 그림 2와 같이 spatial pyramid pooling layer를 CNN 뒤에 연결하여 문제를 해결한다. CNN을 통과하여 추출된 특징지도에 대하여 네 가지 크기의 average pooling을 이용하여 1x1, 2x2, 3x3, 6x6 크기를 가지는 pyramid 형태의 특징지도를 추출한다. 이렇게 생성한 각각의 특징지도는 개별적인 CNN을 통과하고 병합하여 pixel 단위의 soft-max 함수를 통해 parsing 결과를 획득한다.

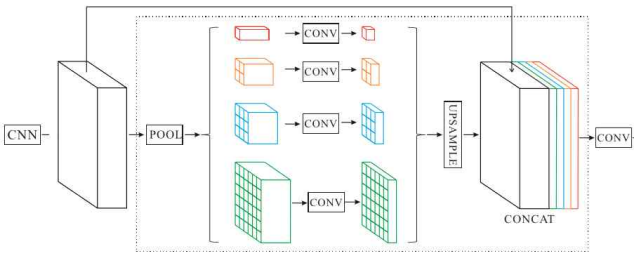

Figure 2. Spatial pyramid pooling module

2.3 실험결과

LIP dataset을 이용하여 기존의 기법 MASK R-CNN[4], JPPNet[5]과 제안하는 기법의 parsing 결과를 비교한다. MASK R-CNN은 instance segmentation 기술이기 때문에 각 클래스 별로 bounding box 정보를 획득하고 데이터를 재구성하여 진행하였다. 아래의 그림 3은 각 기법에 대한 parsing 결과로써 제안하는 기법이 기존의 두 기법에 비해 보다 정교한 parsing 결과를 보여줌을 확인할 수 있다.

3. 결론

본 논문에서는 convolutional neural network와 spatial pyramid pooling layer를 이용한 human parsing 기법을 제안한다. human parsing은 영상 전체의 공간적 분포 및 특성 그리고 서로 다른 클래스에 해당하는 pixel 간 연관성을 고려해야하기 때문에 기존의 convolution 연산으로는 제한이 있다. 따라서 spatial pyramid pooling을 통해 이러한 문제를 해결한다. LIP dataset을 이용하여 기존의 기법과 비교한 결과 제안하는 기법의 parsing 결과가 보다 정교한 것을 확인할 수 있었다.

ACKNOWLEDGMENT

본 논문은 2016년도 도시문제 해결형 기술개발 지원사업 (과제번호 PS160010)을 지원받아 수행한 결과입니다.

참 고 문 헌

[1] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, “You only look once: Unified, real-time object detection”, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.779-788, 2016

[2] W. Liu, D. Anguelov, C. Szegedy, S. Reed, C. Y. Fu and A. C. Berg, “Ssd: Single shot multibox detector”, In European conference on computer vision, Springer, Cham. pp. 21-37, October, 2016

[3] J. Long, E. Shelhamer and T. Darrell, “Fully convolutional networks for semantic segmentation”, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440, 2015

[4] K. He, G. Gkioxari, P. Dollár and R. Girshick, “Mask r-cnn”. In Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE, pp. 2980-2988, October, 2017

[5] X. Liang, K. Gong, X. Shen, and L. Lin, “Look into Person: Joint Body Parsing &Pose Estimation Network and A New Benchmark”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018

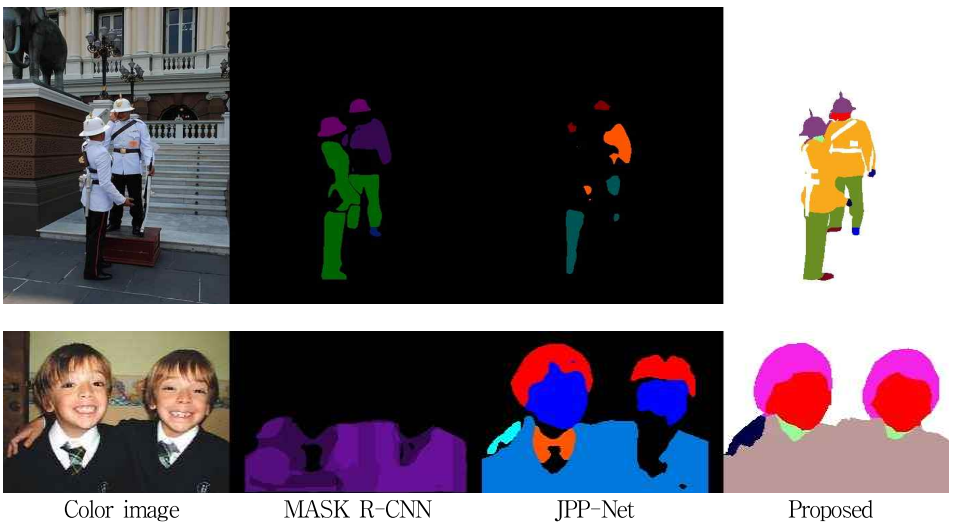

Figure 3. Human parsing results