RAFT 를 이용한 딥러닝 기반 Optical flow 예측 방법 구현 및 고찰

Hyeonseok Chae Wonjun Kim
illusionetic@gmail.com wonjkim@konkuk.ac.kr
Konkuk University

요 약

최근 영상신호처리에 대한 딥러닝 기술이 비약적으로 발전함에 따라 다양한 방면으로 시도되고 있다. 그 중 machine level vision 에서 인지 기능을 하는 optical flow 를 end-to-end 학습 방식으로 제시하여 고성능 결과물을 도출하는 RAFT(Recurrent All-pairs Field Transform for Optical flow, 2020)에 대해 분석하고자 한다. RAFT 는 입력된 두 이미지에 대한 4D correlation volume 을 구축하여 모든 픽셀에 대한 정보를 사용한다. 또한, recurrent neural network 에서 차용한 반복적인 연산 학습 구조를 통하여 결과물인 flow field 의 정확도를 높인다. 해당 모델은 stereo dataset 을 사용하는 다른 모델에 비해 학습 시간이 짧고 용량이 작으면서 error rate 은 낮은 모습을 보인다. 현재 많은 연구에서 optical flow 를 접목하려는 움직임을 보이고 있고 다양하게 활용될 가능성이 다분하다는 점에서 주목할 가치가 있다.

1. 서론

최근 Deep-Learning 기술이 급격하게 성장함에 따라 다양한 분야에서 인공지능을 도입하고 있다. 문제상황에 대한 해결 알고리즘을 학습하기 때문에, 기존의 여러 파이프라인을 거치는 복잡한 방식에서 벗어날 수 있기 때문이다. 이런 최적화 문제를 해결해주고 완성된 결과를 도출하는 end-to-end 학습 방식은 computer Vision 분야에서도 다양하게 시도되고 있다.

현재 computer vision 에서의 딥러닝 연구는 장면을 분석하는 수준에 그쳐 있고 맥락(context)에 대한 이해는 부족한 상황이다. Optical flow 는 관찰자와 장면의 상대적인 움직임과 광원과 객체의 상대성에 의해 발생하는 표면 및 윤곽의 이동을 나타내는 지표로 영상의 맥락을 이해하는 데 있어 중요한 초석이 될 것이라 기대된다. 그동안 optical flow 는 수학적 도식을 통한 전통적인 기법의 알고리즘을 통해 연구가 되어왔다. 하지만 작고 빠른 움직임을 보이는 객체를 인지하지 못한다는 점과 광원의 움직임에 따른 변화를 정확히 구별하지 못한다는 명확한 한계가 존재한다.

RAFT 논문에선 end-to-end 방식으로 모든 픽셀을 이용하여 flow field 를 연산하는 방법을 제안한다. 연속된 두 프레임의 4D Correlation 을 통해 장면간 유사도를 연산하고, 변형된 GRU-cell 을 통해 회귀적으로 delta flow vector 를 update 하는 네트워크를 제시하여 매 학습 단계마다 정확도를 높인다.

본 캡스톤 디자인 보고서의 구성은 다음과 같다. 2 절에서는 표적논문이 제시하는 네트워크 구조에 대해 분석한다. 3 절에서는 해당 모델의 성능을 제시하고 다른 모델과의 비교를 통해 개선점을 제시한다. 4 절에선 optical flow 의 활용방안을 제시하며 앞으로의 비전을 논한다. 5 절에선 결론을 얘기한다.

2. RAFT 상세 내용 소개

입력된 한 쌍의 RGB 이미지($I_1 = (u, v), I_2 = (u', v')$)는 (u, v)좌표계에서 flow field(f^1, f^2)에 대해 $(u', v') = (u + f^1(u), v + f^2(v))$의 관계를 따른다. 해당 모델에선 vector(f^1, f^2)에 초점을 맞추어 미분 가능하고 end-to-end 학습이 가능한 구조로 구성된 세 단계를 제시한다.

2-1. Feature Extraction

학습에 사용될 feature map 은 입력된 연속된 이미지 I_1, I_2 에 대해 convolutional network 로 구해진다. 각 input 에 residual block 을 6 번씩 적용하여 $(\frac{H}{8} * \frac{W}{8} * D)$크기의 고밀도 특징 vector 를 사용할 수 있게 한다.

추가적으로 I_1 에 대해 feature map 과는 동일한 구조를 가졌지만 독립적으로 사용가능한 context map 도 추출하여 추후 iterative updating 부분에서 hidden state 로 사용한다.

2-2. Visual Similarity

Visual Similarity 는 모든 픽셀 쌍의 연관도를 뜻한다. 전 단계에서 구한 두 쌍의 3 차원 feature vector 의 dot product 를 통해 4 차원의 vector 로 변환한 것이 correlation(C)이다. 이러한 correlation 은 크고 작은 움직임들에 대해 민감하게 반응할 수 있게 하기 위해 resolution 에 따라 4-layers 의 pyramid $\{C^1, C^2, C^3, C^4\}$ 를 구축한다. 각 층의 correlation 은 2^k 크기의 kernel 로 pooling 된 형태이며 각 layer 의 correlation 크기는 $H * W * \frac{H}{2^k} * \frac{W}{2^k}$이다. 여기에 correlation lookup 연산을 도입한다. 해당 연산은 4 차원 correlation 을 3 차원 tensor 로 변환하여 학습에 사용할 수 있게 하는 부분으로, flow 의 출발점과 도착점의 정보를 함축적으로 담고있다.

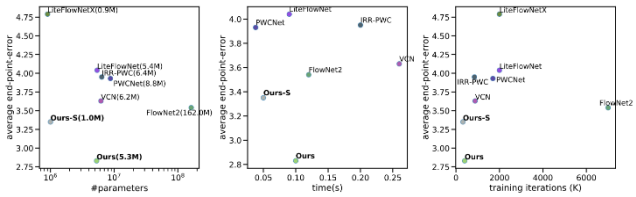
2-3. Iterative Updates

Flow estimator 는 회귀적으로 연산을 시행하여 flow vector 의 변화량 $\Delta f(f_{k+1} = f_k + \Delta f$)를 반복적으로 갱신한다. 각 학습단계마다 0 으로 초기화된 flow vector 에 대해 갱신을 N 번 반복하여 결과의 정확도를 높인다.

해당 과정은 기존 GRU-cell 의 입력단에 convolutional network 를 추가한 구조를 사용한다. Correlation feature, k 번째 flow, context map(hidden state)를 입력으로 GRU-cell 에 의해 activated 되어 매 iteration 마다 flow 와 hidden state 가 출력된다.

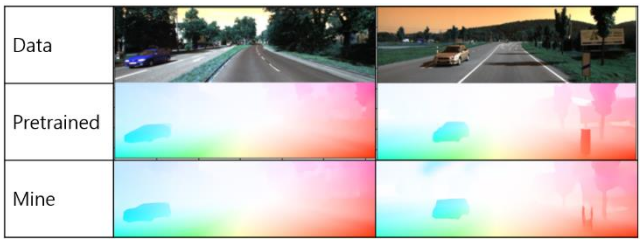
GRU-cell 을 통한 iteration 이 끝나면 N 개의 flow vector 가 저장된다. 마지막으로, 저자는 k 번째 flow 와 ground truth 데이터간 L1 distance 에 지수적으로 증가하는 가중치를 곱하여, 모두 더한 값을 loss function 으로 사용함으로써 iterate 된 횟수에 따라 중요도를 산정한다.

3. 구현결과 및 시도 개선안



RAFT 는 기존에 진행되었던 연구들이 비해 향상된 성능을 자랑한다. 저자가 제시하는 바에 따르면 parameter 수, 학습 반복 횟수, 시간을 비교했을 때 평균 end-point-error(epe)가 낮다고 한다.

다음은 논문 저자가 제시하는 pretrained-model 의 결과와 논문을 토대로 직접 구현한 결과이다.



위 표는 KITTI-datset 에 대해 각각 적용한 flow 를 이미지화 한 것이다. 먼저, Pretrained model 은 flying-chairs, flying-things, sintel dataset 에 대해 교차 학습과 fine-tuning 까지 완료된 상태로 epe 가 5.04 이다. 반면 직접 구현한 모델은 KITTI-dataset 에 대해서만 학습되어 10.54 의 epe 로 약 2 배가량 성능이 저하되었다.

해당 모델이 장점만 가지고 있는 것은 아니다. 가장 큰 단점은 학습에 recurrent neural network(GRU-cell)를 사용하여 deep-learning 의 최대 장점인 병렬구조의 연산을 활용하지 못한다는 점이다. GRU cell 에선 연산 시, 전 단계에서 업데이트된 hidden state 를 입력으로 이용하기 때문에 순차적으로 연산을 진행하고 학습 시간이 늘어날 수밖에 없다.

학습 시간이 지연되는 부분을 개선하기 위해선 iteration 횟수를 줄여야 한다. 하지만 Iteration 횟수는 결과의 정확도와 직결되기 때문에 flow filed 를 0 으로 초기화하지 않고 특정 값에서부터 학습을 시작하는 방법으로 정확도 또한 취하고자 했다.

해당 개선안에선 앞서 구한 correlation neighbor lookup term 이 flow 의 도착점과 주변 픽셀과의 연관도를 말하기 때문에, 해당 값의 크기가 큰 부분으로 이동할 것이라고 간주하여 flow field 를 초기화하여 학습을 시도했다. 하지만 0 이 아닌 다른 값으로 초기화하여 학습한 결과는 epe 가 14.57 로 그리 좋은 방안이 아니었다. 학습에서 optical flow 는 결국 반복적인 연산을

통해 제자리를 찾아가기 때문에 오히려 학습에 방해되는 요인으로 작동했다. 또한 loss function 에 초기 flow vector 에 대한 값도 포함되기 때문에 0 이 아닌 초기값은 오히려 loss 값의 수렴을 저해했다. 그러므로 initial flow filed 는 0 으로 초기화된 상태 혹은 수치화 된 데이터로 주어지는 것이 타당하다고 생각한다.

4. 활용방안

Optical flow 는 여러 프레임으로 이루어진 영상을 이해하는 다양한 연구에 활용될 수 있다. 가장 대표적인 것은 motion estimation 으로 두 영상 간 변환 관계(rotation, translation)를 말해준다. 즉, 프레임 사이의 변화정도에 직접적으로 관여할 수 있기 때문에 motion compensation 과도 연관되어 있고, 앞선 프레임의 정보를 사용하는 video compression 에 주요하게 사용된다.

다음으로, stereo dataset 을 사용하는 다른 학습 모델의 뼈대로 활용이 가능하다. Stereo data 는 보통 disparity 값 산출에 많이 사용되며 disparity 는 depth, normal 등 영상을 3D 로 분석할 수 있는 정보를 제공한다. 영상을 3 차원적으로 분석하는 것은 정확도를 특히 요구한다는 점에 있어서 학습의 기반을 RAFT 모델에 둔다는 것은 타당하다. 이외에도 video stabilization, human pose estimation 등 다양한 방면에서 optical flow 의 접목이 연구되고 있다.

마지막으로, robot level vision system 에서 유용하게 사용될 수 있다. 연속된 프레임에서 가장 큰 정보는 변화량에서 오는데 optical flow 가 장면간 변화량을 인지하는 역할을 하기 때문이다. 하지만, 아직 복잡한 상황을 분석하는 것에는 인간의 판단의 개입이 불가피한 상황이기 때문에 맥락(context)를 학습하는 것은 인공지능의 연구과제로 남아있고, optical flow 는 인공지능의 눈으로서 과제를 수행하는 기초가 될 것이다.

5. 결론

RAFT 는 optical flow 를 구하는 end-to-end 방식으로 학습하는 딥러닝 모델이다. 전체 픽셀의 feature map 에 대해 회귀적으로 정보의 정확도를 높이는 컨셉을 제시하여 parameter 수 및 학습 시간은 줄이고 성능은 높여 활용가능성이 크다는 강점을 가지고 있다. 또한 앞으로 여러 프레임이 연속된 영상을 기계 단계에서 이해하는데 있어 뼈대가 될 것이라 기대된다.

6. 감사의 글

본 연구는 2021 년도 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학지원사업의 결과로 수행되었음(No.2018-0-00213, SW 중심대학(건국대학교)).