CNN을 이용한 영상 프레임 보간과 Fine-tuning을 통한 게임 영상으로의 적용

박성훈 ○ 박범규 ○ 이용규
서울대학교 컴퓨터공학부
{zkxxkffks123, bumkyu00, dldydrb5360}@snu.ac.kr

요 약

영상 프레임 보간(Video Frame Interpolation)은 영상의 연속한 두 프레임 사이에 적절한 이미지 프레임을 삽입하여 영상의 프레임을 증가시키는 기술로 다양한 영상 처리 분야에서 사용될 수 있으며 활발한 연구가 진행 중인 분야이기도 하다. 전통적인 방식의 영상 프레임 보간은 단순히 두 프레임의 투명도를 조절하여 겹치거나(Overlapping), 물체의 이동을 포착해서 해당 부분의 픽셀을 합성하는 방향으로 이루어졌다. 그러나, 이러한 방식을 통해 보간된 이미지는 실체가 뚜렷하지 못하고 두 프레임을 단순하게 겹쳐 놓은 형태로 나타나는 단점이 존재한다. 따라서, 본 논문에서는 현재 state-of-the-art에 가깝다고 알려진 adaptive convolution을 이용한 영상 프레임 보간의 성능을 확인하고 fine-tuning을 통하여 게임 영상에 적용하고자 한다.

1. Introduction

영상 프레임 보간은 영상의 연속한 두 프레임을 적절하게 처리하여 두 프레임이 서로 이어지도록 하는 중간 이미지를 생성하여 삽입하는 방식으로 이루어진다. 이러한 영상 프레임 보간의 가장 전통적인 방식은 두 프레임의 투명도를 조절하여 그대로 겹치는 방식으로 구현되는데 이렇게 생성된 프레임은 원본 프레임의 픽셀 단위의 선형 보간에 불과하기 때문에 적절한 보간 방식이라고 보기 어렵다. 좀 더 적절한 보간 방식은 영상에서 움직임을 예측하여 이 움직임을 적절히 보간하는 것이다. 따라서, 영상에서 움직임을 예측하고 해당 부분의 픽셀을 합성하는 두 단계로 나누어진 보간 방식이 영상 프레임 보간 분야에서 주요한 하나의 방식으로 자리잡았다. 그러나, 이러한 보간 방식은 움직임 예측 기술에 상당부분 의존하는 경향이 있다. [1] 이에 따라, 심층 신경망을 이용하여 위의 두 단계를 합쳐 하나의 단계로 구성하는 방식에 대해 많은 연구가 이루어져 왔다. 그 중에서도 adaptive separable convolution을 이용한 방식이 심층 신경망을 이용하는 방식 중에서는 비교적 오래된 방식이나 현재 state-of-the-art에 가까운 성능을 보이고 있는 것으로 알려져 있다. [2] 하지만, 심층 신경망을 이용한 보간 방식 역시 일반적인 영상에 대한 적용에는 다소 어려움을 겪는 것으로 보인다. 영상의 프레임이 20fps (frame-per-second) 미만인 경우나 물체가 fps에 비해서 지나치게 빠른 경우에는 두 프레임 간 차이가 보간이 어려울 정도로 크게 나타나므로 상당히 흐릿한 이미지가 생성되게 된다.

따라서, 본 논문에서 는 adaptive separable convolution 기반의 pre-trained 모델의 성능을 15fps의 게임 영상에 대한 적용을 통해 확 인하고, 30fps 게임 영상 데이터셋을 이용한 fine-tuning을 통하여 게임 영상에 대한 보간 성능을 개선하고자 한다.

2. Background

Adaptive separable convolution은 커널 기반의 알고리즘으로 CNN을 이용하여 생성한 4개의 커널을 이용하여 두 프레임을 합성함으로써 중간 프레임을 생성한다. Figure1에서는 이 과정을 상세하게 도시하고 있다. 모델의 입력으로 연속한 두 프레임 I_1, I_2 이 주어졌을 때, 모델은 CNN Φ을 통하여 4개의 커널을 생성한다.

$$\langle K_{1,v}, K_{1,h}, K_{2,v}, K_{2,h} \rangle = \Phi(I_1, I_2) \qquad (1)$$

이 4개의 spatially-varying 커널을 이용하여 각 프레임 I_1, I_2을 처리한 결과를 합하여 다음과 같이 최종적인 중간 프레임 $\hat{I}$을 얻을 수 있다.

$$\hat{I} = \Phi(I_1, K_{1,v}, K_{1,h}) + \Phi(I_2, K_{2,v}, K_{2,h}) \qquad (2)$$

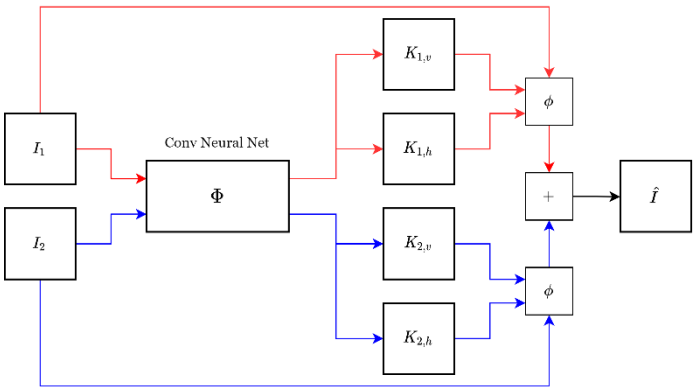


Figure 1. Overview of video frame interpolation with adaptive separable convolutions. [2]

이러한 커널 기반의 CNN을 이용한 보간 방식은 기존의 흐름 기반(flow-based)의 두 단계 보간 방식을 한 단계로

통합하면서 광학적 흐름(optical flow)가 정의되지 않는 오클루젼(occlusion)에 대해서 효과적인 프레임 생성을 가능하게 한다. 이때, 커널의 크기가 너무 작을 경우 배경적인(contextual) 요소를 제대로 고려하지 못할 수 있으며, 반대로 커널의 크기가 너무 클 경우 커널에 담긴 모든 값들의 상관계수(coefficient)들을 적절히 추정 하기가 어려워진다. 본 논문에서 인용한 pre-trained 모델에서는 이 커널의 크기를 51 pixels로 설정하는 것이 좋은 결과를 보임을 확인하였다. [2]

3. Method

2절에서 기술한 adaptive separable convolution 방식의 영상 프레임 보간은 일반적인 영상에 대해서 높은 성능을 보이나, 역동적이고 낮은 프레임을 가진 게임 영상들에 대해서는 상당히 흐릿한 이미지를 생성하는 결과를 보인다. 따라서, 본 논문에서는 pre-trained 모델의 fine-tuning을 통하여 게임 영상에 대한 성능을 개선하고자 한다.

3.1. Training Dataset

Fine-tuning을 위한 데이터셋으로는 FromSoftware Inc의 게임 'Elden Ring'의 30fps 영상을 사용하였다. 해당 게임에서 물체들의 움직임은 상당히 역동적이며 시점 변화로 인한 화면 전환을 수반하기 때문에 기존의 pre-trained 모델로는 적절한 결과를 얻기 어려웠다. 따라서, 본 논문의 취지에 적합하다고 판단하여 데이터셋으로 채택하였다.

3.2. Optimizer and Scheduler

학습을 위한 optimizer는 원 논문 [1]에서와 동일 한 AdaMax를 사용하였으며, learning rate는 0.0005, $\beta_1 = 0.9$, $\beta_2 = 0.999$를 사용하였다. 또한, scheduler를 도입하여 학습 중 learning rate가 서서히 감소하도록 설정하였다. Scheduler로는 stepLR을 사용하였으며 step size는 4, $\gamma = 0.7$로 설정하였다.

3.3. Loss function

모델의 학습을 위해서 사용될 수 있는 loss function 중 가장 간단한 함수는 다음과 같은 $\hat{I}$와 I_{gt} (ground truth image) 사이의 l1 norm이 될 것이다.

$$\mathcal{L}_1 = \left\| \hat{I} - I_{gt} \right\|_1 \qquad (3)$$

그러나, 이러한 단순 픽셀 값의 차이의 합인 l1 norm으로는 사진의 특징적인 요소들이나 배경적인 요소들을 반영할 수 없으므로 전체적으로 흐릿한 결과물을 내놓게 된다. 따라서, 개선된 loss function으로 perceptual loss를 이용한 다음의 loss function을 사용하였다. 이때 사용한 perceptual loss ψ는 VGG16의 pre-trained model에서 상위 6개의 layer를 사용한 loss로서 이미지의 특성(feature)을 추출하는 효과가 있다. 이 perceptual loss와 l1 loss 사이에 크기의 차이가 존재하기 때문에 이를 적절히 보정하기 위하여 계수 α, β를 도입하였다. [2][4]

$$\mathcal{L}_{p,1} = \alpha \left\| \psi(\hat{I}) - \psi(I_{gt}) \right\|_1 + \beta \left\| \hat{I} - I_{gt} \right\|_1 \qquad (4)$$

3.4. Similarity Measurement

모델의 성능을 측정하기 위한 척도로 본 논문은 생성된 중간 이미지 $\hat{I}$와 실제 이미지 I_{gt} 사이의 UQI (Universal Image Quality Index)를 사용하였다. 두 이미지 x, y에 대하여 UQI의 정의는 다음과 같다.

$$Q = \frac{4\sigma_{xy}\bar{x}\bar{y}}{\left(\sigma_x^2 + \sigma_y^2\right)\left[(\bar{x})^2 + (\bar{y})^2\right]} \qquad (5)$$

이때, 사용된 각 기호의 의미는 다음과 같다. [3]

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

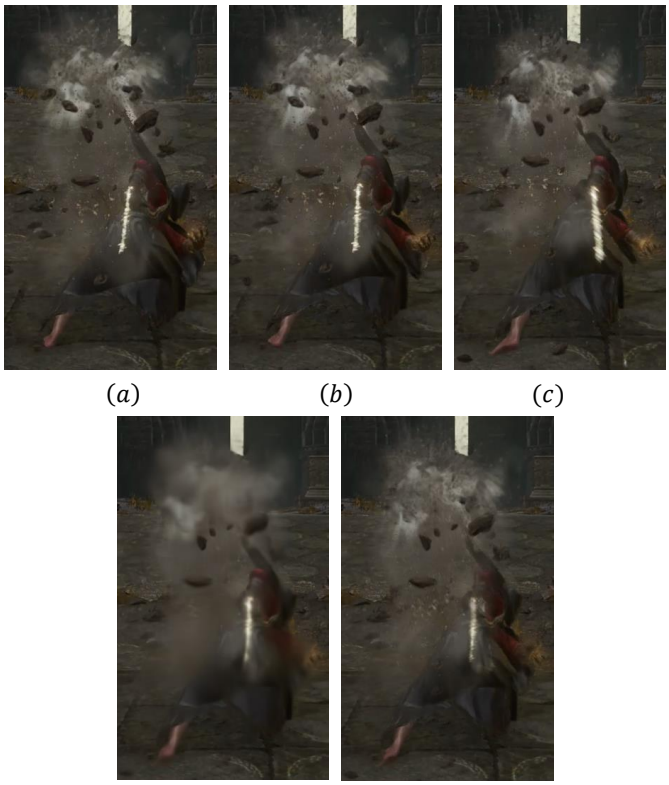
$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2, \qquad \sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N} (y_i - \bar{y})^2$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

4. Result

학습은 optimizer의 learning rate, schedular의 step size와 γ, 그리고 loss function에서의 α, β를 하이퍼 파라메터로 하여 이 값들을 조정하는 방식으로 수행되었다. 이때, $\alpha = 0.2$, $\beta = 0.001$일 때 가장 좋은 결과를 얻을 수 있었다. Train set, valid set, test set은 각각 402, 80, 360개의 이미지로 구성되었으며, 매 학습은 12 epoch 로 이루어졌다. 또한, 미니배치(mini batch)의 크기는 100으로 설정되었다.

Figure2의 (a), (b), (c)는 연속한 세 프레임을 나타내고 있다. 이 세 프레임 중에서 (a), (c)가 모델의 입력으로 주어지며 ground truth 이미지 (b)를 추정하여 생성하게 된다. (d)는 fine-tuning 이전의 원래 모델의 결과를, (e)는 fine-tuning된 모델의 결과를 나타내고 있다. (d)와 (e)를 비교해보면 (d)에서는 역동적으로 움직이는 것으로 보이는 상단의 먼지 부분이 상당히 흐려지면서 전체적인 화질을 낮추고 있지만 (e)에서는 이러한 흐림 현상이 상당히 감소하여 보다 선명한 모습을 보여주고 있으며, 원본 프레임의 질감을 더욱 잘 반영하여 드러내고 있다.



(a) (b) (c) (d) (e)

Figure 2. (a) First frame (b) Second frame (c) Third frame (d) Result of the original model (e) Result of Ours with fine-tuning.

	$Original$	$Fine-Tuned$
$Test\ set$ UQI	98.938%	98.907%

Table1. UQI of result for test set.

그러나, UQI의 경우에는 상기의 결과와 다소 상반된 결과를 나타내고 있다. Table1을 보면 원래의 모델이 fine-tuning된 모델보다 test set에 대해서 비교적 높은 UQI를 보이고 있다. 이러한 결과가 나타나는 이유는 흐릿한 이미지를 생성하는 것이 뚜렷한 이미지를 생성하는 것보다 일반적으로 낮은 평균 오차를 가질 가능성이 높기 때문이다. 이는 Figure4의 이미지 행렬들의 예를 통해서 확인할 수 있다.

$$I_{gt}=\begin{bmatrix}0&1&0\\0&1&0\\0&1&0\end{bmatrix}, I_{blur}=\begin{bmatrix}0&0.3&0.7\\0&0.3&0.7\\0&0.3&0.7\end{bmatrix}, I_{sharp}=\begin{bmatrix}0&0&1\\0&0&1\\0&0&1\end{bmatrix}$$

Figure 3. Image matrix examples.

I_{gt}가 ground truth 이미지라고 하면 이 이미지는 $col\,2$이 모두 1인 수직선이다. I_{blur} 는 흐릿한 이미지 결과값으로 수직선을 흐릿하게 추정하였으나 그 중심 위치를 $col\,2$가 아닌 $col\,3$으로 잘못 추정하였다. 마지막으로 I_{sharp}는 뚜렷한 이미지 결과값으로 수직선을 뚜렷하게 추정하였으나 역시 그 중심 위치를 $col\,2$가 아닌 $col\,3$으로 잘못 추정하였다. 이때, 각각의 추정 이미지와 ground truth 이미지의 차의 L1-norm을 계산하면 I_{blur} 의 경우 4.2, I_{sharp}의 경우 6으로 I_{sharp}가 더 크게 나타난다. 즉, 육안으로는 뚜렷한 이미지가 ground truth 이미지를 더 잘 추정하는 것으로 보이더라도, 뚜렷한 이미지는 픽셀 위치 오차에 따른 패널티를 더욱 크게 받으므로 전반적인 유사도가 낮아지는 결과를 가져오는 것으로 보인다. 따라서, test set에 대한 UQI의 저하는 보다 뚜렷한 이미지 생성에 따른 trade-off로 사료되며 만약 이미지 유사도의 기준에서 정확한 픽셀값보다 선명도에 가중치를 높게 두고 싶다면 이에 적합한 다른 유사도 척도의 도입이나, 흐릿함에 패널티를 주는 추가적인 계산이 필요할 것으로 생각된다.

5. Conclusion

본 논문은 adaptive separable convolution 기반의 pre-trained 모델의 성능을 15fps의 게임 영상에 대한 적용을 통해 확인하고, 30fps 게임 영상 데이터셋을 이용한 fine-tuning을 통하여 해당 게임의 영상에 대한 보간 성능의 개선을 확인하였다. 학습을 위한 loss function으로는 VGG perceptual loss를 통한 feature loss와 단순 L1-norm을 이용한 loss의 선형 조합을 사용하여 색 자체에 대한 정보와 배경적인 정보를 동시에 반영할 수 있게 하였다. 또한, 결과 이미지의 유사도 비교를 위한 척도로 UQI를 도입하였다. 그 결과 기존의 모델이 15fps의 영상에 대해서 다소 흐릿한 이미지를 생성하는 것에 반해 fine-tuning된 모델은 뚜렷한 이미지를 생성하는 것을 확인하였다. 그러나, test set에 대한 UQI는 fine-tuning된 모델이 다소 낮은 수치를 보이는 것을 확인하였는데 이는 흐릿한 이미지를 생성하는 것이 선명한 이미지를 생성하는 것보다 일반적으로 낮은 평균 오차를 가질 가능성이 높기 때문이다. 따라서, 이러한 UQI의 저하는 보다 뚜렷한 이미지 생성에 따른 trade-off로 보이며, 이를 해결하기 위해서는 다른 형태의 유사도 척도의 도입이 필요할 것으로 사료된다.

References

[1] Simon Niklaus, Long Mai, Feng Liu, “Video Frame Interpolation via Adaptive Convolution”, arXiv:1703.07514, 2017.

[2] Simon Niklaus, Long Mai, Oliver Vang, “Revisit-ing Adaptive Convolutions for Video Frame Inte-rpolation”, arXiv:2011.01280, 2020.

[3] Zhou Wang, Alan C. Bovik, “A Universal Image Quality Index”, IEEE signal processing letters, 2002.

[4] Simon Niklaus, Long Mai, Feng Liu, “Video Frame Interpolation via Adaptive Separable Convolution”, arXi v:1708.01692, 2017.