# 초저속 비디오 변환 서비스를 제공하는 웹 시스템

김동건 김도현 최해철
국립한밭대학교 정보통신공학과
dongeon94@naver.com, kd4278@naver.com, choihc@hanbat.ac.kr

# Web System providing Super slow Motion Video Transformation

Donggeon Gim, Dohyeon Kim and Haechul Choi
Dept. of Information and Communication Engineering, Hanbat National University

## 요약

최근 고주사율 디스플레이 시장 확대와 실감콘텐츠에 대한 요구에 따라, 높은 프레임율의 동영상 콘텐츠에 대한 관심이 증가하고 있다. 본 논문은 이용자의 비디오를 초슬로우 비디오로 변환해주는 웹 기반 서비스 시스템을 제안한다. 이는 사용자가 웹을 통해 비디오를 업로드하면, 딥러닝 기반의 비디오 프레임 보간 알고리즘을 이용하여 초고프레임율의 동영상으로 변환하며. 변환된 초저속 비디오를 웹을 통해 보여주거나 파일 포맷으로 제공한다. 제안 시스템은 복잡한 연산을 요구하는 딥러닝 네트워크 모듈과 사용자와의 상호작용을 위한 웹 페이지 모듈로 구성되었다. 프레임 보간을 위해서, State-of-the-art 기술인 딥러닝 기반의 Real-Time Intermediate Flow Estimation for Video Frame Interpolation 방법이 활용되었으며, 웹페이지는 HTML, CSS, Javascript, Flask를 사용하여 구축되었고, Flask를 활용하여 두 모듈이 연동되었다. 제안 웹 기반 시스템을 통해, 사용자는 딥러닝 네트워크 구동에 필요한 별도의 지식 없이 통신 자원만으로 고실감의 경험과 편의성을 제공받을 수 있다.

## 1. 서론

근래에 세계에 많은 핸드폰 제좌들이 국내의 삼성을 필두로 이번 2021년도 애플까지 그들의 스마트폰 제품에서 마케팅 포인트로 밀어붙히고 있는 것이 있다. 바로 120Hz 주사율의 디스플레이이다. 이 120Hz의 디스플레이는 기존 60Hz 디스플레이에 비교해 1초에 2배 이미지를 주사함으로써 사용자에게 발전된 고실감의 경험과 편의성을 제공한다. 또한 그 이전부터는 모니터를 필두로 고주사율 디스플레이를 시장이 확대되어지고 있는데, [1] DSCC에 따르면 스마트폰 제조에 있어 120Hz 주사율 디스플레이를 구현하는데 최적의 기술로 평가받는 LTPO 방식의 OLED 패널이 지금 대부분의 모바일 디스플레이 시장을 점유하고 있는 LTPS 방식의 OLED 패널의 점유율을 2023년 역전할 것으로 전망하고 있다. 이렇듯 고주사율 디스플레이의 보급 성장함에 따라 그로 인해 고주사율 디스플레이에 걸맞은 높은 FPS(Frame Per Second)의 동영상의 수요가 늘어날 것을 기대할 수 있다.

본 논문에서는 앞으로 높은 FPS 동영상의 수요를 해결할 수 있는 딥 러닝 기반 동영상 보간 웹 페이지 서비스를 소개하고자 한다. 발전한 딥 러닝 네트워크의 기술을 사용하여 기존 이미지와 이미지 간을 예측하여, 중간 이미지를 보간하고 기존의 일반적인 동영상도 높은 FPS의 동영상이나 초저속 비디오로 변환을 해주는 서비스이다. 또한 딥 러닝 네트워크의 높은 연산 수요를 고려해 웹을 통하여 서비스를 제공하기 때문에, 상대적으로 자원이 제한된 모바일 장치 등, On-device에서 동영상 데이터를 송신하여, Edge-sever에서 네트워크 연산을 수행하여 변환된 데이터를 사용자가 다운로드할 수 있게 구현하였다.

## 2. 본론

### 2.1 개발 환경

개발 환경 OS로는 리눅스를 사용하였으며, Docker와 Anaconda를 사용 개발 환경 세팅에 있어 비교적 유동성이 좋은 가상환경 소프트웨어 플랫폼을 사용하였다. 웹 페이지로는 일반적으로 사용되는 HTML, CSS 그리고 Javascript를 사용하였으며, 웹 프레임워크(Web Framework)로는 Flask를 사용하였다. Flask는 파이썬으로 작성된 마이크로 웹 프레임워크(Micro Web Framework)이다. 그렇기에 간단하게는 Python Web Framework라고 할 수 있다. 또한, HTML, CSS 그리고 Javascript 등과 같은 기존의 웹 프레임워크에서 똑같이 사용되어지는 언어들을 그대로 사용할 수 있다. 본 연구에서 제공할 웹 서비스는 딥 러닝 네트워크의 구동을 필요로 하기 때문에 프로그래밍 언어로는 파이썬(Python)을 사용한다. 그리하여 앞서 말한 파이썬으로 작성된 플라스크는 본 연구의 주목표가 되는 딥 러닝 네트워크 연동에 적합하다고 할 수 있다.

### 2.2 네트워크 선정

현재 딥 러닝 기반 동영상 보간 관련된 오픈 소스 코드들이 몇 가지 있다. 본 연구는 그 중에서 [2] DAIN(Depth-Aware Video Frame Interpolation), [3] NVIDIA Super SloMo, [4] RIFE(Real-Time Intermediate Flow Estimation for Video Frame Interpolation) 총 3가지 딥 러닝 네트워크의 성능 평가 및 비교를 통해서 웹 서비스 적합한 딥 러닝 네트워크를 하나를 선정하였다.

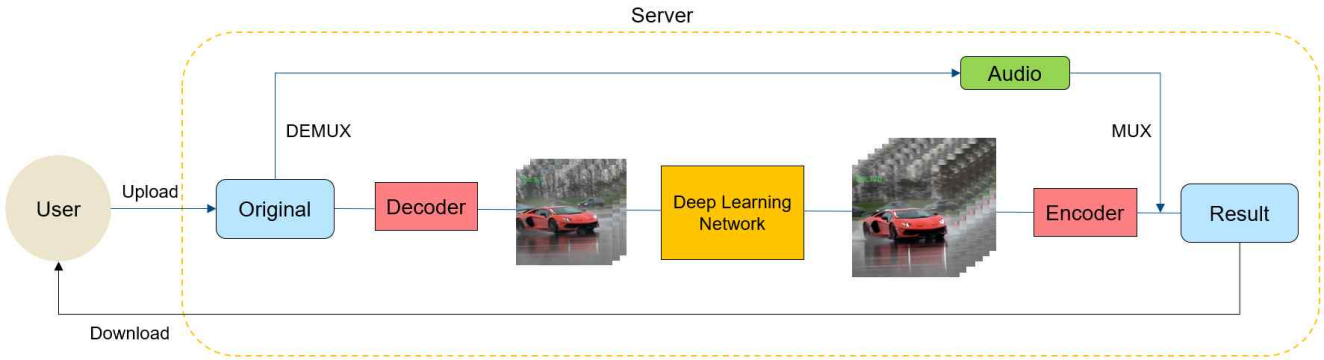


Fig 1. 시스템 구조

Original Super SloMo DAIN RIFE

| Sequences | Super SloMo | DAIN | RIFE |
|---|---|---|---|
| F1 cars* | 34.5363 | 37.8965 | 37.5285 |
| wrc car race | 31.1850 | 32.1687 | 32.0276 |
| Eagles | 33.8039 | 36.0577 | 35.6890 |
| Tennis | 36.9508 | 40.9091 | 40.7244 |
| **Averages(dB)** | **34.1190** | **36.7580** | **36.4923** |
| *Time(fps) | 1.38 | 0.55 | 11.43 |

Fig 2. 각 네트워크별 객관적 성능 평가

객관적 성능 비교 평가 기준으로는 PSNR과 시간당(sec) 보간할 수 있는 이미지 량(fps)을 사용하였다. 성능 비교 평가에 사용된 데이터는 FHD 해상도(1920 * 1080)로 이루어진 30fps 및 50fps 동영상을 사용하였고, 연산에 쓰인 그래픽카드로는 V100을 사용하였다. Fig 2을 보면 알 수 있듯이, 총 3가지의 네트워크 중 DAIN 네트워크의 PSNR 성능이 가장 좋게 평가되었다. 하지만 동작 시간의 관점에서 보면 RIFE 네트워크의 성능이 다른 네트워크에 비해 약 20배가량 성능이 뛰어난 것으로 볼 수 있었다. PSNR 값은 35 이상이 되면 사람이 인식할 수 있는 주관적 성능 비교 평가에서의 그 차이가 미비하고, 두 네트워크의 PSNR 값이 0.3 이하로 상대적으로 용인할 수 있다고 판단하여 동작 시간에서 20배의 차이로 많은 우위를 점하고 있는 RIFE 네트워크를 선정하기로 하였다.

### 2.3 시스템 구조

Fig 1.에서 보듯이 사용자는 자신이 변환시키고 싶은 동영상을 웹페이지를 통해서 업로드하게 된다. 업로드된 동영상은 서버에 업로드되어지고, 서버는 파일의 존재 여부를 확인하고 데이터가 확인되면 동영상 변환 연산에 들어간다. 처음 단계에서는 DEMUX를 통해서 데이터가 오디오 파일이 있다면 오디오 파일을 추출한다. 이 오디오 파일은 나중을 위하여 임시로 저장한다. 그 다음으로는 디코더(Decoder)를 통해 동영상을 여러 장의 이미지로 분할 하고, 그렇게 분할된 이미지를 차례대로 딥 러닝 네트워크 모듈에 삽입한다. Fig 3.을 보면 RIFE 딥 러닝 네트워크의 구조를 볼 수 있다. 이 네트워크는 차례로 입력된 $I_0$, $I_1$ 값을 통해 새로운 이미지를 만드는 반복 작업을 통해 보간된 이미지들을 기존 이미지들의 사이에 배열한다. 배열된 이미지들은 인코더(Encoder)를 통해 영상으로 다시 병합되고, 병합된 동영상에 이전에 추출한 오디오 파일을 MUX 시킨다. 결과물은 그대로 사용자에게 다운로드 될 수 있도록 제공한다.

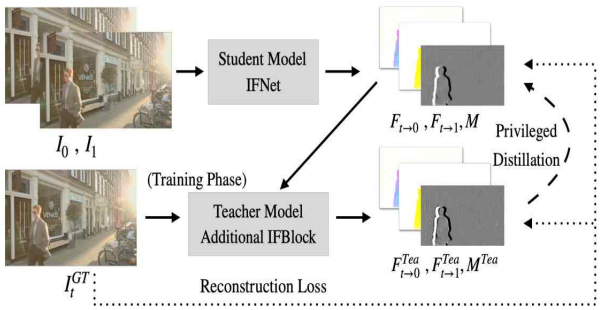


Fig 3. RIFE 네트워크 구조

## 3. 결과

사용자는 지정된 주소를 통해 웹페이지를 방문하여 보간하고 싶은 동영상을 업로드 한다. 이 때 사용자는 자신의 동영상을 페이지에 구성된 Rate Select Option을 통해 몇 배율로 동영상을 보간하고 싶은지 고르고, 그 다음으로는 재생속도라고도 할 수 있는 FPS 설정값을 입력할 수 있다. Rate 비율은 2배에서 16배까지 2의 제곱형태로 배율 조정할 수 있다. FPS는 최소 10FPS부터 최대 320FPS까지 조정할 수 있으며, 사용자가 값을 입력하지 않으면 자동으로 사용자가 올린 동영상에 알맞은 FPS값으로 제공한다. 그렇게 Upload 된 동영상은 서버에서 딥 러닝 네트워크를 통해 이미지를 보간하고, 보간된 이미지와 기존 이미지를 병합해 변환된 동영상을 만든다. 변환된 동영상이 완성되면 웹 페이지에서 redirection을 통해 사용자에게 변환된 동영상을 다운로드할 수 있는 페이지로 이동시키고, 사용자는 Download 버튼을 통해 변환된 동영상

을 다운로드한다. 이 때 다운로드 페이지는 사용자가 원본 동영상과 변환된 동영상의 차이점을 조금 더 직관적으로 경험할 수 있게 두 동영상을 통해 비교 동영상을 짧게 제작하여 페이지에서 실행해서 보여준다.

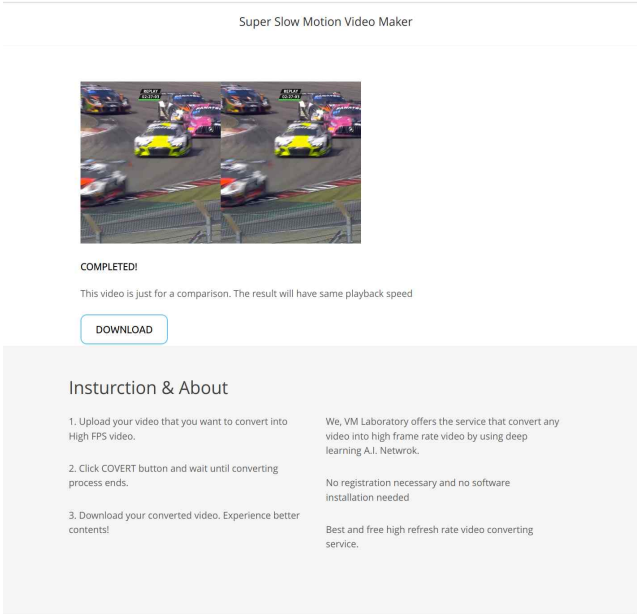


**FIg 4. 웹 페이지 결과화면**

## 4. 결론

앞으로 더 많은 고주사율 디스플레이의 공급이 확대됨에 따라 그에 따른 높은 FPS를 가진 동영상의 수요들도 늘어날 것으로 기대된다. 본 논문은 그런 시장의 기대수요를 예측함에 따라 현재 존재하는 낮은 FPS를 가진 동영상들도 더 나은 고실감의 경험과 편의성을 제공할 수 있는 높은 FPS를 가진 동영상으로 변환해줄 수 있는 서비스를 제안하고 있다.

딥 러닝 네트워크의 발전이 계속해서 빠른 속도로 성장하고 있음에 따라, 모바일 장치에서의 딥 러닝 네트워크의 연산 수요가 늘어날 것으로 예상하는데, 그 높은 연산 수요를 처리하기에는 모바일 장치의 한정적인 자원이 한계가 있다. 본 연구는 그런 제한 사항을 타개할 수 있는 웹 페이지와 딥 러닝 네트워크 간의 구동 연동에 성공함으로써 접근성이 좋은 웹 페이지를 통해 어디에서나 On-device 를 통해 이용할 수 있고, 그로 인해 사용자는 비교적 많은 양의 연산이 필요한 딥 러닝 네트워크의 기능을 통신 자원만 있다면 사용할 수 있게 되었다.

그렇기에, 본 논문에서의 동영상 보간 서비스와 딥 러닝 네트워크 연동은 미래에 사용자들에게 더 발전된 시각적 경험과 편의성을 제공해주는 것뿐만 아니라 앞으로의 제한된 모바일 장치에서의 딥 러닝 네트워크 이용 가능성을 크게 보여줬다고 할 수 있다.

## 참고 문헌

[1] DSCC, “Smartphone Report Reveals and Predicts Latest Smartphone Trends – High Refresh Rate Penetration Surging”

[2] W. Bao, W. Lai, C. Ma, X. Zhang, Z. Gao and M. Yang, "Depth-Aware Video Frame Interpolation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3698-3707, doi: 10.1109/CVPR.2019.00382.

[3] H. Jiang, D. Sun, V. Jampani, M. Yang, E. Learned-Miller and J. Kautz, "Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 9000-9008, doi: 10.1109/CVPR.2018.00938.

[4] Z. Huang, T. Zhang, W. Heng, B. Shi and S. Zhou, “RIFE: Real-Time Intermediate Flow Estimation for Video Frame Interpolation.”, arXiv. cs.CV, Aug. 2021.