Neural Network 기반 VCM의 성능평가

*박성욱, *이해림,†이주영, †정세윤, *조승현
*경남대학교, †한국전자통신연구원
e-mail : qso4545@gmail.com, lhl971212@naver.com, leejy1003@etri.re.kr,
jsy@etri.re.kr, scho@kyungnam.ac.kr

Performance Evaluation of VCM based on Neural Network

*Seonguk Park, *Haelim Lee, †Jooyoung Lee, †Se-Yoon Jeong and *Seunghyun Cho
*Kyungnam University, †ETRI

요약

최근 스마트시티, 자율 주행 자동차 등 기계에 의해 소비되는 영상 데이터의 양이 증가함에 따라 기계의 임무 수행 능력을 향상시킬 수 있는 압축기술이 필요하게 되었다. 그런데, 전통적 방식의 영상 코덱은 사람의 인지 화질 특성을 고려해 개발된 기술이기 때문에 기계의 임무 수행에 필수적인 정보 외에도 불필요한 정보가 존재한다. 따라서 사람이 아닌 기계의 임무 수행에 대해 효율적으로 영상을 압축하기 위한 비디오 코덱 기술이 필요하다. 이와 관련하여, 최근 MPEG에서 Video Coding for Machines라는 영상 압축기술에 대한 표준화가 논의되고 있다. 본 논문에서는 기계를 위한 영상 압축기술의 연구배경과 연구를 통해 전통적인 영상 압축 코덱 방식과 neural network 기반 압축 코덱 방식에 대해 각각의 방식이 머신비전 임무를 수행한 정확도를 기준으로 영상 압축성능을 비교해 효율적인 압축 코덱 방식에 대해 분석한다.

1. 서론

1-1. VCM 배경 및 표준화 진행 상황

최근에 Surveillance, Intelligent Transportation, Smart City, Intelligent Industry, Intelligent Content와 같은 다양한 산업 분야가 발전함에 따라 기계에 의해 소비되는 영상 데이터양이 증가하고 있다. 이에 반해, 현재 사용 중인 전통적인 영상 압축방식은 시청자가 인지하는 시각(Human Vision)의 특성을 고려해 개발된 기술이기에 불필요한 정보들을 포함하고 있어 기계 임무 수행에 비효율적이다. 따라서, 기계 임무 수행에 대해 효율적으로 영상을 압축하기 위한 비디오 코덱 기술에 관한 연구가 집중되었다.

멀티미디어 부호화 국제표준화 그룹인 MPEG(Moving Picture Experts Group)에서 2019년 7월 127차 회의부터 VCM(Video Coding for Machine) 기술이 논의되었다. VCM은 사람이 보는 시청자 시각 기준이 아닌 기계의 데이터 소비시각(Machine Vision)에 대한 기준의 영상 부호화 기술이다. 최근 업데이트된 2020년 10월 MPEG 132차 회의에서 이전 131차 회의 평가체제 초안[1]과 요구사항[2]이 수정되었다. 먼저, 평가체제로 VCM 압축성능을 평가하는 임무 수행 dataset 종류가 수정되었다. 하지만 dataset의 라이센스 문제가 정리되지 않아 MPEG 132차 회의에서 평가체제 초안이 완성되진 않았다. 또한, 요구사항으로 전자제품에 대한 consumer electronics 요구사항[3]이 추가되었다. 본 논문에서는 132차 MPEG 회의에서 도출된 VCM 평가체계[4]에 따라 기존의 코덱과 Neural Network 기반 코덱의 비교를 기계 비전 성능을 통해 분석한다.

VCM 압축성능 효율을 비교할 수 있는 머신비전 임무의 종류로 Object Detection, Object Segmentation, Object Tracking의 3가지 임무가 있다.

표 1. VCM 압축성능 평가 머신비전 임무 3가지

임무 수행 방법	Dataset	네트워크 구조
Object Detection	COCO, CityPersons, Open Images Compressed, FLIR Thermal Dataset	Faster R-CNN X101-FPN
Object Segmentation	COCO, Cityscapes, KITTI, Open Images Compressed	R50-FPN Cityscapes, PreMOVOS
Object Tracking	MOT20	JDE-1088×608

표 1에 보인 바와 같이, 성능평가에 사용되는 dataset으로 COCO, CityPersons, Open Images Compressed, FLIR Thermal Dataset, KITTI, MOT20 Dataset을 사용한다. 각각의 네트워크 구조는 Faster R-CNN X101-FPN, R50-FPN Cityscapes, PreMOVOS, JDE-1088×608를 사용한다.

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2020-0-00011, (전문연구실)기계를 위한 영상부호화 기술)
이 논문은 과학기술정보통신부 및 정보통신산업진흥원의 '고성능 컴퓨팅 지원' 사업으로부터 지원받아 수행하였음

본 논문에서는 머신비전 임무 3가지 중 간단한 Object Detection을 사용해 연구과제의 성능평가를 진행한다.

1-2 VCM 구조

최근까지 MPEG 회의에서 제안되어 분류된 VCM 모델 구조로 3가지가 도출되었다[5]. 원본 영상을 입력받아 압축하는 방법, 원본 영상의 features를 추출해 입력받아 압축하는 방법 그리고 앞의 두 가지 모두 입력받아 압축하는 방법이 있다. 본 논문에서는 입력을 원본 영상으로 받는 구조의 두 가지 영상 압축 코덱 성능 비교 연구를 한다.

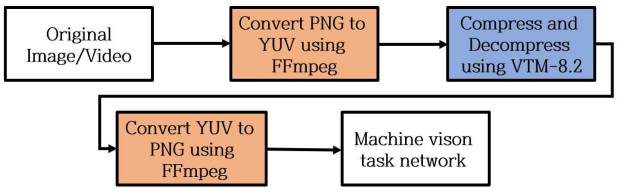

그림 1. 전통적인 영상 압축 코덱

그림 1은 원본 영상을 입력으로 받아 전통 방식(HEVC, VVC 등) 코덱을 사용해 출력을 내는 구조이다. 그림 1을 살펴보면 먼저, 전통적인 영상 압축 코덱은 입력 영상을 YUV 형식으로 받는다. 따라서 압축 코덱에 원본 영상을 입력하기 전 FFmpeg을 통해 PNG 형식인 원본 영상을 YUV로 변환한다. 변환된 영상을 압축 코덱에 입력 후 나온 결과물에 대해 머신비전 네트워크를 이용해 코덱 성능을 평가한다. 머신비전 수행 네트워크는 PNG형식을 입력으로 사용한다. 때문에, 성능평가를 위해 다시 FFmpeg을 통해 YUV형식을 PNG형식으로 변환 후 입력하여 사용한다.

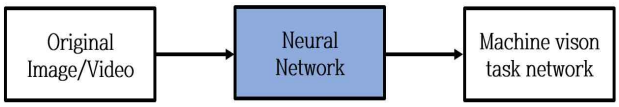

그림2. Neural Network를 이용한 압축 코덱

그림 2는 그림 1의 구조에서 전통적인 압축 코덱 방식을 neural network(NN) 기반 압축 코덱으로 사용해 출력을 내는 구조이다. 그림 2를 보면 NN 기반 압축 코덱은 PNG 형식의 원본 영상을 YUV 형식으로 변환하는 복잡한 과정 없이 원본 영상을 입력받아 압축하여 머신비전 수행 네트워크에 입력해 성능 평가하는 구조이다.

본 논문에서는 연구를 통해 전통적인 압축 코덱 방식과 NN 기반 압축 코덱 방식에 대해 각각의 방식이 머신비전 임무를 수행한 정확도를 기준으로 압축성능을 비교해 효율적인 영상 압축 코덱 방식이 무엇인지 보여준다.

2. 본론

2-1 Neural Network 기반 영상 압축 네트워크

NN 기반 영상 압축기술은 입력 영상에 대해서 엔트로피 양을 최소화하면서 복원이 완료된 영상에 대한 손실을 최소화하는 은닉 벡터(latent vector)를 만들어 내는 파라미터를 학습하는 방법으로 되어 있다. 또한, 그림 3과 같은 방식으로 은닉 벡터를 양자화(quantization)와 엔트로피 부호화(entropy encoding)를 하여 최종 비트스트림(bitstream)을 만들어 낸다.

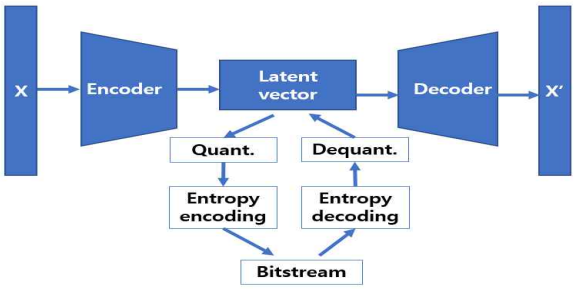

그림 3. NN 기반 압축 예시

본 논문에 앞서, scale hyperprior를 이용한 NN 기반 압축방식[6]을 구현한 공개 소프트웨어[7]에 대하여 성능평가가 이루어진 바가 있다[8]. NN 기반 압축성능평가를 위해 비교할 anchor로 전통적 압축방식의 최신 표준 VVC(Versatile Video Coding)의 테스트 모델인 VTM(VVC Test Model)-8.2[9] 버전을 사용하였다. 그리고 머신비전 임무 수행 네트워크로 Faster-RCNN X101-FPN을 사용하였다. [8]에서 도출된 성능은 머신의 임무 수행정확도 관점에서 NN 기반 압축방식이 VTM-8.2로 압축한 결과보다 비슷한 성능을 보이거나 더 높은 성능을 보인다고 보고되었다. 본 논문에서는 은닉 벡터의 효율적 부호화를 위해 hierarchical prior 네트워크와 joint auto regressive 방식의 압축방식을 이용한 모델[10-11]과 anchor[12]를 비교하여 성능평가를 진행했다.

2-2 머신비전 임무 수행 네트워크

본 논문은 VCM 압축성능 평가를 위한 기계의 임무 수행 네트워크 3가지 중 간단한 방법인 물체감지(object detection)를 수행한다. 물체감지란 영상의 객체를 식별하는 머신비전 기술이다. 영상 입력을 받아 영상 안의 물체를 인식해 박스(boundig box)를 그려 물체에 대해 위치 정보를 얻는 작업과 박스 안의 물체가 어떤 것인지 분류하는 작업 두 가지를 수행하는 기술이다. 현재까지 진행한 물체감지 연구에서 몇 가지 도출된 방법 중 본 논문에서는 Region Proposals 방법을 이용한 물체감지 임무를 수행한다. 사용하는 신경망 모델은 Detectron2[13]에 포함된 Faster R-CNN X101-FPN 네트워크이다. Detectron2은 Pytorch 기반의 Facebook AI Research의 차세대 오픈소스 객체 감지시스템이다. 본 논문에서 사용한 네트워크는 크게 세 개의 블록으로 구조를 설명할 수 있다.

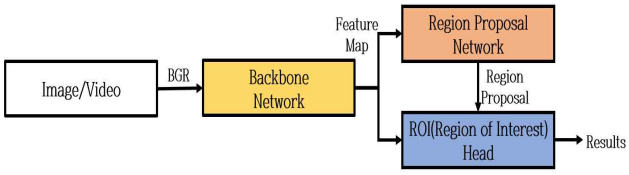

그림 4. Detectron2의 Faster R-CNN X101-FPN 구조

그림 4에 보인 바와 같이, Backbone Network에 입력 영상을 Blue,

Green, Red(BGR) 순서로 입력하고 입력된 영상을 다양한 scale(1/4, 1/8, 1/16, 1/32, 1/64)의 Feature Map으로 추출한다. 다음으로 RPN(Region Proposal Network)에 앞에 추출된 Feature Map을 입력받아 Region Proposal을 진행한다. 감지한 object에 대해 최대 1000개의 신뢰도 점수를 포함한 Region Proposal를 얻는다. 그 후 ROI(Region of Interest) Head는 앞의 두 가지 구조에 출력된 것을 각각 입력받아 feature map을 각 객체의 Region Proposal에 대한 크기로 자르고 왜곡하여 여러 고정 크기 feature로 만든다. 그다음 fully connected layer를 통해 미세조정 된 상자 위치 및 분류 결과를 얻어 NMS(Non-Maximum Suppression)를 통해 최대 100개의 proposal box를 출력한다. 앞의 도출된 결괏값으로 성능평가를 진행한다. 머신 성능측정 척도로 mAP(mean Average Precision)를 이용해 나타낼 수 있다. mAP는 압축률을 나타내는 bpp(bit per pixel)값에 따라 머신 비전 임무 수행 성능을 보여주는 척도이다[14].

2-3 실험방법

본 논문의 실험은 VCM 평가체계[2]에 정의된 실험방법을 따른다. 실험에서 VCM anchor와 비교할 구조는 그림 2와 같이 NN 기반 압축 코덱을 이용한 구조이다. NN 네트워크로는 CompressAI[11]에 구현된 mbt2018-mse-[1-8] 모델을 사용하였고, 숫자 1-8은 모델의 quality level을 의미한다. quality level이 낮을수록 큰 압축률에 해당된다. 머신 비전 네트워크로는 Detectron2[13]에 구현된 Faster R-CNN X101-FPN을 사용하였다. 데이터 세트로는 COCO 2017 Dataset 중 validation set 5000장을 사용하였다.

VCM Anchor는 입력 영상을 FFmpeg[15]을 이용하여 해상도를 100%, 75%, 50%, 25%로 각각 변환 후, 각 해상도 별로 QP (Quantization Parameter)를 22, 27, 32, 37, 42, 47로 바꿔가며 VTM-8.2로 부/복호화를 수행한다. 그 후 복호화가 완료된 영상을 원본 해상도로 재 변환 후 머신 비전 임무 수행 네트워크에 통과시켜 Object Detection에 대한 성능(mAP)을 추출한 후 각 해상도에 대한 QP별 bpp와 mAP를 측정하여 R-P(Rate-Performance) curve를 구한다.

본 논문에서 제시하는 실험은 입력 영상을 FFmpeg[15]을 이용하여 해상도를 100%, 75%, 50% 25%로 변환 후 각 해상도 별로 NN 기반 압축 네트워크를 사용한다. quality level을 1부터 8로 바꿔가며 부/복호화를 수행한다. 복호화가 완료된 이미지들을 원본 해상도로 재 변환 후 머신 비전 임무 수행 네트워크에 통과시켜 object detection에 대한 성능(mAP)을 추출 후 각 해상도에 대한 quality level 별 bpp와 mAP를 측정하여 R-P(Rate-Performance) curve를 구한다.

실험에서 나온 결과를 보고 VCM anchor와 비교하여 머신 비전 임무 수행 성능 기준으로 압축 효율을 비교하였을 때 anchor와 같이 인지 화질에 최적화된 코덱을 사용한 것과 NN 기반 코덱으로 사용했을 때 이점을 분석한다.

3. 실험 결과 및 분석

본 실험에서는 NN 기반 압축모델 [10]인 mbt2018모델과 VTM-8.2로 압축한 결과에 대해서 압축 효율 대비 머신의 임무 수행정확도를 비교하였다.

먼저 그림 5는 기존의 영상 압축 코덱인 VVC의 test model인 VTM-8.2과 NN 기반의 압축모델인 [6-10] 간의 압축 효율 대비 객관적 화질 성능 지표인 PSNR(Peak Signal-to-Noise Ratio)의 그래프 결과이다. VTM-8.2로 압축한 이미지에 대한 PSNR 결과가 NN 기반의 압축모델보다 압축 효율이 더 좋은 것을 확인했다.

그림 6은 해상도가 100%인 이미지들에 대해서 VTM-8.2로 압축한 성능과 NN로 압축한 성능에 대해서 bpp와 머신의 임무 수행정확도 mAP를 비교한 결과이다.

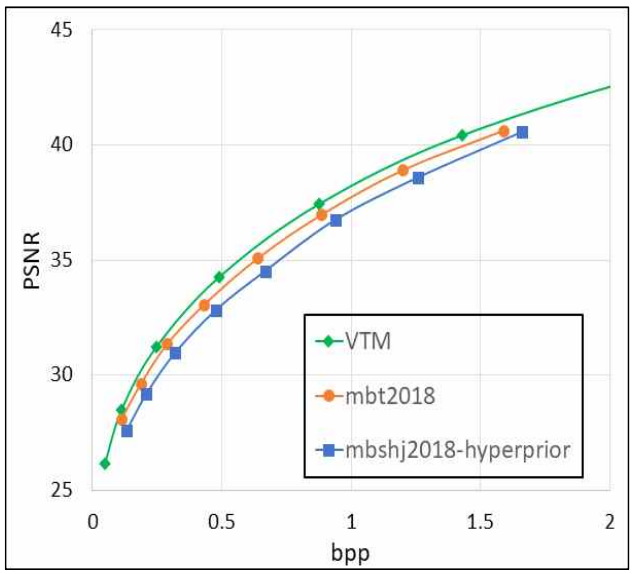

그림 5. 압축모델별 bpp 대비 PSNR 결과

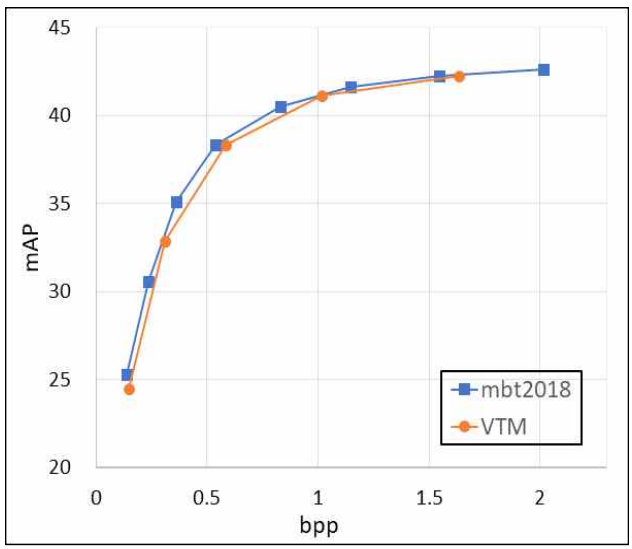

그림 6. VTM과 mbt2018 압축모델의 물체 검출 R-P curves 비교

그림 5와 그림 6을 비교하면 두 가지의 방법으로 압축했을 때, PSNR 측면에서는 전통적 영상 압축방식인 VTM-8.2가 더 높은 성능을 보였지만, 머신의 임무 수행 정확도 측면은 NN 기반의 압축모델이 더 높은 성능을 보이거나 비슷한 성능을 보인다. VVC와 같은 전통적인 압축방

식의 특징으로는 블록 단위의 압축을 하는 방식이므로 복원된 이미지에 블록 형태의 노이즈(blocking noise)가 NN 기반의 압축기술보다 많이 발생하여 머신의 임무 수행 정확도에 대한 성능이 낮게 나온다고 판단된다.

그림 7은 동일한 데이터 세트에 대한 다양한 해상도(100%, 75%, 50%, 25%)를 NN 기반 코덱으로 압축 후 머신의 임무 수행 정확도 결과를 보여준다. 원본 크기로 압축한 이미지보다 해상도를 줄여 압축한 이미지가 압축 효율 대비 머신의 임무 수행 정확도 성능이 몇몇 구간에서 더 높은 결과를 보여주고 있다.

그림 8은 그림 7의 결과에 대해서 최적의 성능을 보이는 해상도와 Quality level 쌍을 보여주는 Pareto Front 그래프이다. 머신 비전의 입장에서는 최대의 성능을 내기 위해서 최적의 해상도와 Quality level 쌍을 찾는 것이 중요하다.

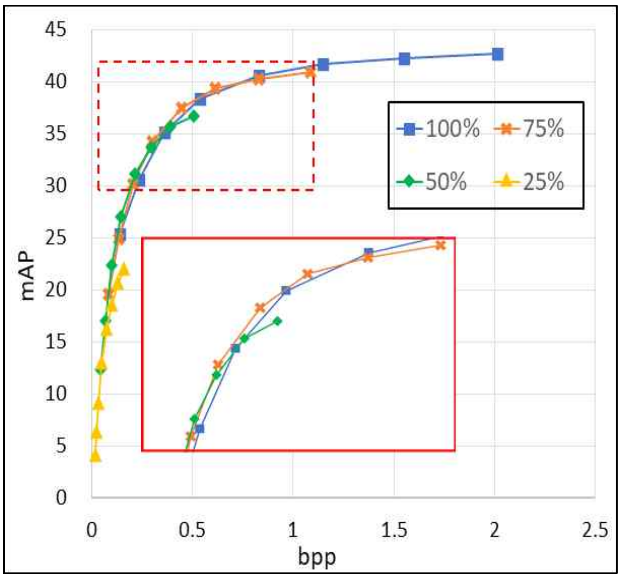

그림 7. 멀티 해상도 NN 기반 압축 후 물체 검출 R-P curves 비교

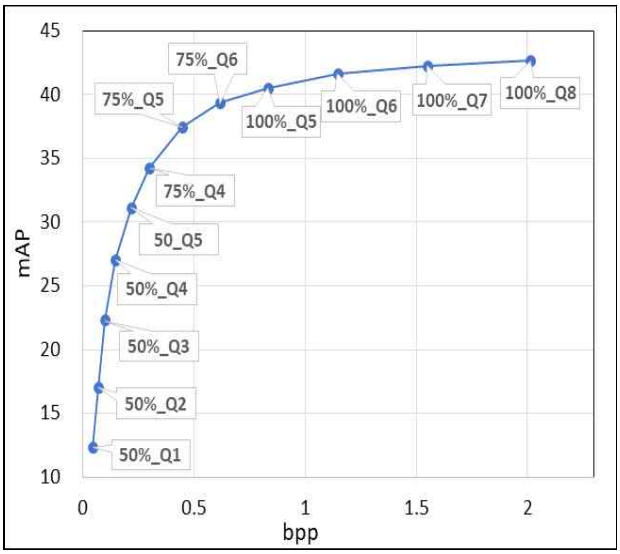

그림 8. 그림 7의 Pareto Front

4. 결론 및 향후 연구

본 논문에서는 사람이 보는 인지시각 기준이 아닌 기계의 시각에 대한 기준의 영상 부호화 기술인 VCM에 대해서 다루고 있다. 기존의 전통적인 비디오 압축 코덱으로 압축 후 머신의 임무 수행 정확도에 대한 결과와 NN 기반 압축 코덱을 사용 후 머신의 임무 수행정확도 결과를 분석했다. 전통적 영상 압축 코덱과 NN 기반 압축 코덱을 객관적 화질(PSNR) 측면에서 비교하였을 때 전통적 압축 코덱의 성능이 높지만, 머신의 임무 수행정확도 측면에서는 NN 기반 영상 압축 방식이 전통적 영상 압축 코덱에 비해 더 좋은 성능 또는 비슷한 성능을 보인다. 또한, 머신의 임무 수행 정확도는 같은 영상에 대해서 해상도의 크기와 압축률에 따라 최적의 성능을 낼 수 있다.

향후 연구로는 영상 압축 코덱 및 머신의 임무 수행 네트워크가 NN 기반으로 되어 있어 두 개의 NN을 joint optimization 하였을 때 머신의 임무 수행정확도 결과의 향상 가능성을 검증해 볼 필요가 있다.

참고문헌

[1] MPEG w19506, “Use cases and draft requirements for Video Coding for Machines”, Online, June, 2020.
[2] MPEG w19507, “VCM Evaluation Framework for Video
[3] MPEG MDS19841, “Use cases and draft requirements for Video Coding for Machines”, Online, October, 2020.
[4] MPEG MDS19843, “Draft of Call for Evidence for Video Coding for Machines”, Online, October, 2020.
[5] 최창균 외, “VCM 구조 분석 및 VVC 기반 Feature 부호화 성능 분석”,제30회 신호처리 합동학술대회, 2020년 9월.
[6] Balle J, Minnen D, Singh S, et al. "Variational image compression with a scale hyperprior" international conference on learning representations, 2018.
[7] https://github.com/tensorflow/compression/tree/master/examples
[8] MPEG m54366, “[VCM] Coding Experiments fo End-to-end Compression Network in VCM”, Online, June, 2020.
[9] https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/releases/VTM-8.2
[10] D. Minnen, J. Ballé, and G. Toderici, ‘Joint Autoregressive and Hierarchical Priors for Learned Image Compression’, 2018
[11] https://github.com/InterDigitalInc/CompressAI
[12] MPEG m54349, “[VCM] Anchor generation results for object detection on COCO dataset”, June, 2020.
[13] https://github.com/facebookresearch/detectron2
[14] 반현민 외, “VCM Anchor 성능평가 및 분석”, 제30회 신호처리 합동학술대회, 2020년 9월.
[15] https://ffmpeg.org/