ETRI

2018
Electronics and
Telecommunications
Trends

정지영상 및 동영상 인지화질 측정 기술 동향

Technology Trends on Image/Video Perceptual Quality Assessment

이대열 (D.Y. Lee, daelee711@etri.re.kr) 실감 AV 연구그룹 연구원
김종호 (J.H. Kim, pooney@etri.re.kr) 실감 AV 연구그룹 선임연구원
정세윤 (S.Y. Jeong, jsy@etri.re.kr) 실감 AV 연구그룹 책임연구원/PL
조승현 (S.H. Cho, shcho@etri.re.kr) 실감 AV 연구그룹 책임연구원
김휘용 (H.Y. Kim, hykim5@etri.re.kr) 실감 AV 연구그룹 책임연구원/그룹장
최진수 (J.S. Choi, jschoi@etri.re.kr) 실감 AV 연구그룹 책임연구원/PL

Assessment technologies regarding the perceptual quality of images and videos have been receiving significant attention, as they serve as essential tools for monitoring and improving the quality of various media services. In this paper, we review the technology trends of recent studies on the perceptual quality assessment of images and videos, and discuss the future direction of this research field.

* DOI: 10.22648/ETRI.2018.J.330302

*이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임[No. 2017-0-00072, 초실감 테라미디어를 위한 AV 부호화 및 LF 미디어 원천기술 개발].

Ⅰ. 서론

디지털 영상은 획득 단계에서부터 시청자에게 도달하기까지 다양한 단계를 거치며, 이 과정 중 캡처 장비의 물리적 한계 혹은 네트워크의 한정된 대역폭 등으로 인해 정보가 손실되거나 왜곡이 유입될 수 있다. 예를 들어, 획득 단계에서는 카메라 센서 잡음, 노출 제어, 카메라 모션 등으로 인해 영상에 왜곡이 유입될 수 있고 획득 이후에는 저장 용량 혹은 전송 대역폭을 고려한 영상 압축 기술로 인해 왜곡이 발생할 수 있다. 또, 영상이 채널을 통해 전송되는 과정에서도 정보 손실이 일어날 수 있으며, 최종 렌더링 단의 디스플레이 장치로 인한 왜곡이 유발될 수도 있다.

상기 다양한 단계에서 발생하는 왜곡 및 정보 손실은 영상의 시청 화질에 영향을 미치기 때문에, 멀티미디어 서비스 제공에 있어 각 단계의 왜곡 정도를 파악하고 개선하기 위한 노력은 매우 중요하다고 할 수 있다.

ITU-T VCEG 및 ISO/IEC MPEG 등의 국제비디오 표준화 단체의 경우 Mean Squared Error(MSE), Peak Signal-to-Noise Ratio(PSNR), Structural Similarity Index(SSIM) 등을 포함한 다양한 화질 지표를 영상 압축 기술의 성능 평가 및 최적화 도구로써 활용하고 있다. Netflix, Amazon, Hulu 등의 OTT 사업자들의 경우 학계와 활발한 교류를 통해 네트워크 스트리밍 환경에서의 체감 품질(QoE: Quality of Experience) 예측 및 개선 연구를 진행 중에 있다[1], [2]. ITS(Institute for Telecommunication Science) 산하 영상 화질 전문가 단체인 VQEG(Video Quality Experts Group)은 명암비와 색역대를 비약적으로 향상시킨 High Dynamic Range (HDR)/Wide Color Gamut(WCG) 영상 또는 Augmented Reality(AR)/Virtual Reality(VR) 등의 차세대 몰입형 영상 등에 대한 화질 연구를 진행 중이다. 이처럼 멀티미디어 서비스 다양한 단계에서의 왜곡 종류 및 인지화질 영향도 분석 연구는 활발히 이루어지고 있으며, 연구 동향을 살펴보면 사람이 실제로 인지하는 화질에 대한 이해를 기반으로 미디어 서비스의 질을 높이려는 접근이 이루어짐을 알 수 있다.

본고에서는 상기 영상 인지화질 연구의 기본 개념에 대해서 알아보고, 다양한 정지영상 및 동영상 화질 측정 기술 동향 분석 및 향후 기술 발전 방향에 대해 논해보고자 한다.

Ⅱ. 인지화질 측정 기술 개요

인지화질 측정 기술은 영상으로부터 추출한 다양한 시간적, 공간적 특징정보(feature)를 활용하여 영상의 주관적 화질을 예측하는 기술이다. 기존 제안된 MSE 혹은 PSNR 등의 객관적 화질 지표의 경우 영상의 픽셀 값 왜곡 정도라는 직관적인 수치를 통해 화질을 정량화하지만, 인지화질을 정확히 반영하는 데에는 한계가 있었다[3]. 이에 UT Austin, Tampere University, Shizuoka University, VQEG 등에서는 여러 영상에 대한 대규모 주관적 화질 평가를 수행하여 각 영상에 대한 인지화질 점수를 수집하고 해당 데이터베이스를 공개하였다[4]-[11]. 상기 데이터베이스들을 토대로 다양한 정지영상 및 동영상 인지화질 지표들이 연구되었으며, 대부분의 기술들이 상기 데이터베이스들로 학습을 진행하거나 예측 정확도를 검증한다. 인지화질 측정 기술의 분류, 활용 가능한 데이터베이스 종류, 성능 측정 방법 관련 상세 사항은 아래에 후술된다.

1. 인지화질 측정 기술 분류

인지화질 측정 기술은 크게 전기준법(FR: Full Reference), 반기준법(RR: Reduced Reference), 그리고 무기준법(NR: No Reference)으로 분류할 수 있다. 전기준법 기술은 왜곡 영상과 더불어 원본 영상을 함께 가지고 있는 경우, 원본과 왜곡 영상을 직접 비교하는 방식으로 대부분 기술에서 차용하는 방식이다.

하지만 원본영상 전체를 가지고 있지 않은 경우에는 사용이 제한되는 단점이 있다. 반기준법의 경우 영상의 일부 특징만을 전송하여 이를 토대로 영상의 화질을 예측하는 기술로, 영상 화질 예측을 위해 필요한 데이터양이 적어 비용 효율적이다. 무기준법의 경우 왜곡 영상만을 가지고 화질 수준을 예측해야 하므로 정확도가 높은 기술을 만들기가 매우 어렵지만, 보다 넓은 범위의 응용에 활용될 수 있다는 점에서 장점을 가진다.

2. 인지화질 데이터베이스 종류

정지영상 및 동영상 인지화질 연구에 주로 활용되는 데이터베이스는 〈표 1〉, 〈표 2〉와 같다. 대부분의 데이터베이스들은 압축 왜곡 혹은 네트워크로 인한 전송 에러(TE: Transmission Error)가 포함된 영상의 화질을 다루며, 각 영상에 대한 인지화질 점수를 Mean Opinion Score(MOS) 혹은 Differential Mean Opinion Score (DMOS) 등의 형태로 나타낸다. 여기서 MOS/DMOS 는 여러 사람의 화질점수의 평균을 구한 것으로, MOS의 경우 수치가 높을수록 화질이 좋음을 의미하며, DMOS의 경우 원본영상 대비 MOS값의 차이를 나타내는 수치이므로, 수치가 낮을수록 화질이 좋음을 의미한다. 대부분 데이터베이스에서는 각 영상의 MOS/DMOS값과 더불어 표준편차 수치도 함께 제공하고 있으므로, 각 영상에 대한 사람들의 화질 평가 성향이 얼마나 다양했는지를 가늠할 수 있다. 화질평가 방법으로는 크게 Absolute Category Rating(ACR), Double Stimulus Impairment Scale(DSIS), Subjective Assessment Methodology for Video Quality(SAMVIQ) 등의 방법이 많이 사용되는데 각 방법에 대한 간략한 설명은 다음과 같다. ACR의 경우 평가 대상 영상을 한번만 보여주며, 그에 대한 화질 점수를 매기게 하는 방식이다. ACR 방식을 차용하면서도 DMOS 형태의 점수일 경우가 있는데, 이때는 원본 영상을 평가 영상들 사이에 무작위로 섞어서 평가한 후, 원본 영상과 평가 영상 점수의 차이를 계산한 경우이다. DSIS는 영상 압축기술 연구 분야에서 활발히 쓰이는 주관적 화질 방법으로, 원본 영상을 먼저 실험참가자에게 보여준 후, 평가 영상을 보여주며 원본 대비 화질을 평가하게 하는 방식으로, 사용자가 원본 영상의 화질 수준에 대해서 알고 있는 상태에서 평가 영상의 화질을 측정한다는 점이 특징이다. SAMVIQ는 원본 영상을 명시적으로 알려준 후 다양한 평가 대상 영상의 화질을 측정하는데, 영상을 정지하거나 재시청할 수 있으며, 점수를 재조정할 수 있는 것이 특징이다. 상기 주관적 화질 평가 방법 관련 상세 사항은 [3], [12] 에 안내되어 있다.

〈표 1〉 정지영상 인지화질 데이터베이스

데이터베이스	원본영상	왜곡종류	화질평가 방법	점수 형태
LIVE Image[4]	29개	JPEG, JPEG2000, Gaussian blur, white noise(WN), transmission error(TE)	ACR	DMOS(0~100)
TID 2008[5]	25개		Custom	MOS(0~9)
TID 2013[6]	25개		Custom	MOS(0~9)
CSIQ Image[7]	30개		Custom	DMOS(0~1)

〈표 2〉 동영상 인지화질 데이터베이스

데이터베이스	원본영상	해상도	왜곡종류	화질평가 방법	점수 형태
LIVE Video[8]	10개	768×432	AVC, MPEG2, TE	ACR	DMOS(0~100)
CSIQ Video[9]	12개	832×480	AVC, HEVC, WN, TE	SAMVIQ	DMOS(0~100)
NFLX Public[10]	9개	1,920×1,080	AVC, resizing	DSIS	DMOS(0~100)
VQEG HD3[11]	9개	1,920×1,080	AVC, MPEG2, TE	ACR	DMOS(0~5)

3. 인지화질 측정 기술 성능 지표

정지영상 및 동영상 화질 측정 기술의 성능은 〈표 1〉, 〈표 2〉 등의 데이터베이스에서 제공되는 실제 인지화질 데이터와 얼마나 연관성이 높은지를 통해서 계산한다. 대표적으로 쓰이는 지표로는 Pearson Linear Correlation Coefficient(PLCC), Spearman's Rank Ordered Correlation Coefficient(SROCC), Root Mean Squared Error(RMSE) 등이 있다. 각 지표의 의미에 대해서 살펴보면, PLCC는 화질 측정 기술과 실제 인지화질 데이터 간의 전반적인 선형 관계를 평가하며, SROCC는 화질 측정 기술과 인지화질 데이터 간의 순위 유지 정도를 평가한다. 두 지표의 경우 절댓값 기준 0에서 1 사이의 값을 가지며, 값이 1에 가까울수록 실제 인지화질 데이터와 통계적 유사성이 높음을 의미한다. RMSE의 경우 화질 측정 기술이 예측한 MOS/DMOS값과 실제 MOS/DMOS 데이터값의 차이 정도를 계산하므로 0에 가까울수록 성능이 높음을 의미한다[3]. 대부분의 정지영상 및 동영상 화질 측정 기술들의 경우 상기의 세 지표를 모두 활용하여 실제 인지화질 데이터와의 통계적 유사성을 평가하는 편이다.

Ⅲ. 정지영상 인지화질 측정 기술 동향

본 절에서는 인지화질 측정 기술 중, 정지영상의 화질 측정 기술 동향을 알아보도록 하겠다. 정지영상 화질 측정 기술 중에는 전기준법 기술인 MSE와 PSNR 등이 일찍이 제안되었는데, 해당 지표들이 가지는 직관적인 의미와 계산 용이성 등으로 인해 아직도 보편적으로 쓰이고 있으나, 여러 연구들을 통해서 밝혀진 바와 같이 영상의 인지화질을 반영하는 데에는 한계가 분명히 있는 지표들이다[3]. UT Austin에서는 단순한 픽셀의 밝기값 차이뿐만 아니라 영상 내 명암 및 구조의 보존 정도가 인지화질과 깊게 연관되어 있음에서 착안, 영상처리 분야에서 널리 알려진 SSIM 지표를 개발하여 인지화질 예측 정확도 측면에서 유의미한 개선을 이끌어 냈고, 이를 필두로 하여 지금까지 다양한 종류의 정지영상 화질 측정 지표들이 연구되었다. 최근에는 다양한 기계학습 기반 기술들이 영상처리 기술 분야에 적용되면서, 정지영상 화질 측정 기술에도 기계학습이 적용된 사례들을 볼 수 있다[13], [14].

특히 정지영상 화질 측정 분야의 경우, 동영상보다 다루어야 할 정보량이 적은 편이기 때문에 별도의 도메인 지식을 활용하지 않고도, 정지영상과 인지화질 점수 데이터만을 활용하여 신경망이 스스로 그 상관관계를 학습하게끔 하는 데이터-드리븐(Data-Driven) 기법 많이 제안되었으며, 해당 기술들의 높은 예측 성능으로 인해 데이터-드리븐 기법도 엄연히 하나의 카테고리로 자리잡게 되었다. 본 절에서는 영상 신호처리 지식을 기반으로 직접 추출한 특징정보(Hand-Crafted Feature)를 활용하여 화질을 측정하는 모델 기반(Model-Based) 기법과 데이터-드리븐 기법의 대표 기술들을 알아보고, 각 기법의 특징에 대해서 살펴보도록 하겠다.

1. 모델 기반(Model-Based) 기법

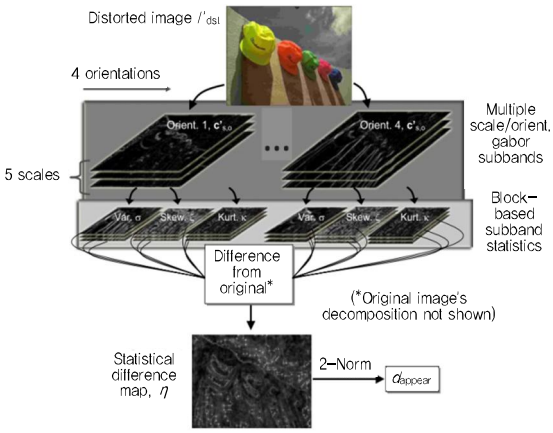
모델 기반 기법은 영상의 다양한 공간적 특징정보를 추출하고, 이를 이용하여 왜곡 정도를 파악하는 기법이다. 모델 기반 기법에도 Support Vector Regression (SVR), 혹은 인공 신경망(Neural Network) 등의 기계학습 기술들이 사용되는 사례가 있으나, 영상처리 혹은 사람의 시각시스템(HVS: Human Visual System) 등의 지식을 기반으로 추출한 특징정보에 기계학습을 적용한다는 점에서 순수 데이터-드리븐 기법과 차별성을 가진다. 모델 기반 기법 중 대표적인 기술 몇 가지를 살펴보면, Multiscale SSIM(MS-SSIM), Most Apparent Distortion(MAD), 그리고 Blind Image Integrity Notator (BLIIND) 등이 있다.

MS-SSIM은 기존의 SSIM 기술이 디스플레이 해상도 혹은 시청 거리 등의 시청 환경에 대한 고려가 없음에 주목하여, 멀티스케일(Multi-scale)에서의 SSIM을 측정하고 시청환경에 따라 적응적으로 가중치 합하여 화질을 측정할 수 있도록 만든 전기준법 지표이다[15]. MS-SSIM 기술이 활용하는 특징정보는 SSIM과 동일하게 영상 블록의 밝기, 명암, 및 구조의 유사도로 모두 공간 도메인(Spatial domain) 상에서 계산 가능한 특징정보들이며, 복잡도가 크게 높지 않으면서도 예측 정확도가 준수하여 다양한 정지영상 화질 측정 연구의 비교 대상으로 활용된다.

MAD는 HVS의 특성을 적극적으로 고려한 전기준법 지표로, 영상의 전체적인 화질에 따라 사람이 화질을 평가하는 전략이 다를 것이라는 가정으로 지표를 설계하였다[16]. 예를 들어 고화질의 영역에서는 HVS가 영상에 존재하는 왜곡 정도에 따라 화질을 평가할 것이며, 저화질의 영역에서는 HVS가 산개해있는 왜곡 가운데서 영상의 내용이 어느 정도 식별 가능한가를 통해 화질을 평가할 것이다. 고화질 영역에서는, 먼저 영상을 인지 밝기/명암(Perceived luminance/Contrast)으로 변환한 뒤 영상의 왜곡 정도를 계산한다. 수식으로는 아래와 같이 표현된다.

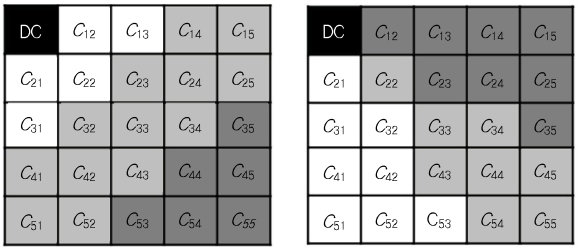
$$\hat{I} \approx (I^{2.2})^{\frac{1}{3}} \times CSF.$$

디지털 영상은 각 픽셀의 밝기를 8bit 혹은 16bit로 표현할 수 있는 정수 형태로 가지고 있는데, I란 이러한 디지털 영상을 의미하며, I에 적용되는 2.2승은 실제 디스플레이에서의 밝기 정도로 표현해주기 위한 sRGB 디스플레이 감마(gamma)값이다. 디스플레이에서의 밝기 값은 그 뒤에 적용되는 1/3승 계산을 통해 사람의 눈에 들어오는 인지 밝기로 변환이 된다. 그 이후 CSF(Contrast Sensitivity Function) 함수 적용을 통해 사람이 대비(contrast)별로 민감하게 느끼는 주파수 영역을 판별한다. 이렇게 최종 계산된 $\hat{I}$는 디지털 영상 I를 사람이 실제 인지하는 영상 정보로 변환한 것이며, 원본 영상과 왜곡 영상 간의 인지 밝깃값 $\hat{I}$에 대한 차이 정도를 계산함으로써 인지적 관점에서의 왜곡 정도를 계산 한다. 상기의 과정을 통해 계산된 고화질 영역에서의 왜곡 정도는 d_{detect}라고 칭한다. 저화질 영역에서는 (그림 1)과 같은 계산을 수행하는데, 주로 영상의 왜곡이 산재해 있는 가운데서도 영상의 구조 등을 식별 가능한지 등을 보기 위해 네 방향의 Gabor 필터를 다섯 해상도 단계에 대해서 적용하며, 적용된 결괏값에 대해서는 블록 별로 표준편차, 비대칭도(Skewness), 첨도(Kurtosis) 등을 계산하여 원본 및 왜곡 영상 간의 특성 차이를 분석한다. 이렇게 계산된 차이 값은 d_{appear}라고 칭한다. 최종 MAD값은 d_{detect}와 d_{appear} 를 함께 고려하여 계산하며, 영상의 전체적 화질이 좋을수록, d_{detect}에 더 큰 가중치를 주는 방식으로 계산한다. MAD 지표 또한 MS-SSIM 과 더불어 정지영상 화질 측정 분야에서 기준(Anchor)이 되는 기술이며, HVS에 중심을 두는 화질 측정 기술들의 근간이 되는 기술이다.



(그림 1) MAD 지표, 저화질 영역에서의 흐름도[16]

BLIIND 는 앞서 소개된 기술들과는 다른 무기준법 기반의 방식으로, 원본 영상과의 비교 없이 입력된 왜곡 영상의 주파수 분포가 얼마나 자연스러운 영상 같은지를 판단하여 화질을 예측하는 방식이다[17]. 본 기술을



(a) Sub-band 별 DCT 계수 분포 (b) Orientation 별 DCT 계수 분포

(그림 2) BLIIND에서 활용되는 특징정보[17]

비롯한 다양한 무기준법 화질 측정 기술들의 경우 Natural Scene Statistics(NSS)라는 가정을 기반으로 한다. NSS 가정이란, 자연스러운(Natural) 영상의 주파수 분포는 특정한 통계적 특성을 따를 것이며, 여기서 벗어나는 영상은 부자연스러운, 화질이 좋지 않은 영상일 것이라는 가정이다. BLIIND 기술의 경우, (그림 2)와 같이 영상을 블록 단위로 나누고, 각 블록별 DCT 계수를 전체적으로, 방향별(Orientation)로, 주파수 대역별(Sub-band)로 나누어 계수의 분포 특성을 파악한다. 이때, DCT 계수 분포는 주로 중앙부에 몰려있는 형태이므로, Generalized Gaussian Distribution(GGD)라는 중앙부의 첨도가 높은 분포 모델에 피팅(Fitting) 한 후, 그때의 GGD 수식 파라미터를 특징정보로써 활용한다[17]. 수집된 DCT 계수 분포 기반의 특징정보들은 Support Vector Machine(SVM) 혹은 기타 확률모델을 이용하여 MOS 점수에 매핑된다. 본 기술은 영상을 공간적 영역에서가 아닌 주파수 영역에서 정보를 추출한다는 점과

〈표 3〉 LIVE Image 데이터베이스에서의 성능수준

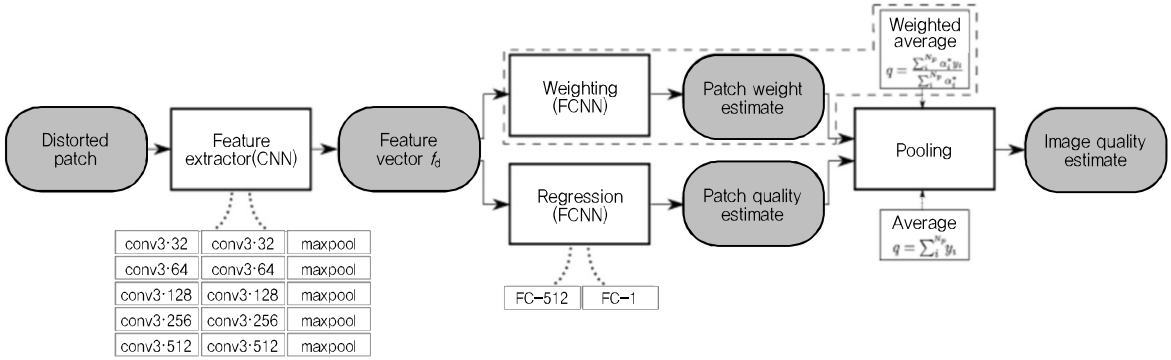
화질 지표	SROCC	PLCC
PSNR	0.866	0.856
SSIM	0.913	0.906
MS-SSIM[15]	0.953	0.945
MAD[16]	0.956	0.949
BLIIND[17]	0.930	0.931
CNN IQA v1[13]	0.972	0.960
CNN IQA v2[14]	0.980	0.970

원본 영상 없이도 주파수 계수의 분포를 활용하여 화질을 적정 수준 예측한다는 점에서 매우 의미가 있는 기술이다. 상기 소개된 모델 기반 기법들의 LIVE Image 데이터베이스에서의 성능 수준은 〈표 3〉에 제시되어 있다.

2. 데이터-드리븐(Data-Driven) 기법

데이터-드리븐 기법은 영상에 대한 특징정보 추출 없이 딥러닝 기술을 이용하여 영상과 MOS 점수 간의 상관관계를 학습하는 기법이다. 데이터-드리븐 기법 중 대표적인 기술은 Fraunhofer HHI에서 제안한 Convolutional Neural Network(CNN) 기반 기술들[13], [14] 이다. 먼저 무기준법 기반 기술 CNN IQA v1[13]이 제안되었으며, 그 이후 전기준법 기반 기술 CNN IQA v2[14] 연구를 통해 성능을 소폭 상승시켰다.

CNN IQA v1 기술의 흐름도는 (그림 3)과 같다. 본 기술은 도메인 지식을 활용환 알고리즘 기반 특징정보



(그림 3) CNN IQA v1 기술 흐름도[13]

추출 없이, 기계학습 분야에서 널리 활용되는 Visual Geometry Group(VGG) 신경망 구조[18]와 화질점수 데이터만을 활용하여 자동으로 특징정보를 추출하며, 정지영상 전체 정보를 활용하는 것이 아니라 무작위로 선정된 몇 개의 영상 패치들로 화질을 예측한다. 참고로 LIVE Image를 비롯한 다양한 정지영상 화질 데이터베이스의 경우, 정지영상 전체에 대한 화질점수는 제공하나, 정지영상을 이루는 영상 패치에 대한 점수는 제공하지 않는다. 하지만 CNN IQA v1의 경우 각 영상 패치의 특징정보를 활용하여 영상 패치의 화질점수를 예측하고, 이를 어떻게 가중치 합해야 정지영상 전체의 MOS 값으로 매핑 되는지를 학습하기 때문에, 학습이 진행될수록 화질 측면에서의 각 영상 패치의 중요 정도(Patch weight)를 판단할 수 있게 된다. 이는 정지영상과 그 화질점수만을 이용하여 역으로 영상의 주요 영역(Saliency map)을 유추할 수 있음을 시사하므로, 매우 의미 있는 접근이라고 할 수 있겠다. (그림 3)에 표현된 바와 같이 CNN IQA v1은 무기준법으로 영상의 화질을 평가하였으며, 영상 패치 점수를 조합하는 단계에서 상기의 Patch weight를 적용하는 방안과 단순 평균을 취하는 방안을 모두 적용해보았으며 그 결과, 단순 평균을 취하는 경우 화질 예측 정확도가 더 높게 나타났다. 하지만 유사 프레임워크를 전기준법 방식으로 적용한 CNN IQA v2[14]의 경우에는 Patch weight를 적용함으로써 예측 정확도가 더 올라감을 확인하였다. 영상 패치의 개수는 768×512 사이즈의 정지영상 기준, 32×32 패치 32개 이상일 때부터 예측 정확도가 수렴함을 확인하였다. CNN IQA v1과 CNN IQA v2 기술의 성능 수준은 〈표 3〉에 표기되어 있다. 상기 서술한 바와 같이 데이터-드리븐 기술의 경우 예측 성능이 높은 장점도 가지고 있지만, 정지영상과 화질점수 데이터만을 가지고 패치 단위 화질 평가 혹은 영상 내 중요 영역 판별 등 매우 활용성이 높은 정보들을 유추할 수 있는 장점을 가지므로 앞으로의 발전 방향이 더 기대되는 연구 기법이다.

Ⅳ. 동영상 인지화질 측정 기술 동향

동영상 화질 평가의 경우, 정지영상 화질 평가 기술을 단순히 여러 장으로 확장한 것으로 생각할 수 있겠지만, 동영상의 인지화질을 정확하게 예측하기 위해서는 공간적 특징정보뿐만 아니라 시간적 특징정보에 대한 고려가 매우 중요해진다. 이에 다양한 연구 기관에서는 기존 개발된 우수한 정지영상 화질 측정 기술상에 각자의 방법으로 시간적 요소를 추가하는 방향으로 연구를 진행하고 있다. 동영상 화질 측정 기술에서도 기계학습이 활발히 활용되고 있는데, 정지영상의 경우와는 달리 순수 데이터-드리븐 기법을 차용하는 시도는 적고 대부분이 모델 기반 기법을 차용한다. 동영상은 정지영상 대비 정보량이 매우 많아 깊은 신경망을 적용하기에는 요구되는 메모리양이 너무 큰 이슈가 있으며, 정지영상과 비교하였을 때 마스킹(Masking) 효과도 훨씬 크기 때문에 모델 기반 기법을 활용하여 주요영역 및 특징정보를 일정 부분 가이드 할 필요성이 있기 때문에 그런 것으로 판단된다.

동영상 화질 측정 기술들은 크게 두 가지로 분류될 수 있는데, 영상 전체를 시청 후 하나의 화질 점수로 표현을 하는 시청 후(Retrospective) 화질 측정 분야와 영상 중간중간의 화질을 지속적으로 측정하는 연속적(Continuous-time) 화질 측정 분야로 나뉜다. 연속적 화질 측정 기술들의 경우, 주로 스트리밍 환경에서 비디어 버퍼링(Stall)과 같은 이벤트로 인한 시청자의 경험을 측정하기 위한 연구가 대부분인데, 이처럼 평가 대상 영상에 버퍼링이 포함되는 경우, 영상의 화질 이슈라고 보기 힘든 측면도 있기 때문에 화질 대신 체감 품질(QoE)이라는 용어를 쓰는 경우도 있다. 본 절에서는 각 분야의 대표 기술들을 알아보고, 각 기법의 특징에 대해서 살펴보도록 하겠다.

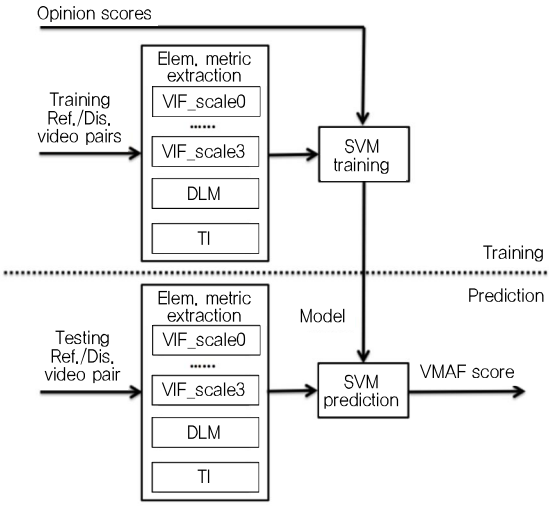
1. 시청 후(Retrospective) 화질 측정

시청 후 화질을 측정하는 기술들은 10초가량 영상 전체의 시/공간적 특징정보를 분석한 이후, 이를 토대로 영상 전체에 대한 하나의 MOS값을 예측한다. 대표적인 기술들로는 Spatiotemporal MAD(ST-MAD), Video BLIINDS(V-BLIINDS), VQM-VFD, VMAF(Video Multimethod Assessment Fusion) 등이 있다.

ST-MAD 기술의 경우, 정지영상 화질 측정 기술인 MAD[16]를 기반으로 한 기술로, 영상을 하나의 입체 볼륨(volume)으로 간주하고 이를 다양한 방향에서 슬라이스(slice)하여 그 단면에 나타난 왜곡 정도를 측정한다[19]. 이때 단면이 우리가 통상 알고 있는 영상 프레임일 경우에는 MAD와 동일하게 왜곡 정도를 측정하고 이를 Spatial MAD(S-MAD) 라고 칭한다. 반면, 단면이 각 프레임의 행 혹은 열을 모아놓은 평면인 경우에는 이를 각 행 혹은 열이 시간적인 전개에 따라서 어떻게 변해가는가를 보여주는 시간적 정보로 간주, 해당 평면에 대한 d_{appear}값만을 계산하여 왜곡 정도를 측정한다. 그 이후 각 평면의 d_{appear}값에 대한 가중치 합을 수행하는데, 사람이 주로 프레임 중앙부에서 일어나는 모션에 집중함을 고려하여 프레임 중심부에 속하는 행 혹은 열에 높은 가중치를 적용한다. 이렇게 계산된 최종값은 Temporal MAD(T-MAD)라고 칭하며, S-MAD와 T-MAD를 조합함으로써 동영상의 화질을 나타내는 ST-MAD를 계산할 수 있다. ST-MAD 기술의 경우 시간적 방향의 단면 분석을 통해 시간적 왜곡 정도를 고려하였고, 이를 통해 동영상의 인지화질을 보다 정확히 측정할 수 있음을 증명했다는 점에서 의미가 있는 기술이다.

V-BLIINDS의 경우 정지영상 화질 측정 기술 BLIIND[17]를 기반으로 한 기술로, 시간적 특성을 고려하기 위해 인접한 영상 프레임 간의 차분값(Frame-difference)을 이용하는 비교적 간단한 방법으로 기술의 활용 범위를 동영상으로까지 확대하였다[20]. BLIIND 기술과 유사하게, Frame-difference 영상에 대한 전체적, 방향별, 그리고 주파수 대역별 DCT 계수 분포를 GGD 수식에 피팅하고, 그때의 파라미터값을 동영상 MOS값에 매핑하는 과정을 거친다. BLIIND와 마찬가지로 무기준법 기술이기 때문에 원본 영상 없이도 동영상의 화질을 예측할 수 있다는 점에서 활용성이 높은 기술이라고 할 수 있다.

VQM-VFD와 VMAF는 전기준법 기반의 동영상 화질 측정 기술이며, 추출한 특징정보에 기계학습을 이용하여 MOS값을 예측한다는 점, 시간적 특성 고려를 위해 Temporal Information(TI) 특징정보를 다룬다는 점 등에서 서로 유사성이 있는 기술들이다. 여기서 TI란 V-BLIIND에서 사용했었던 Frame-difference 영상을 원본 영상과 왜곡 영상 각각에 대해서 구한 후, 그 둘 간의 차이 정도를 구한 것을 의미한다. VQM-VFD는 미국 NTIA에서 연구된 동영상 화질 지표로, 영상 내 전반적인 엣지 그리고 수직 수평 성분이 강한 엣지 등을 추출하여, 영상 내 블러(Blur) 혹은 블록 아티팩트(Block Artifact) 등으로 인한 화질 효과를 보고자 하며, 기술의 마지막 단에는 Fully Connected(FC) 신경망이 붙어 상기 공간적 특징정보 및 TI를 하나의 MOS 점수로 매핑한다. 참고로 VQM-VFD에 기본 탑재되는 FC 신경망 weight의 경우, 83명의 실험 참가자가 다섯 해상도의 11,255종 비디오에 평가한 대규모 데이터베이스를 기반으로 학습되었다[21]. VMAF의 경우 Netflix사에서 USC, UT Austin 등의 학교와의 연구를 통해 개발 중인 동영상 화질 측정 기술로, 영상의 특성에 따라 기 존재하는 여러 화질 지표를 적응적으로 융합하는 Multi-Method Fusion(MMF) 방식을 사용한다. (그림 4)에서 보이는 바와 같이 VMAF는 Visual Information Fidelity (VIF), Detail Loss Measure(DLM), TI 등의 기 존재하는 지표를 가중치 합하여 계산되며, 가중치 합 방식은 SVR을 이용하여 학습한다[1]. VMAF는 특히 JVET 등



(그림 4) VMAF 기술 흐름도[1]

〈표 4〉 LIVE Video 데이터베이스에서의 성능수

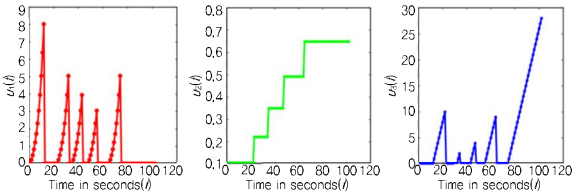
화질 지표	SROCC	PLCC
PSNR	0.416	0.453
SSIM	0.688	0.641
VMAF[1]	0.667	0.664
MS-SSIM[15]	0.696	0.639
ST-MAD[19]	0.825	0.833
V-BLIIND[20]	0.722	0.824
VQM-VFD[21]	0.761	0.763

에서 활발히 기고되고 있으며, 시간적 특성 등을 더욱 고려하도록 기술 개선이 지속적으로 이루어지고 있어 앞으로 주목이 되는 기술이다. 상기 소개된 동영상 화질 측정 기술들의 LIVE Video 데이터베이스에서의 성능 수준은 〈표 4〉에 제시되어 있다.

2. 연속적(Continuous-Time) 화질/QoE 측정

연속적 화질/QoE 측정 기술은 네트워크 스트리밍 환경에서 동영상 압축 혹은 버퍼링으로 인한 체감 품질을 측정하기 위한 연구 분야로, 현재 및 과거 상태(State)를 기반으로 시청자가 느끼는 품질을 가늠하는 기술로, 최적의 체감 품질 제공을 위한 영상의 해상도 및 압축률 제어 등에 활용될 수 있는 기술이다.

연속적 화질/QoE 연구에 활용되는 데이터베이스로



(a) 버퍼링 지속시간 (b) 현재까지 버퍼링 발생 회수 (c) 정상 재생 지속 시간

(그림 5) 연속적 QoE 측정 기술에 활용되는 특징정보[25]

는 LIVE Mobile Stall Video Database-II[22], Waterloo QoE Database[23], LIVE-Netflix Video QoE Database[24] 등이 있으며, 상기 데이터베이스들의 경우 동영상에 대한 하나의 화질값만을 제공하는 통상의 데이터베이스들과는 달리 영상의 각 프레임에 대한 화질 정보를 제공한다. 동영상의 각 프레임에 대한 주관적 화질 점수를 수집한 방법을 살펴보면, 실험참가자에게 동영상을 시청하게 한 후, 그 아래 놓인 스코어 바(Score bar)를 실시간으로 조정할 수 있도록 한다[22].

연속적 화질/QoE 측정 기술들은 다양한 특징정보들을 활용해 연속적으로 각 프레임의 화질/QoE을 예측하며 실제 주관적 데이터와의 통계적 유사성을 통해 그 성능을 평가한다. 이때 활용되는 성능 지표로는 시청 후 화질 측정 기술과 마찬가지로 SROCC와 PLCC 등이다.

연속적 화질/QoE 측정에 활용되는 특징정보는 주로 버퍼링과 연관된 것이 많은데, 여러 기술에서 공통으로 활용되는 특징정보로는 (그림 5)와 같이 버퍼링 지속 시간, 현재까지 버퍼링 발생 횟수, 정상 재생 지속시간 등이 있다. 이외에도 각 프레임에 대한 정지영상 화질 측정 결과 등이 특징정보로 활용되기도 한다. (그림 5)에서 나타나는 바와 같이, 연속적 화질/QoE 측정에 활용되는 특징정보는 각 프레임에 대한 수치를 가지고 있는 연속적인 정보이며, 특정 시점에서의 화질/QoE 예측을 수행하고자 할 때 주어진 정보는 현재 그 이전 시점의 특징정보 값들이다. 특히 사람의 인지적 기억(Memory) 효과를 고려하면, 현재 상태와 과거 상태 정보들을 어떻게 잘 조합하여 현재 시점 체감 품질을 예측할지에 대한

고민이 매우 중요하다고 할 수 있다[25].

상기의 일환으로, [25]에서는 Hammerstein Wiener (HW)[26] 모델이라는 과거 및 현재 상태의 가중치를 합하는 모델을 활용하여 연속적 화질/QoE를 측정하였으며, 관련 연구 [2]에서는 상기의 HW 모델과 더불어 Recurrent Neural Network(RNN)[27]와 Non-linear Autoregressive Neural Network(NARX)[28] 신경망을 활용하고 각 방법의 성능을 비교하였다. 그 결과 HW와 NARX를 활용하였을 때의 연속적 화질/QoE 예측 정확도가 우수하였으며, RNN을 이용하였을 때가 성능이 조금 떨어졌음을 보였다[2].

연속적 화질/QoE 측정 관련 논문들([2], [22], [24], [25])을 살펴보면 특히 최근 들어 활발히 연구됨을 알 수 있으며, 본 기술은 OTT와 같은 미디어 서비스의 질을 높일 수 있는, 산업적 가치가 매우 큰 기술이기 때문에 앞으로의 발전 방향이 더욱 기대된다.

V. 결론

지금까지 본고에서는 영상 인지화질 측정기술의 개념, 그리고 정지영상 및 동영상 화질 측정 기술의 동향에 대해서 살펴보았다. 최근의 정지영상 및 동영상 화질 측정 기술 동향을 보면, 기존의 신호 왜곡 정도를 계산하는 수준을 넘어서 사람이 실제로 인지한 영상의 화질을 정확히 반영하고자 하며, 이 과정에서 주관적 화질 데이터와 기계학습 기술을 활용하기도 한다. 또한, 영상 내 노이즈, 압축 아티팩트, 네트워크 버퍼링 등의 다양한 요소로 체감 품질을 측정하고 예측하는 등, 미디어 서비스 다양한 단계에서의 인지화질 영향도를 분석하고 개선하고자 하는 노력이 이루어지고 있다. 앞으로는 영상의 공간해상도 및 프레임률 증가뿐만 아니라, HDR/WCG 혹은 AR/VR 영상과 같은 새로운 축에서의 실감성이 향상된 영상들이 소개되고 서비스될 것이며, 이에 대응하여 화질 측정 기술 분야는 인지화질뿐만 아니라, 몰입형 환경에서의 시각적 피로도 등 다양한 요소들을 종합적으로 고려하는 방향으로 발전할 것으로 전망된다.

약어 정리

AVC	Advanced Video Coding
CNN	Convolutional Neural Network
CSF	Contrast Sensitivity Function
DMOS	Differential Mean Opinion Score
HEVC	High Efficiency Video Coding
HVS	Human Visual System
ITS	Institute for Telecommunication Science
JPEG	Joint Photographic Experts Group
MAD	Most Apparent Distortion
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MSE	Mean Squared Error
MS-SSIM	Multi Scale SSIM
NARX	Non-linear Autoregressive Neural Network
PLCC	Pearson Linear Correlation Coefficient
PSNR	Peak Signal to Noise Ratio
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SROCC	Spearman's Rank Order Correlation Coefficient
SSIM	Structural Similarity
SVR	Support Vector Regression
VCEG	Video Coding Experts Group
VGG	Visual Geometry Group
VMAF	Video Multimethod Assessment Fusion
VQEG	Video Quality Experts Group

참고문헌

[1] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, “Toward a Practical Perceptual Video Quality Metric,” The Netflix Tech Blog, July 24, 2017. http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html

[2] C.G. Bampis, Z. Li, I. Katsavounidis, and A.C. Bovik “Recurrent and Dynamic Models for Predicting Streaming

Video Quality of Experience," *IEEE Trans. Image Process.*, vol. 27, no. 7, July 2018, pp. 3316-3331.

[3] L. Xu, W. Lin, and C.C. Kuo, *Visual Quality Assessment by Machine Learning*, Singapore: Springer, 2015.

[4] LIVE Image Quality Assessment Database. http://live.ece.utexas.edu/research/quality/subjective.htm

[5] N.N. Ponomarenko, "Tampere Image Database, Version 1.0, 2008," Feb. 22, 2010. http://www.ponomarenko.info/tid2008.htm

[6] N.N. Ponomarenko, "Image Database for Evaluation of Full-Reference Image Visual Quality Assessment Metrics," Mar. 23, 2014. http://www.ponomarenko.info/tid2013.htm

[7] CSIQ Lab, "CSIQ Image Database," Apr. 29, 2016. http://vision.eng.shizuoka.ac.jp/mod/page/view.php?id=23

[8] Laboratory for Image & Video Engineering, "LIVE Video Quality Assessment Database," 2009. http://live.ece.utexas.edu/research/quality/live_video.html

[9] CSIQ Lab, "CSIQ Video Database," May 8, 2016. http://vision.eng.shizuoka.ac.jp/mod/page/view.php?id=24

[10] Z. Li, "NFLX Public Database," GooGle Drive, Mar. 2, 2016. https://drive.google.com/drive/u/0/folders/0B3YWNlCYMBIweGdJbERlUG9zc0k

[11] Institute for Telecommunication Science, "VQEG, HDTV Database." http://www.its.bldrdoc.gov/vqeg/projects/hdtv/

[12] ITU-R BT.500-13 (2012) ITU, *Methodology for the Subjective Assessment of the Quality of Television Pictures*, 2012.

[13] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A Deep Neural Network for Image Quality Assessment," *IEEE Trans. Image Proccess.*, Phoenix, AZ, USA, Sept. 28, 2016, pp. 3773-3777.

[14] S. Bosse, D.Maniry, K.-R. Muller, T. Wiegand, and W. Samek, "Deep Neural Networks for No-reference and Full-Reference Image Quality Assesment," *IEEE Trans. Image Proccess.*, vo. 27, no. 1, Jan. 2017, pp. 206-219.

[15] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale Structural Similarity for Image Quality Assessment," *IEEE Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 9-12, 2003, pp. 1398-1402.

[16] E.C. Larson and D.M. Chandler, "Most Apparent Distortion: Full-Reference Image Quality Assessment and the Role of Strategy," *J. Electron. Imag.*, vol. 19, no. 1, 2010, pp. 0110061:1-0110061:21.

[17] M.A Saad, A.C. Bovik, and C. Charrier, "Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, Aug. 2012, pp. 3339-3352.

[18] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv:1409.1556, Sept. 2014.

[19] P.V. Vu, C.T. Vu, and D.M. Chandler, "A Spatiotemporal Most-Apparent-Distortion Model for Video Quality Assessment," *IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sept. 11-14, 2011, pp. 2505-2508.

[20] M.A. Saad, A.C. Bovik, and C. Charrier, "Blind Prediction of Natural Video Quality," *IEEE Trans. Image. Process.*, vol. 23, no. 3, Mar. 2014, pp. 1352-1365.

[21] S. Wolf and M.H. Pinson, "Video Quality Model for Variable Frame Delay (VQM_VFD)," U.S. Dept. Commer., Nat. Telecommun. Inf. Admin., Boulder, CO, USA, Tech. Memo TM-11-482, Sept. 2011.

[22] D. Ghadiyaram, J. Pan, and A. Bovik, "A Subjective and Objective Study of Stalling Events in Mobile Streaming Videos," *IEEE Trans. Circuits Syst. Video Technol.*, Nov. 2017.

[23] Z. Duanmu, A. Rehman, K. Zeng, and Z. Wang, "Quality-of-Experience Prediction for Streaming Video," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Seattle, WA, USA, July 11-16, 2016, pp. 1-6.

[24] C.G. Bampis, Z. Li, A.K. Moorthy, I. Katsavounidis, A. Aaron, and A.C. Bovik, "Study of Temporal Effects on Subjective Video Quality of Experience," *IEEE Trans. Image Process.*, vol. 26, no. 11, Nov. 2017, pp. 5217-5231.

[25] D. Ghadiyaram, J. Pan, and A.C. Bovik, "Learning a Continuous-Time Streaming Video QoE Model," *IEEE Trans. Image Process.*, May. 2018, pp. 2257-2271.

[26] J.A. Nelder, "The Fitting of a Generalization of the Logistic Curve," *Biometrics*, vol. 17, no. 1, 1961, pp. 89-110.

[27] J.L. Elman, "Finding Structure in Time," *Cognitive Sci.*, vol. 14, no. 2, 1990, pp. 179-211.

[28] C.G. Bampis and A.C. Bovik, "An Augmented Autoregressive Approach to HTTP Video Stream Quality Prediction," arXiv: 1707.02709, July 2017.