일반논문-09-14-1-04

구조적 유사도 기반의 인간의 시각적 특성을 이용한 비디오 품질 측정 기준

박 진 철[a], 이 상 훈[a]‡

Structural Similarity Based Video Quality Metric using Human Visual System

Jincheol Park[a], and Sanghoon Lee[a]‡

요 약

최근 에러의 가시도를 측정하던 기존 패러다임의 한계를 극복하고자 structural similarity (SSIM) metric이 제안되어 우수한 성능을 보이고 있다. 하지만 SSIM은 기존에 활발히 연구되어오던 인간시각체계의 민감도에 대한 특성을 완전히 배제함으로써 새로운 한계점을 노출한다. 본 논문에서는 포비에이션 포인트로부터의 거리, 평균 휘도 값, DCT 계수, 모션 정보를 이용하여 통합된 시각적 가중치를 정의하였고 이를 SSIM과 자연스럽게 결합함으로써 성능을 개선하였다. VQEG 멀티미디어 그룹의 테스트 플랜을 이용한 테스트를 통해 본 논문의 품질측정 기준이 기존의 SSIM 보다 주관적 화질평가의 결과와 연관도가 더 높음을 보임으로써 성능 향상을 증명하였다.

Abstract

Recently, the structural similarity (SSIM) index metric is proposed. In the present paper, a new framework, which is called visual SSIM (VSSIM), is proposed by incorporating crucial human factors into the SSIM. The human factors are foveation, luminance, frequency and motion information. The performance of VSSIM is evaluated by subjective quality test compliant with the Video Quality Expert Group (VQEG) multimedia group test plan. It shows that the visual SSIM is more correlated with the subjective quality result than the conventional SSIM.

Keyword : 구조적 유사도, 인간 시각 체계, spatial weight, IAF (information allocation function) weight, MV weight

I. 서 론

최근 들어 무선 통신, 영상 압축 기술, 컴퓨터 네트워크 등의 급격한 발전은 무선 환경에서의 영상통신을 가능하게 하였다. 또한 무선 네트워크의 발전은 단순한 무선 환경뿐만 아니라 이동성이 요구되는 경우에서도 영상통신이 가능하게 하였으며 이에 따라 기존에 음성이나 데이터가 통신 트레픽의 대부분을 차지하였던 상황에서 비디오 소스도 통신 트레픽의 중요한 부분을 획득하고 있다. 하지만 비디오 소스는 다른 종류의 트레픽들과 구분되는 특성들이 있다. 이는 비디오 화질 측정을 위한 특별한 기준의 필요성을 요구한다. 첫째 비디오 소스에서 발생하는 트레픽은 다른 어떤 data의 전송 량보다 월등히 많다. 따라서 고용량의 비디오 소스를 압축하고 전송하는 과정에서 필연적인 데이터 손실과 왜곡이 발생하기 때문에 비디오 코딩과 중간 네트웍들의 성능 검증과 모니터링의 필요성을 유발한다. 둘째

a) 연세대학교 전기전자공학부
Department of Electrical Engineering, Yonsei University
‡ 교신저자 : 이상훈(slee@yonsei.ac.kr)
※ 이 논문은 2008년도 정부(지식경제부)의 재원으로 IITA의 지원을 받아 수행된 연구임(IITA-2008-(C1090-0801-0011)). 또한, 이 논문은 2008년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. R01-2007-000-11708-0)

영상 통신에서의 최종 수신단은 사람의 눈이다. 따라서 이는 사람의 눈이 인지하는 데이터이기 때문에 단순히 손실 없이 전송된 비트 수에 의해서만이 아니라 시각적 특성에 의해서도 실제 최종 수신단에 전송되는 정보량은 다르다. 따라서 통신에서 주로 사용하는 Mean Square Error (MSE) 처럼 단순히 전송된 신호의 양만을 측정하는 것이 아닌 최종 수신단 입장에서 정보의 질을 측정 하는 별도의 기준이 필요하다.

인간시각체계의 특성에 가장 근접한 품질 측정은 주관적 화질 평가이다. 이것은 직접 관찰자들이 비디오 품질을 평가하는 방식이다. 이에 대한 방식은 이미 VQEG의 테스트 플랜에서 권고되어 표준화 되어있다[10]-[13]. 하지만 이 방식은 매우 많은 시간과 비용을 필요로 하기 때문에 실제 영상 통신에서의 응용에 한계가 있다. 따라서 주관적 품질 평가의 결과에 근접하면서 자동적으로 품질을 측정하는 많은 객관적 품질 들이 연구되어 왔다. 가장 고전적인 비디오 품질 기준으로는 MSE로부터 유도되는 PSNR이 있지만 인지되는 비디오 품질과 매우 상관도가 떨어지며 영상 품질 기준으로써 부적절하다는 것은 이미 공공연한 사실이다.

현재 대표적인 객관적 비디오 품질 측정 기준으로는 크게 perception based metric, error sensitivity metric, SSIM index approach로 나눌 수 있다. 이 중에서 가장 활발한 연구가 진행된 기준은 perception based metric이다[14]. 이는 품질 손실에 영향을 미치는 여러 가지 인지 특성들을 추출하여 그 정도를 측정하고 선형적으로 결합하는 방식이다. 예를 들면 blurring, jerky/unnatural motion, global noise, blocking artifact, color distortion 등과 같은 인지 특성들을 추출한다. 이러한 특성으로 정의되는 파라미터들의 크기를 측정하고 가중 합으로 결합하는 모델을 개발하여 영상 화질을 측정하는 방식이다. 이 방식은 VQEG에서 ITU-T에 권고하여 2004년에 ITU-T J.144 표준으로 채택되었으며 Video Quality Metric (VQM)이라고 부른다. Error sensitivity metric은 가장 오래 전부터 연구되어오던 방식으로 품질 손실은 에러 시그널의 크기와 직접적으로 연관되어 있다는 가정하에 MSE의 개념으로 참조영상과 왜곡영상간 에러의 크기를 구한다[1][9]. 그 다음 인간시각체계 모델로 에러 시그널의 민감도를 구한 후 이를 에러의 크기에 가중치를 부여 함으로써 품질을 측정한다. 이 때 이용되는 인간시각체계모델은contrast sensitivity function (CSF)와 masking effect 가 있으며 이것들을 이용하여 에러의 민감도를 구한다. 이는 원래 Daly, Lubin, Watson 등이 이미지 품질 측정에 초첨을 맞춰서 연구해 오다가 HVS의 시간적 개념들을 추가하여 Tan, Van den Branden Lambrecht, Winkler 등이 비디오 품질 측정으로 발전 시켰다[1]. 하지만 이러한 에러 가시도를 측정하는 품질 기준들은 그 동안의 무수한 노력에도 불구하고 널리 통용될 만큼 뛰어난 성능을 보이지 못할 뿐 아니라 많은 한계점들을 노출한다.

한편 최근에 기존의 접근들과는 다르게 HVS은 영상을 인지할 때 신호의 구조적 정보를 추출한다는 새로운 가설이 Wang에 의해 제기 되었다[2]-[4]. 따라서 품질 손실은 어떤 에러의 타입에 의해서라기 보다는 신호 자체의 구조적 왜곡에 의해 발생한다는 가정하에 구조적 왜곡을 측정하는 구조적 유사도 측정 (SSIM: Structural SIMilarity) 방식이 제안 되었다. 이는 더 정확한 품질 왜곡을 예측한다. 하지만 SSIM은 여전히 추가적인 성능 개선의 여지를 담고 있다. SSIM은 인간시각체계가 영상에서 구조적 정보를 인지한다는 HVS의 특성은 고려하지만 그 신호의 특성에 따라서 인간시각체계에 중요한 정도가 다르다는 특성은 고려하지 않았다. 하지만 영상 안에서 각 신호의 특성에 따라 인간시각체계가 갖는 중요도가 다르다는 것은 이미 많은 논문의 연구 결과를 통해서 증명된 사실이다. 이를 바탕으로 본 논문에서는 Wang의 가설에 HVS의 시각적 중요도에 대한 특성을 추가하여 개선된 품질측정기준을 제안한다. 신호 자체가 갖는 성분이 시각적으로 얼마나 중요냐에 따라서 신호가 겪는 구조적 왜곡의 정도가 동일하더라도 실제 인간시각체계에 미치는 영향은 다르다는 가정하에 신호 자체가 갖는 시각적 중요도에 의해 HVS에 미치는 구조적 왜곡의 영향을 조절한다. 본 논문에서는 SSIM의 강점과 HVS의 민감도특성을 자연스럽게 조합하는 프레임웍을 제시함으로써 성능을 향상시킨 것에 기여도가 있다.

본 논문은 그 동안 많은 연구가 진행되었던 비디오 품질 측정들의 조사를 바탕으로 그들의 한계점을 파악하여 개선된 품질 기준을 제안하였으며 영상통신에서의 비디오 소스를 평가하는데 실험의 초점을 맞췄다. 본 시스템을 평가하

기 위한 주관적 화질평가에는 VQEG multimedia group의 test dataset중에서 다수가 사용되었으며 VQEG multimedia group test plan에 부합되는 방식으로 진행되었다. 실험결과에서는 본 논문에서의 품질 기준이 SSIM보다 주관적 화질평가 결과와 더 높은 상관도를 보임으로써 기존의 SSIM에 인간시각체계의 특성을 적용한 기준이 인간시각체계에 더욱 근접함을 증명하였다.

II. Structural Similarity (SSIM)

SSIM은 이전에 많이 연구되던 품질 기준들과는 달리 하향식 접근을 통해 화질을 측정을 하는 방식이다. 하향식 접근이란 인간 시각 체계가 화질을 평가할 때 영상 신호의 요소들 하나하나를 평가하는 것이 아니라 전체적인 관점에서부터 시작해서 영상 품질을 평가한다 가정하에 이루어지는 것이다. 따라서 이전의 방식들이 영상신호의 구성 요소들의 에러의 가시도를 측정함으로써 품질의 손실을 측정하는 bottom up 접근 이었다면, SSIM은 사람의 눈이 영상을 볼 때 에러를 추출하는 것에 집중하는 것이 아니라 영상 안에서의 구조적 정보를 추출하는데 집중한다는 새로운 가설을 통해 접근한다. 따라서 N×N 크기의 window 안에서 original signal을 $x=\{x_i \mid i=1,2,...,N\times N\}$ 로 distorted signal을 $y=\{y_i \mid i=1,2,...,N\times N\}$ 로 표현 했을 때 다음처럼 평균 l(x,y), 콘트라스트 c(x,y), 그리고 상관도 s(x,y)의 유사도를 구하고 이들을 서로 곱함으로써 구조적 유사도를 구한다.

$$l(x,y)=\frac{2\mu_x\mu_y}{\mu_x^2+\mu_y^2},\ c(x,y)=\frac{2\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2},\ s(x,y)=\frac{2\sigma_{xy}}{\sigma_x\sigma_y} \quad (1)$$

$$S(x,y)=l(x,y)\cdot c(x,y)\cdot s(x,y)=\frac{4\mu_x\mu_y\sigma_{xy}}{\left(\mu_x^2+\mu_y^2\right)\left(\sigma_x^2+\sigma_y^2\right)} \quad (2)$$

여기에서 μ_x와 μ_y는 각각 신호 x와 y의 평균값이며 이는 신호의 밝기를 반영한다. σ_x와 σ_y는 각각 신호 x와 y의 분산이며 이는 신호의 콘트라스트의 정도를 반영한다. σ_{xy}는 신호 x와 y 사이의 공분산이며 이는 두 신호간의 상관도를 반영한다.

SSIM은 특히 error sensitivity approach 방식에서의 한계점들을 훌륭하게 보완한다. 하지만 SSIM은 정지영상에 기반을 두기 때문에 비디오 소스에서의 성능은 확실하지 않은 상태이다. 하나의 비디오 프레임 안에서 각 신호마다 갖는 특성에 따라 HVS의 특성에 의한 중요도가 다르다는 것은 이전의 논문들을 통해 충분히 증명되어있다[9]. 그럼에도 불구하고 여기에서는 기존의 인간시각체계 모델들이 완전하지 않다는 이유로 대부분의 신호들을 동일하게 취급한다. 따라서 error sensitivity based metric의 경우와는 다르게 비디오 소스 안에서 각 프레임간의 구조적 유사도는 훌륭하게 측정 하지만 모든 신호를 동일하게 취급하기 때문에 인간시각체계의 특성을 잡아내지는 못하는 경우를 볼 수 있다. 하지만 SSIM은 영상신호의 유사도를 아주 정확하게 측정 하기 때문에 신뢰성 있는 시각체계 모델에 의한 시각적 중요도의 적용을 통해 추가적인 성능 향상을 기대할 수 있다.

III. Visual Structural Similarity (VSSIM)

본 논문에서는 신호 자체가 시각적으로 얼마나 중요한 정보를 담고 있는가에 따라 동일한 정도의 구조적 왜곡에서도 시각적으로 미치는 영향의 정도는 다르다는 가정하에 신호 자체의 시각적 중요도를 정의한다. 따라서 각 신호 자체가 지니고 있는 특성들이 HVS에 얼마나 큰 영향력을 가지고 있는지를 중요도로 정의하고 적용함으로써 각 신호마다의 구조적 왜곡의 전체 스코어에 대한 영향력을 중요도에 따라 다르게 반영시킨다. 물론 본 논문에서의 HVS 모델도 이상적일 수는 없지만 이전에 다른 인지적 영상 품질 측정 방식들에서의 모델들보다 신뢰성 있는 성능을 보이며 이는 기존 SSIM의 분명한 성능향상을 유도함을 본 과제의 실험 결과를 통해서 알 수 있다. 이는 실제 주관적 화질평가가 주어진 이미지 자체의 수학적 관계를 통해서 직접적으로 유도되기 보다는 HVS를 통해서 어떻게 인지되는가에

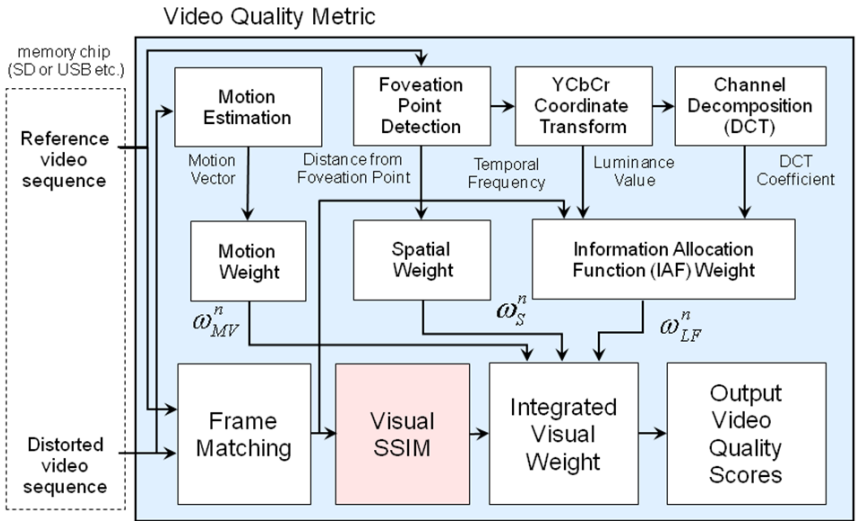

그림 1. 제안하는 영상 품질 기준의 블록 다이어그램
Fig. 1. A block diagram of the proposed video quality metric

의해 정해지기 때문이다. 따라서 본 논문은 이러한 시각적 가중치를 사용하는 영상품질 기준들의 기본 원리를 SSIM의 우수한 두 영상간의 구조적 유사성 측정 능력에 적용함으로써 하향식 접근과 상향식 접근방식들의 장점을 동시에 취한것이라 할 수 있다. 따라서 추후에 훌륭한 시각적 모델들이 나온다면 추가적인 성능 향상도 기대할 수 있다.

그림 1는 제안하는 영상품질 기준의 블록 다이어그램을 나타낸다. 여기에서의 모든 과정은 8x8 의 블록 단위로 진행된다. 이는 HVS의 특성과 SSIM을 자연스럽게 결합하는 것을 가능하게 한다. 구조적인 정보를 추출하기 위해서는 8x8의 샘플 윈도우가 필요한데 이것이 매 블록이 된다. 시각적 중요도 또한 매 블록 단위로 정의될 수 있다. 따라서 매 블록마다 구해진 SSIM의 결과는 매 블록마다의 시각적 중요도에 의해서 전체 score에 반영되는 정도가 달라지게 된다. 여기에서의 SSIM은 두 영상간의 구조적 유사도를 비교하는 기본 측정 장치처럼 사용되어 기존의 픽셀단위로 단순히 두 영상간의 거리를 측정했던 방식을 대체한다. SSIM이 error sensitivity metric으로부터 직접적으로 개선된 점은 HVS model의 한계점을 극복한 것이 아니라 두 영상간의 비교방식이 우수하다는 것이기 때문이다.

그림 1에서 볼 수 있듯이 제안하는 품질 기준에서는 3가지의 시각적 중요도가 적용된다. 첫째 spatial weight를 구하기 위해서 인간 시각 체계 모델 중 포비에이션에 대한 특성을 이용한다. 포비에이션 이란 사람이 어떠한 지점을 응시할 때 그 지점에서 멀리 있는 신호일 수록 민감도가 떨어진다는 정신물리학에서의 시각적 개념이다[5][6]. 수학식 (3)은 이에대한 모델링 함수이다.

$$f_{totalBW}(x)=\min\left(\frac{e_2\ln\left(\frac{1}{CT_0}\right)}{\alpha\left[e_2+\tan^{-1}\left(\frac{d(x)}{Nv}\right)\right]},\frac{\pi Nv}{360}\right) \tag{3}$$

v는 영상의 폭으로 계산한 영상으로부터 사람의 눈까지의 거리, $d(x)$는 foveation point로부터 신호 x까지의 픽셀단위의 거리, N는 픽셀단위의 영상의 폭이다. $v=d(x)/N$이며 결과적으로 $\tan^{-1}(\frac{d(x)}{N_v})$는 그림에서의 e가 된다. 나머지는 상수이며 이들은 심리생리학에서 실험적으로 구한 조정 파라미터들이다. 값은 $e_2=2\bullet 3$, $\alpha=0.106$, $CT_0=1/64$이다. 이를 이용하여 foveation weight는 다음처럼 구한다.

$$w(x_n)=\frac{f_m(x_n)}{f_m(x^r)} \tag{4}$$

여기에서 x_n은 n 번째 블록의 중앙에 위치한 픽셀을 나타내며, x^r는 foveation point 가 된다. 사람이 응시하는 위치를 의미하는 foveation point를 찾는 방법은 움직임 정보 혹은 얼굴 검색 등 여러 가지 방법들이 연구되어 있다[5][6]. 구체적인 방법에 대해서는 이 논문의 범위를 넘어서기 때문에 언급하지 않겠다.

Information allocation function(IAF)은 원래 연속신호의 콘트라스트를 유한한 셋으로 매핑 할 때 각 주파수 도메인과 콘트라스트 도메인에서의 신호 변화에 대한 민감도를 고려하여 양자화 계수을 정함으로써 비트 할당을 하기 위한 정보를 제공하는 것이다[7][8]. IAF는 DCT 영역에서의 주파수와 휘도성분에 대한 각 DCT 계수의 크기 값을 이용하여 각 DCT 계수들의 중요도를 제공하기 때문에 이들의 중요도를 조합하면 한 블록이 갖는 신호 자체의 중요도를 구할 수 있다. IAF의 기본 원리는 다음과 같다. HVS의 구분 능력에 대한 한계는 주파수와 콘트라스트의 incremental threshold에 의해 정해진다. 만일 DCT 계수의 변화가 대응하는 incremental threshold 보다 작다면 HVS은 그 변화를 인지하지 못한다. 여기에서 각 incremental thresholds는 DCT 계수의 주파수가 무엇이냐에 따라서 그리고 평균 휘도 값에 대한 각 계수들의 비율에 따라서 정해지며 결국 IAF는 이 두 가지 incremental thresholds의 역수로 정해진다. IAF의 수학적 모델은 다음과 같다.

$$IAF(f,a_{f_a}/a_0)=\frac{K}{CSF(f)^{-1}+\frac{a_{f_a}}{a_0}G_f(f,a_{f_a}/a_0)} \quad (5)$$

여기에서 CSF(f)는 CSF를 나타내며, K는 상수, a_{f_a}는 주파수 f_a 에서의 DCT 계수, a_0는 평균 휘도 값, $G(\cdot)$는 평균 휘도값에대한 CSF의 비선형적 조정 함수 이다[7]. 식 (5)의 모델을 보면 IAF는 물리적으로 신호가 저주파 대역에 있으면서 평균 휘도 신호가 클 때 큰 값을 갖는다. 이는 고전적인 CSF를 이용하면서 평균 휘도 성분에 따라 그 영향을 조정해 주는 것이라 할 수 있다. IAF 모델을 이용하면 각 DCT 계수들의 시각적 중요도를 구할 수 있으며 한 블록의 시각적 중요도는 다음처럼 그 블록의 DCT 계수들의 합을 정규화하여 구한다.

$$w_{LF}(x_n)=\begin{cases}\frac{1}{MaxIAF}\sum_{a_f\in Z}IAF(f,a_f/a_0), & a_0>0\\ 0, & a_0=0\end{cases} \quad (6)$$

여기에서 Z는 0이 아닌 DCT 계수들의 집합이다. $MaxIAF$는 정규화 하기 위한 값으로써 하나의 프레임 안에서 가장 큰 값을 갖는 블록의 IAF 합 값이 된다.

모션정보는 비디오에서 가장 중요한 정보 중에 하나라는 것은 너무나도 자명한 사실임에도 불구하고 이를 직접적으로 이용하여 민감도를 구한 HVS 모델이나 품질 기준은 거의 없는 실정이다. 이것은 모션에 대한 시각적 영향을 모델링 하고 품질 기준에 적용하는 것이 복잡하고 어렵기 때문이다. 하지만 모션정보를 충분히 담고 있는 파라미터인 모션 벡터가 있으며 이는 블록마다 구해질 수 있다. 따라서 이를 파라미터로 이용하면 각 블록마다의 모션에 의한 시각적 중요도를 자연스럽게 구할 수 있다. 본 논문에서는 기본 실험에 의해서 하나의 블록이 갖는 모션 벡터의 크기에 대한 민감도를 모델링 하였다. 모션의 크기가 큰 신호는 HVS에 정확히 인지되지 않는다는 가설을 바탕으로 exhaustive한 실험을 통해 모델링 하였다. 일반적으로 시각적 모델에서 많이 사용되는 사인파 그레이팅 (sine wave grating)을 이용하여 모션 벡터의 크기에 따라 달라지는 인지 가능한 콘트라스트의 크기를 정규화 하여 중요도를 구하였다. 따라서 모션 벡터가 큰 블록이 담고 있는 신호는 HVS에 정확히 인식되지 않기 때문에 여기에서의 구조적 손상도 큰 영향을 미치지는 못하며 이는 그 모션 벡터의 시각적 중요도가 될 수 있다. 다음의 수식은 실험에 의해 구한 모델링 함수이다.

$$w_M^k=\frac{-1.54762|MV_k+37|}{37} \quad (7)$$

하지만 비디오 프레임 안에는 하나의 object만이 모션을 갖는 상황만 있는 것은 아니다. 때로는 무수히 많은 object

들의 모션이 비디오 프레임 안에서 동시에 존재할 수도 있고 카메라의 움직임에 의해서 화면 전체가 움직이는 상황이 있을 수 도 있다. 문제를 간단히 하기 위해서 일단 scene change가 발생하는 상황은 고려하지 않겠다. 화면상의 어떠한 object가 같은 움직임을 갖더라도 정지된 배경 위에 있느냐 아니면 복잡한 움직임을 갖는 배경위에 있느냐에 따라서 HVS에 인식되는 정도는 다를 것이며 주위의 복잡한 움직임은 특정 object의 움직임의 인지에 간섭을 유발할 것이다. 따라서 위의 model에 frame의 종합적인 움직임의 정보가 추가적으로 고려되어야 하며 다음의 식에서처럼 한 frame의 평균 MV의 크기를 함께 고려하여 MV weight를 구하였다.

$$w_{MV}^{k} = \frac{1}{2}\left(\frac{1+w_{MV}^{k}}{2-\overline{w_{MV}^{k}}}\right)$$
$$\overline{w_{MV}^{k}} = \frac{-1.54762\left|\overline{MV_k}\right|+37}{37} \quad (8)$$

$\overline{w_{M}^{k}}$는 평균 모션 벡터를 모델링 함수에 대입하여 구한 값이다.

IV. Experiment Result

VQEG Multimedia (MM) Group에서는 영상 품질 메트릭의 성능을 검증하기 위해 주관적 화질 평가의 결과와 비교를 하기위한 테스트 플랜을 제공한다. 본 논문에서는 이 테스트 플랜을 바탕으로 주관적 화질 평가를 실시했으며 제안하는 영상 품질 메트릭의 성능을 평가하였다. 제안하는 영상 품질 메트릭인 VSSIM의 개선된 성능을 보이기 위해 최근 이슈가 되고 있는 SSIM과 고전적으로 널리 사용되던 PSNR과 비교를 하였다. 테스트를 위해 총 18개의 테스트 시퀀스가 사용되었으며 28명의 관찰자가 동원되었다. 각 테스트 시퀀스는 H.263 코덱을 이용하여 여러 가지 다양한 비트율과 프레임 율을 적용하여 인코딩 하였다. 주관적 화질 평가를 위해서는 adjectival categorical rating (ACR) 방식을 이용하였는데 이는 테스트 영상을 한번씩만 보여주

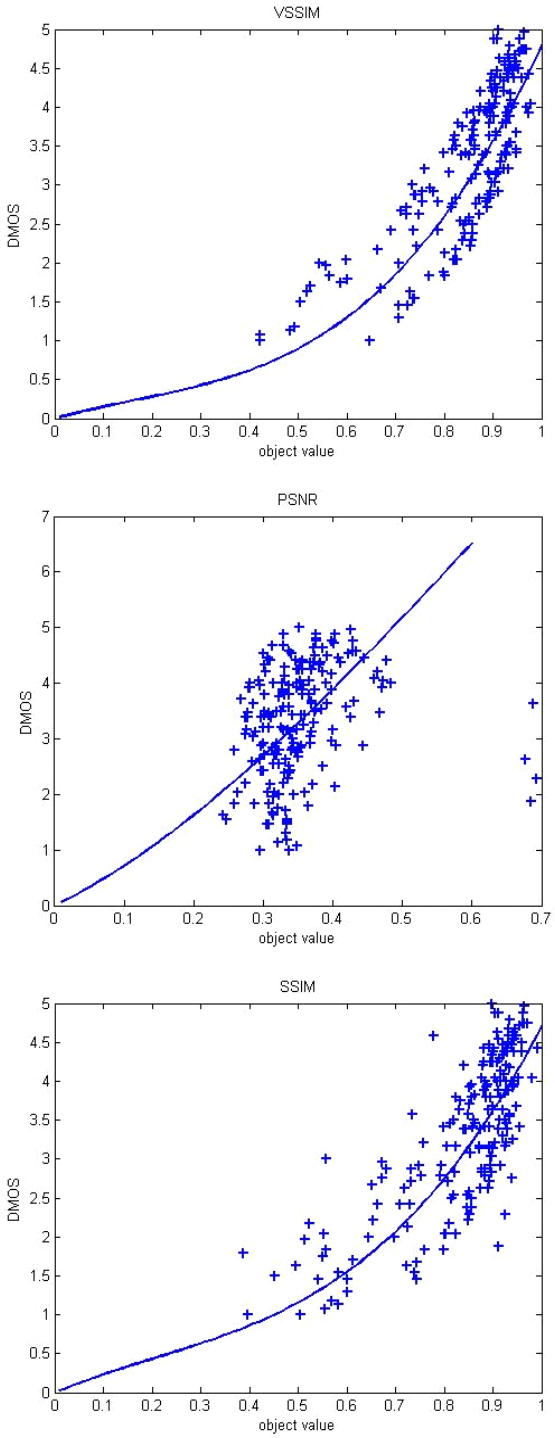

그림 2. 스캐터 플롯
Fig. 2. Scatter plot comparison based on the VQEG MM test plan

고 Excellent, Good, Fair, Poor, Bad 중에서 하나에 관찰자가 보팅하는 방식으로써 VQEG Multimedia (MM) Group에서 주관적 화질 평가를 위해 사용하는 방식이다. 그림 2는 주관적 화질 평가를 통해 구한 스캐터 플롯이다.

이 스캐터 플롯에서 본 논문에서 제안하는 영상 품질 측정 기준인 VSSIM의 결과와 주관적 화질 결과인 DMOS와의 상관도가 기존의 SSIM과 PSNR 에서의 경우보다 월등함을 쉽게 볼 수 있다. 이러한 스캐터 플롯을 이용하여 객관적 결과와 상관도를 수치적으로 나타내는 메트릭으로는 pearson linear correlation coefficient, spearman's rank-order correlation coefficient, outlier ratio를 이용하였다 [10]-[13]. Pearson linear correlation coefficient은 주관적 화질 결과와 객관적 결과와의 상관도를 구하는 기준이며 spearman's rank-order correlation coefficient는 두 결과의 단조성 혹은 일관성을 측정하기 위한 기준이다. Outlier ratio는 주관적 테스트 결과가 기준 값으로부터 일정 범위 이상 멀어진 결과들의 비율을 구하는 기준이다. 이러한 3가지 기준을 이용하여 비교한 결과는 그림 3과 같다. 그림에서 볼 수 있듯이 Peason correlation의 경우 기존의 SSIM 보다 0.02 정도가 높으며 Spearman's Rank-order correlation의 경우 SSIM 0.04 높음으로써 개선된 성능을 보인다. Outlier의 경우 SSIM과 동일하게 안정된 결과를 보인다.

여기에서 비교한 3가지 결과 중 Spearman's Rank-order correlation에서 특히 높은 개선을 보이는데 이는 본 논문에서의 가중치를 적용함으로써 주관적 화질 평가와 객관적 화질 평가간의 일관성이 높아졌기 때문이다. 다시 말하면 기존 SSIM에서 전체적인 구조적 왜곡의 정도가 비슷한 점수를 산출하였다 하더라도 구조적 왜곡의 분포가 시각적인 민감도에 따라 다를 때 실제 시각적으로도 다른 화질을 보이게 된다. 예를 들어 어떤 프레임에서 전체적인 SSIM 점수가 같다 하더라도 높은 SSIM 점수가 관심 영역에 집중적으로 분포되어 있는지 아니면 관심영역에서 벗어난 주변부에 분포되어 있는지에 따라 실제 시각적으로 화질의 차이가 있게 된다. 따라서 이러한 시각적 민감도에 영향을 줄 수 있는 중요한 요인을 본 논문에서는 spatial weight, IAF weight, MV weight로 한정하고 논지를 진행 했으며 이외의 시각적 특성들이 추가된다면 더 많은 상황을 고려할 수 있게 되고 결과적으로 성능을 더욱 개선시킬 수 있다.

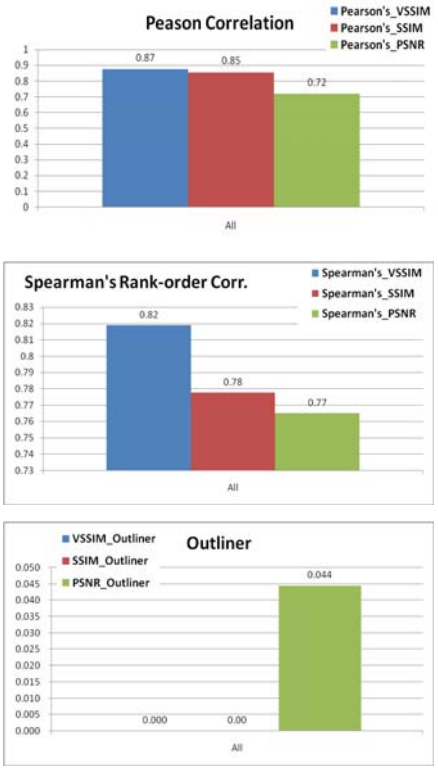

그림 3. 주관적 화질 평가 결과
Fig. 3. Performance comparison of the video quality assessments based on the VQEG MM test plan

V. Conclusion

영상 통신의 비중이 커지고 있는 상황에서 품질 측정은 매우 중요한 기술이다. 이를 위해 많은 품질기준들이 학계와 산업계를 통해 제안되었지만 아직 만족할만한 성능을 보여주지 못하는 실정이다. 최근 에러의 가시도를 측정하던 기존 패러다임의 한계를 극복하고자 structural similarity (SSIM)이 제안되어 우수한 성능을 보이고 있다. 하지만 SSIM은 기존에 활발히 연구되어오던 인간 시각 체계의 민감도에 대한 특성을 완전히 배제함으로써 새로운 한계점을 노출한다. 본 논문에서는 포비에이션, 밝기, 콘트라스트, 움직임 정보를 이용하여 SSIM과 자연스럽게 결합 함으로써 성능을 개선하였다. VQEG multimedia group의 test data

set을 이용한 test는 본 논문의 metric이 기존의 SSIM 보다 주관적 화질평가의 결과와 상관도가 더 높음을 보여준다. 추후에 심리 물리학의 인간 시각 체계에 대한 발전된 결과가 나오면 더욱 정확한 인간시각체계에 의해 추가적인 성능 향상도 기대할 수 있다.

참 고 문 헌

[1] Z. Wang, A. C. Bovik, “Modern Image Quality Assessment,” Morgan & Claypool Publishers, 2006.

[2] Z. Wang, H. R. Sheikh and A. C. Bovik, “Objective Video Quality Assessment” in The Handbook of Video Databases: Design and Application, B. Furht and O. Marqure (Editors) , Boca Raton, Florida: CRC Press, pp. 1041-1078, Sept. 2003.

[3] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image Quality Assessment: From error visivility to structural similarity,” IEEE Trans. Image Processing , vol. 13, no. 4, pp. 600-612, Apr. 2004.

[4] Zhou Wang, Ligang Lu and A. C. Bovik “Video quality assessment based on structual distortion measurement,” Signal Processing Image Communication, vol.19, no.2, pp.121-132, Feb. 2004.

[5] S. Lee, M. S. Pattichis and A. C. Bovik, “Foveated video quality assessment,” IEEE Trans. Multimedia, pp. 129-132, vol. 4, Mar. 2002.

[6] Z. Wang and A. C. Bovik, “Embedded Foveation Image Coding,” IEEE Trans. Image Process., pp. 1397-1410, vol. 10, no.10, Oct. 2001.

[7] J. Malo, A. M. Pons, and J. M. Artigas, “Subjective image fidelity metric based on bit allocation of the human visual system in the DCT domain,” Image Vis. Comput., vol. 15, no. 7, pp. 535.548, 1997.

[8] J. Malo, J. Gutierrez, I. Epifanio, F. Ferri, and J. M. Artigas, “Perceptual feed-back in multigrid motion estimation using an improved DCT quantization,” IEEE Trans. Image Process., vol. 10, no. 10, pp. 1411.1427, Oct. 2001.

[9] S. Winkler, “Issues in vision modeling for perceptual video quality assessment,” Singnal Processing, vol. 78, pp. 231-252, Oct. 1999.

[10] ITU-R, “Methodology for the subjective assessement of the quality of television pictures,” Recommendation ITU-R BT.500-11, 2002

[11] ITU-T, “Subjective video quality assessment methods for mulitmedia applications,” Recommendation ITU-T P.910, 2002

[12] http://www.vqeg.org

[13] Video Quality Expert Group, “Multimedia Group Test Plan,” 2006.

[14] M. Pinson and S. Wolf. “A New Standardized Method for Objectively MeasuringVideo Quality, IEEE Transactions on Broadcasting, VOL. 50, NO.3, pp. 312-322, Sept., 2004.

저 자 소 개

박 진 철

- 2006년 : 숭실대학교 전기전자공학과 학사
- 2008년 : 연세대학교 전기전자공학과 석사
- 2008년 ~ 현재 : 연세대학교 전기전자공학과 박사과정
- 주관심분야 : 무선네트워크, 멀티미디어 통신, 영상화질평가

이 상 훈

- 1989년 : 연세대학교 전자공학과 학사
- 1991년 : 한국과학기술원 전기전자공학과 석사
- 1991년 ~ 1996년 : KT 연구개발센터 연구원
- 2000년 : The University of Texas at Austin, Electrical Engineering 박사
- 2000년 ~ 2002년 : Lucent Technologies 연구원
- 2003년 ~ 현재 : 연세대학교 전기전자공학부 부교수
- 주관심분야 : 무선네트워크, 멀티미디어 통신, 센서 네트워크