움직임 분석 기반의 시각인지 모델을 이용한 비디오 코딩 방법

오 형 석^{a)}, 김 원 하^{a)*}

Video Coding Method Using Visual Perception Model based on Motion Analysis

Hyung Suk Oh^{a)}, and Wonha Kim^{a)*}

요 약

본 논문에서는 인간 인지 기반 비디오 코딩을 위한 비디오 처리 방법을 개발한다. 제안하는 방법은 율-왜곡(rate-distortion) 최적화의 영향뿐만 아니라 제한적인 시, 공간 해상도, 지역적인 움직임 이력(history), visual saliency에 의한 인간 시각 인지를 고려한다. 이러한 인간의 인지적인 효과들을 고려하기 위하여 본 논문에서는 움직임 패턴을 모델링하고 Hedge 알고리즘을 사용하여 움직임 패턴을 결정하는 기법을 개발한다. 그 다음, 제안한 움직임 패턴과 기존의 visual saliency와의 결합을 통하여 인간 시각 인지 모델을 수립한다. 제안된 인간 시각 인지 모델을 구현하기 위하여 기존의 foveation filtering 방법을 확장한다. 시각적 자극이 덜한 지역만을 부드럽게(smoothing)하는 기존의 foveation filtering 기법과 비교하여 제안하는 foveation filtering 기법은 인간 시각 인지 모델에 따라 지역적으로 부드럽게 또는 지역적 특성을 향상시킴으로써, 시각적 자극이 덜한 지역에서 줄여진 대역폭을 효과적으로 시각적 자극이 큰 지역에서 사용하도록 이동 시킬 수 있는 장점이 있다. 제안된 방법의 성능은 전반적인 비디오 화질을 만족할 뿐만 아니라 인간이 인지하는 화질의 품질을 12%~44% 향상시킨다.

Abstract

We develop a video processing method that allows the more advanced human perception oriented video coding. The proposed method necessarily reflects all influences by the rate-distortion based optimization and the human visual perception that is affected by the visual saliency, the limited space-time resolution and the regional moving history. For reflecting the human perceptual effects, we devise an online moving pattern classifier using the Hedge algorithm. Then, we embed the existing visual saliency into the proposed moving patterns so as to establish a human visual perception model. In order to realize the proposed human visual perception model, we extend the conventional foveation filtering method. Compared to the conventional foveation filter only smoothing less stimulus video signals, the developed foveation filter can locally smooth and enhance signals according to the human visual perception without causing any artifacts. Due to signal enhancement, the developed foveation filter more efficiently transfers the bandwidth saved at smoothed signals to the enhanced signals. Performance evaluation verifies that the proposed video processing method satisfies the overall video quality, while improving the perceptual quality by 12%~44%.

Keyword : visual saliency, human visual perception, video coding, hedge

a) 경희대학교 전자정보대학 전자전파공학과

College of Electronics and Radio Engr. Kyung Hee University

* 교신저자 : 김원하 (wonha@khu.ac.kr)

※ 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.20110027611)

· 접수일(2011년12월28일), 수정일(1차:2012년2월21일, 2차:3월9일), 게재확정일(2012년3월21일)

1. 서론

비디오 코덱은 율-왜곡(rate-distortion)을 최적화하기 위하여 복잡한 지역이나 움직임이 많은 지역에 더 많은 대역폭을 할당함으로써 최적의 비디오 화질을 제공한다. 그러나 주관적인 화질 테스트에서는 율-왜곡 최적화와 관련된 화질 왜곡 측정 방법으로 계산한 화질이 인간이 실제로 인지하는 화질과 차이가 있다는 것을 쉽게 알 수 있다^[1]. 이는 인간 시각 시스템 (Human Visual System: HVS)이 율-왜곡 최적화를 통한 전반적인 비디오 화질과 인간이 무의식적으로 인지하는 시각적 자극이 관련되어 있는 것으로 분석할 수 있다^[2-4]. 따라서 전반적으로 최적의 화질을 만족하고 동시에 시각적 자극이 높은 지역의 해상도를 유지하는 인간 인지 기반 비디오 코딩 방법이 필요하다.

시각적 자극이 높은 지역의 좋은 화질이 더 좋은 시각적

화질을 제공한다는 것을 기반으로 기존의 인간 인지 기반 비디오 코딩 방법들은 시각적 자극이 낮은 지역의 고주파 성분을 제거하거나 또는 양자화 파라미터 (QP)을 조절하여 시각적 자극이 더 높은 지역의 대역폭을 많이 할당하도록 하고 있다^[5,6]. 따라서 기존의 visual saliency 모델들은 특정한 지역에서 인간의 시각적 자극이 무의식적으로 얼마나 높게 반응하는지를 측정하는 방법을 제안하고 있다. Walther et al.는 시각적으로 자극적인 특성, 즉, 색깔, 휘도와 지역적 방향 정보들을 이용하였다^[7]. 이 방법은 움직임의 특성을 고려하지 않았기 때문에 정지영상의 분석에는 유용하나, 비디오 처리에는 비효율적이다. 움직임의 특성을 분석한 motion saliency의 대표적인 연구는 Jacobson et al., Harel et al.과 Ma et al.이다^[8-10]. Jacobson은 자동으로 saliency 지역들을 찾는 discriminant center-surround 모델-중심 블록과 주위 블록사이의 상호 정보량(mutual in-

그림 1. 눈 추적기로 측정된 saliency 맵, (a) 'Soccer' 비디오의 saliency 맵 (b) 'Soccer' 비디오

Fig. 1. The saliency map obtained by an actual eye tracker, (a) saliency map in 'Soccer' sequence (b) the corresponding frame

그림 2. 'Mother and daughter' 비디오에서 눈 추적기로 측정된 saliency 맵과 [9]의 saliency 모델에 의한 saliency 맵과의 비교

Fig. 2. Comparison of the saliency map by an actual eye tracker and the saliency map generated by the saliency model in [9] for the 'Mother and daughter' sequence

formation)을 계산하여 saliency를 찾는 방법-을 이용하였다^[8]. Harel은 움직임과 화소 크기가 주변 지역들과 유사하지 않은 지역에 더 높은 saliency를 주었다^[9]. Ma는 인간의 관심 지역은 움직임이 크고 움직임의 변화가 크며, 공간적으로 동일한 움직임 특성이 있는 지역으로 관찰하였다^[10]. 따라서 Ma가 제안한 saliency 모델은 한 지역의 움직임 크기와 움직임의 변화들을 엔트로피로 계산하였다.

기존의 motion saliency 모델뿐만 아니라 인간의 시각은 제한적인 시,공간 해상도와 단기 움직임 기억으로부터 영향을 받는다^[11,12]. 인간 시각 시스템(HVS)이나 영상 기기들은 단위 시간당 획득할 수 있는 정보가 한정되어 있기 때문에 움직임의 속도가 커지면 획득하는 시간적 정보의 양이 증가하는 반면에 공간적 정보의 양은 감소하게 된다^[11]. 즉, 텍스트쳐 에너지는 움직임이 커질수록 텍스트쳐 에너지가 감소되어 흔들림(blurring)이 발생된다. <그림 1>(a)는 눈 추적기(eye tracker)로 측정된 saliency 맵을 나타내며, 흰색은 가장 높은 saliency를 나타낸다. <그림 1>(b)에서 보는 것과 같이 큰 움직임이 있는 지역이 가장 높은 자극을 나타내지만, 빠른 움직임으로 인한 흔들림(blurring) 때문에 화질이 좋지 않은 것을 볼 수 있다. 따라서 시각적 자극이 높은 빠른 움직임 지역과 비디오 화질의 관계는 반드시 일치하는 것은 아니다. 또한, 인간은 지역적 움직임을 단기적으로 기억하기 때문에 움직임으로 인한 단기 기억을 고려해야 한다. <그림 2>는 눈 추적기에 의한 saliency 맵과 [9]의 saliency 모델을 통한 saliency 맵을 비교한다. <그림 2>의 위쪽 그림에서 보듯이 아이의 얼굴은 움직임과 멈춤을 반복하며, 아이의 얼굴이 멈춘다고 할지라도 인간은 그 얼굴을 계속 관찰하는 특성이 있는 것을 알 수 있다. 그러나 <그림 2>의 아래쪽 그림에서는 아이 얼굴의 움직임이 없을 때 saliency를 유지하지 못하는 것을 알 수 있다. 따라서 인간에게 더 적당한 visual saliency 모델은 움직임으로 인한 단기 기억을 반영해야 한다^[12].

본 논문에서는 인간 인지에 더 적당한 비디오 코딩 방법을 제공하는 비디오 처리 방법을 제안한다. 먼저, visual saliency, 제한적인 시,공간 해상도와 단기 움직임 기억을 모두 고려하는 새로운 움직임 패턴을 제안한다^[13]. 또한 움직임 패턴을 실시간으로 결정하기 위하여 Hedge 알고리즘을 사용한다^[14]. 그 다음, 제안된 움직임 패턴과 기존의 sali-

ency 모델을 결합하여 시각 인지 모델(visual perception model)을 개발한다. 개발된 시각 인지 모델은 기존의 foveation filtering 방법을 개선하여 구현한다. 기존의 foveation filtering 방법은 오직 smoothing만을 수행하는 반면, 본 논문에서 제안하는 foveation filtering 방법의 특징은 제안한 시각 인지 모델에 따라서 지역적 특성을 부드럽게 또는 향상을 동시에 수행하는 것이다^[6]. 따라서 제안하는 foveation filter는 시각적 자극이 덜한 지역에서 줄여진 대역폭을 좀 더 효율적으로 시각적 자극이 큰 지역으로 이동시킬 수 있다.

논문의 구성은 다음과 같다. 2장에서는 새로운 움직임 패턴을 제안하며, Hedge 알고리즘을 사용하여 제안한 움직임 패턴을 실시간으로 결정하는 방법을 제시한다. 3장에서는 제안한 움직임 패턴과 기존의 saliency를 결합한 시각 인지 모델을 개발하며, 시각 인지 모델을 이용한 비디오 처리 방법을 제시한다, 4장에서 개발된 비디오 처리 방법의 성능을 검증한다. 마지막으로 5장에서 본 논문의 결론을 도출한다.

II. 움직임 패턴 모델

본 절에서는 인간 인지에 따라서 움직임 패턴을 모델링하고, 움직임 패턴을 실시간 학습 방법인 Hedge 알고리즘을 사용하여 실시간으로 움직임 패턴을 결정한다.

1. 인간 인지 기반 움직임 패턴

인간은 움직임 변화와 움직임의 속도에 의한 텍스트쳐 에너지의 변화에 따라 움직임 패턴을 끊임없이 업데이트하며, 업데이트된 움직임 패턴을 다시 기억하게 된다. 이를 바탕으로 움직임 패턴을 normal mode, fast mode, hysteretic mode로 분류한다^[13]. <그림 3>은 제안하는 움직임 패턴 분류를 보여준다.

움직임이 있는 지역들은 일반적인 움직임이 있는 지역과 빠른 움직임이 있는 지역으로 분류된다. 인간 시각 시스템(HVS)이나 영상 기기들은 단위 시간당 획득할 수 있는 정보가 한정되어 있기 때문에 움직임의 속도가 커지면 획득

그림 3. 제안하는 움직임 패턴 분류

Fig. 3. Case study of the proposed moving pattern classification

하는 시간적 정보의 양이 증가하는 반면에 공간적 정보의 양은 감소하게 된다^[11]. 즉, 텍스처 에너지는 움직임이 커질수록 텍스처 에너지가 감소되어 흔들림이 발생되며, 공간 정보가 많은 복잡한 지역에서는 적은 움직임에도 흔들림이 발생된다. 따라서 일반적인 움직임이 있는 지역들은 움직임에 의하여 흔들림이 발생하지 않는 지역인 반면에 빠른 움직임이 있는 지역은 흔들림이 발생하는 지역이다. 또한, 인간은 기억된 움직임 패턴을 상기시키며 지역을 인지한다^[12]. 따라서 이전에는 움직임이 있었지만 현재 움직임이 없는 지역에서 인간은 잠시 동안 그 지역이 계속 움직이고 있다고 생각하게 되고, 그 지역의 움직임 패턴을 이전의 움직임 패턴으로 간주하게 된다. 만일 그 지역의 움직임 패턴이 기억에서 사라지기 전까지 움직임이 감지되지 않는다면 그 지역은 배경이 된다. 따라서 hysteretic mode는 움직임 패턴이 기억에서 사라지기 전까지의 움직임이 없는 지역의 모드이다.

2. Hedge 기반 움직임 패턴 결정

본 절에서는 움직임 패턴을 결정하는 방법을 개발한다. 움직임 정보와 텍스처 에너지 변화와 같은 움직임 특성들은 각 시간에서 움직임 상태를 결정한다. 인간은 시간적으로 변화하는 움직임 상태에 따라서 움직임 패턴을 연속적으로 갱신하기 때문에 움직임 패턴은 움직임 상태의 이력(history)과 관련이 깊다. 즉, 인간이 인지하는 움직임 패턴은 실시간 학습 방법으로 결정되어야 하며, 이를 위하여 본 논문에서는 실시간 학습 방법으로 Hedge 알고리즘을 사용한다^[14]. <그림 4>는 움직임 패턴을 결정하기 위한 제안하는 체계를 보여준다. 제안하는 방법은 블록의 움직임 특성들을 이용하여 블록의 움직임 상태를 결정하고, Hedge 알고리즘으로 블록의 움직임 상태를 실시간으로 학습하여 블록의 움직임 패턴을 결정한다. 그 다음, 동일한 움직임 패턴에 속하는 주위 블록들을 무리 지어(clustering) 객체의 움직임 패턴을 결정한다.

그림 4. 움직임 패턴을 결정하기 위하여 제안하는 체계

Fig. 4. The proposed framework for determining moving patterns

t 번째 frame에서 k 번째 블록을 B_k^t 라 한다. B_k^t 의 움직임 벡터를 MV_k^t 라 하고, s_k^t 를 B_k^t 의 움직임 상태라 한다. 움직임 상태는 normal state(NS), fast state(FS), stationary state(SS)로 구분한다. Normal state인 블록은 motion blur 현상이 없고 움직임이 있는 블록이며, fast state인 블록은 움직임에 의한 motion blur 현상이 있는 블록이다. Stationary state는 움직임이 없는 블록이다.

Motion blur 현상이 있는 지역을 간단하고 효율적으로 감지하는 방법은 지역의 텍스처 에너지와 시간적 잔여 에너지(temporal residual energy)의 비율을 측정하는 것이다. 시간적 잔여 에너지는 임의의 지역과 그 지역의 움직임 보상 지역의 차이 에너지를 뜻한다. Motion blur 현상은 지역의 텍스처 에너지를 없애기 때문에 motion blur 현상이 보이는 지역에서 텍스처 에너지는 시간적 잔여 에너지보다 작아지게 된다. Fast mode는 motion blur 현상이 있는 지역에서 나타나게 되고, normal mode는 motion blur 현상이 없는 지역에서 나타나게 된다. 따라서 motion blur 현상이 있는 블록의 텍스처 에너지가 시간적 잔여 에너지보다 작다는 특성을 이용하여 텍스처 에너지와 시간적 잔여 에너지사이의 비율을 계산한다. 블록 B_k^t 의 energy 비율 R_k^t 는 다음과 같이 계산된다^[15].

$$R_k^t = \frac{\text{Temporal residual energy}}{\text{Texture energy}} = \frac{\frac{1}{N \times N} \left(\sum_{q \in B_k^t} \{B_k^t(q) - B_k^t(q + MV_k^t)\} \right)^2}{\left(\frac{1}{N \times N} \sum_{q \in B_k^t} \{B_k^t(q)\}^2 \right) - \left(\frac{1}{N \times N} \sum_{q \in B_k^t} \{B_k^t(q)\} \right)^2} \quad (1)$$

여기서, q 는 B_k^t 의 내부 화소이다.

에너지 비율 R_k^t 과 MV_k^t 를 이용하여 움직임 상태 s_k^t 는 다음과 같이 구분된다.

$$s_k^t = \begin{cases} \text{Normal state (NS)}, & \text{if } MV_k^t \neq 0, R_k^t \leq 1 \\ \text{Fast state (FS)}, & \text{if } MV_k^t \neq 0, R_k^t > 1 \\ \text{Stationary state (SS)}, & MV_k^t = 0 \end{cases} \quad (2)$$

각 시간에서 발생하지 않는 움직임 상태를 수치화하기

위하여 l_k^t 을 시간 t 에서 k 번째 블록의 nonoccurrence 벡터라 정의한다. 또한 주어진 구간 $[t-T, t]$ 에서 관측된 움직임 상태의 이력을 기록하기 위하여 기억 벡터(record vector) L_k^t 을 정의한다. 즉,

$$l_k^t = \begin{cases} [0, 1, 1], & \text{if } s_k^t = NS \\ [1, 0, 1], & \text{if } s_k^t = FS \\ [1, 1, 0], & \text{if } s_k^t = SS \end{cases} \text{이고, } L_k^t = \sum_{r=t-T}^t l_k^r \quad (3)$$

여기서 T 는 주어진 구간의 크기이다. 그러므로 $L_k^t(0)$, $L_k^t(1)$, $L_k^t(2)$ 는 각각 발생하지 않은 normal state, fast state, stationary state의 개수이다.

각 움직임 상태의 이력에 근거하여 Hedge 알고리즘은 자주 발생하는 상태의 확률을 지속적으로 증가시킴으로써 움직임 상태의 확률을 업데이트한다. $w_k^t = [w_{0,k}^t, w_{1,k}^t, w_{2,k}^t]^T$ 를 비중 벡터라 하고, $w_{0,k}^t$, $w_{1,k}^t$, $w_{2,k}^t$ 를 각각 normal state, fast state, stationary state의 비중이라 한다. Hedge 알고리즘은 각 움직임 상태의 비중을 다음과 같이 업데이트 한다.

$$w_k^t = \begin{bmatrix} w_{0,k}^t \\ w_{1,k}^t \\ w_{2,k}^t \end{bmatrix} = \begin{bmatrix} w_{0,k}^0 \cdot \exp^{-\eta L_k^t(0)} \\ w_{1,k}^0 \cdot \exp^{-\eta L_k^t(1)} \\ w_{2,k}^0 \cdot \exp^{-\eta L_k^t(2)} \end{bmatrix} \quad (4)$$

여기서 $w_{0,k}^0$, $w_{1,k}^0$, $w_{2,k}^0$ 는 각 움직임 상태의 초기 비중이고 $w_{0,k}^0 + w_{1,k}^0 + w_{2,k}^0 = 1$ 이다. η 는 학습률(learning rate)이며, $\eta > 0$ 이다.

각각의 비중들은 각 움직임 상태의 확률을 추정하기 위하여 정규화된다. 확률 벡터 p_k^t 는 다음과 같이 계산된다.

$$p_k^t = \begin{bmatrix} p_{0,k}^t \\ p_{1,k}^t \\ p_{2,k}^t \end{bmatrix} = \frac{1}{\|w_k^t\|} \cdot \begin{bmatrix} w_{0,k}^t \\ w_{1,k}^t \\ w_{2,k}^t \end{bmatrix} \quad (5)$$

여기서 $p_{0,k}^t$, $p_{1,k}^t$, $p_{2,k}^t$ 는 각 움직임 상태에 해당하는 확률이다.

따라서 식(5)를 이용하여 움직임 패턴을 결정한다. 움직임 패턴은 현재 움직임 상태와 각 움직임 상태의 발생 패턴으로부터 만들어진다. 현재 움직임은 블록의 경우, 블록의 움직임 패턴은 normal state의 상태가 연속적으로 관측될 경우에 normal mode(NM)인 반면 fast state의 상태가 연속적으로 관측될 경우에 블록의 움직임 패턴은 fast mode(FM)이 된다. 또한 현재 움직이지 않는 블록의 경우, 주어진 구간에서 대부분의 움직임 상태가 normal 또는 fast state라면 블록의 움직임 패턴은 hysteretic mode(HM)이 된다. 그 외의 패턴은 background(BG)가 된다. 그러므로 가장 높은 확률을 가지는 움직임 상태와 현재 움직임 상태에서부터 움직임 패턴 BMP_k^t 는 다음과 같다.

$$BMP_k^t = \begin{cases} \text{NM if } \max\{p_{i,k}^t\} = p_{0,k}^t, s_k^t \neq SS \\ \text{FM if } \max\{p_{i,k}^t\} = p_{1,k}^t, s_k^t \neq SS \\ \text{HM if } \max\{p_{i,k}^t\} \neq p_{2,k}^t, s_k^t = SS \\ \text{BG if } \max\{p_{i,k}^t\} = p_{2,k}^t, s_k^t = SS \end{cases} \quad (6)$$

각 프레임의 전 블록들의 움직임 패턴을 결정한다. 다음 객체는 유사한 움직임 패턴을 지닌 블록을 무리를 지어 구성한다. 동일 객체에 속하는 블록들을 구성하는 기준으로는 서로 연결되어 있고, 블록의 움직임 상태가 동일하고 움직임 방향이 비교적 유사한 것으로 하였다. 움직임 방향의 유사성을 결정하는 임계치 값은 fuzzy 2-partition 방법을 사용하였다^[16]. 이러한 방법을 적용하고 나서 노이즈 블록과 잘못 처리된 블록을 처리하기 위하여 블록을 한 화소단위로 하여 dilation, erosion 과 hole filling의 morphological processing을 적용하였다^[17]. <그림 5>는 객체와 객체의 움직임 패턴 결과를 보여준다.

3. 움직임 패턴 결정의 성능

제안한 움직임 패턴 결정의 성능은 시청자의 주관에 달려 있기 때문에 비디오의 원 움직임 패턴을 생성하기 위하여 실험 참가자 30명을 모집하였다. 실험에서 사용된 비디오는 "silent", "paris", "ice", "pedestrian", "night"이며, 비디오에서 실험 참가자들 중 90% 이상 공통으로 선택된 지역을

원 움직임 패턴으로 하였다. 움직임 패턴 검출의 정확성을 판단하기 위하여 precision-recall을 사용하였다^[18]. R_{gt} 와 R_{pr} 을 각각 원 움직임 패턴 블록 세트, 제안 방법에 의하여 검출된 블록 세트라 한다. 제안 방법의 의하여 검출된 움직임 패턴 지역의 세트는 $R_{gt} \cap R_{pr}$ 이며, precision 확률 p 와 recall 확률 r 은 다음과 같이 계산된다.

$$p = \frac{|R_{gt} \cap R_{pr}|}{|R_{pr}|}, r = \frac{|R_{gt} \cap R_{pr}|}{|R_{gt}|} \quad (7)$$

Precision 확률은 제안 방법이 얼마나 정확하게 움직임 패턴을 검출하는지를 나타내는 반면, recall 확률은 제안 방법이 얼마나 많이 움직임 패턴을 검출하는지를 나타낸다. 따라서 precision 확률 p 와 recall 확률 r 을 동시에 결합하여 F score로 표현한다. F score는 다음과 같다.

$$F = \frac{2pr}{p+r} \quad (8)$$

그림 5. 객체 clustering과 morphological processing에 의하여 결정된 객체와 그 객체의 움직임 패턴

Fig. 5. Demonstration of the moving patterns of objects determined by using object clustering and morphological processing

F-score가 1로 다가갈수록, 알고리즘의 검출 정확성은 증가한다. <표 1>에서 보듯이, 제안 방법의 F-score는 0.9보다

높다는 것은 제안 방법의 정확성이 90%이상이라는 것을 뜻하는 것으로써 제안한 움직임 패턴 결정은 충분히 정확하다고 판단할 수 있다.

표 1. 테스트 비디오의 F-score

Table 1. F-scores for the test video sequences

Sequences	Normal Pattern	Aliased Pattern
Paris	94%	95%
Ice	93%	97%
Silent	94%	97%
Pedestrian	91%	94%
Night	95%	96%

III. Band-limited 비디오 코딩을 위한 처리 방법

본 절에서는 전반적인 비디오 화질, 무의식적으로 인지

하는 시각적 자극, 객체의 해상도에 민감한 인간의 시각적 인지에 따라서 최적하게 대역폭을 사용하는 비디오 처리 방법을 개발한다.

1. 제안하는 시각 인지 모델

기존의 인간 인지 기반 비디오 코딩 방법이 시각적 자극 기반 saliency만을 고려하는 반면, 제안하는 인간 인지 기반 비디오 코딩 방법은 전반적인 비디오 화질, 무의식적으로 인지하는 시각적 자극, 객체의 해상도를 모두 고려한다. 따라서 본 논문에서는 제안한 움직임 패턴과 기존의 saliency 모델을 결합한 새로운 시각 인지 모델을 개발한다. 제안하는 시각 인지 모델은 일반적인 움직임이 있는 지역에 saliency를 할당하고, 빠른 움직임이 있는 지역의 saliency를 제외시키며, hysteretic 움직임 지역에는 이전의 감지된 saliency를 할당하는 방법이다.

<그림 6>과 <그림 7>은 기존의 saliency 맵과 제안하는 시각 인지 모델에 의한 saliency 맵을 비교한다. 기존의 sali

그림 6. Saliency 맵의 비교, (a) 'Night' 영상 (b) [9]의 saliency 모델로 측정된 saliency 맵과 그에 상응하는 자극 (c) 제안 방법에 의하여 측정된 시각 인지 맵과 그에 상응하는 자극

Fig. 6. Comparison of the saliency maps by [9] and the visual perception maps by the proposed method, (a) original frames in 'Night' sequence, (b) Saliency maps by [9] and corresponding motion stimuli, (c) Visual perception maps by the proposed method and corresponding motion stimuli

그림 7. Saliency 맵의 비교, (a) 'Paris' 영상 (b) [9]의 saliency 모델로 측정된 saliency 맵과 그에 상응하는 자극 (c) 제안 방법에 의하여 측정된 시각 인지 맵과 그에 상응하는 자극

Fig. 7. Comparison of the saliency maps by [9] and the visual perception maps by the proposed method, (a) original frames in 'Night' sequence, (b) Saliency maps by [9] and corresponding motion stimuli, (c) Visual perception maps by the proposed method and corresponding motion stimuli

ency 방법으로 제안하는 새로운 시각 인지 모델에서는 graph based visual saliency (GBVS) 방법을 사용한다^[9]. <그림 6>에서 GBVS saliency 맵은 빠른 움직임 지역에 높은 saliency가 나타나는 반면, 제안 방법은 움직임 패턴에 의하여 빠른 움직임 지역을 구별하고, 구별된 빠른 움직임 지역에는 saliency를 할당하지 않는다. <그림 7>에서 GBVS saliency 맵은 움직임과 멈춤을 반복하는 왼쪽 사람의 얼굴의 saliency가 끊임없이 나타나지 않는 것을 볼 수 있다. 따라서 얼굴 인식과 같은 알고리즘이 없으면 기존의 saliency 모델로는 얼굴의 saliency를 유지할 수 없으나 제안 방법에 의하여 생성된 saliency 맵에서는 얼굴 인식과 같은 알고리즘 없이 얼굴의 saliency를 유지한다는 것을 알 수 있다. 그러므로 제안 방법이 인간 시각 인지에 더 적합하다고 판단할 수 있다.

2. 제안하는 시각 인지 모델의 구현

기존의 saliency 기반 비디오 코딩 방법은 foveation filter로 구현된다. Foveation filter의 기본 원리는 인간이 집중하는 지역으로부터 거리가 멀어질수록 시각적 민감도를 낮추는 것이다^[6]. 따라서 기존의 foveation filtering 방법은 인간이 집중하는 지역으로부터 거리가 멀어질수록 지역의 특성을 부드럽게 하면서 대역폭을 줄이는 것이다.

기존의 foveation filtering 방법과 다르게 제안하는 foveation filtering 방법은 빠른 움직임 지역 또는 배경을 부드럽게 함과 동시에 높은 saliency가 있는 지역에 선명도 향상 기법을 사용하여 지역적 특성을 향상시킨다. 빠른 움직임이 있는 지역은 motion blur 현상이 있기 때문에 지역

적 특성을 없애는 강한 필터를 적용한다. 배경과 같은 지역은 인간이 집중하는 지역은 아니지만 비디오의 전반적인 화질을 유지하는데 중요한 지역이므로 인간이 거의 인지하지 않는 범위 내에서 지역적 특성을 없애는 약한 필터를 적용한다. 또한, normal 움직임 지역 또는 hysteretic 움직임 지역에서는 saliency의 강도에 따라 지역적 향상 필터를 적용한다. 본 논문에 적용된 선명도 향상 기법은 객체의 자세 사향을 향상시키기 때문에 더 많은 대역폭이 그 지역에서 사용되기 때문에 제안하는 foveation filtering 방법은 좀 더 효율적으로 시각적 자극이 덜한 지역에서 줄여진 대역폭을 시각적 자극이 높은 지역으로 이동시킬 수 있는 장점이 있다.

제안하는 foveation filter를 구현하기 위하여, 본 논문에서는 DCT 영역에서 구현된 선명도 향상 기법으로 α -rooting 기법을 사용한다^[19]. α -rooting 기법은 고주파 성분을 증가시킴으로써 영상의 선명도를 향상시킨다. 반대로, 지역적 특성을 없애기 위하여 고주파 성분을 감소시킨다. 서로 다르게 처리되는 블록들 사이의 결함(artifact)을 방지하기 위하여 블록의 저주파 성분을 수정하지 않는다. $F(u, v)$ 를 $N \times N$ 블록에서 (u, v) 위치의 DCT 성분이라면, 선명도 향상 가중치 $w^\alpha(u, v)$ 는 다음과 같다.

$$w^\alpha(u, v) = \left| \frac{F(u, v)}{255} \right|^{\alpha-1}, \quad 0 < \alpha < 1 \quad (9)$$

여기서 α 는 선명도 강도를 조절하는 파라미터이다. α 가 작을수록 에지성분이 강조된다.

σ 를 가우시안의 편차라 하고, $G(u, v, \sigma)$ 를 (u, v) 위치의 가우시안 로우패스(low-pass) 성분이라 한다. 그러면, σ 가 클수록 고주파 성분들은 더 크게 제거된다. 따라서 제안하는 foveation filtering이 적용된 DCT 성분 $\tilde{F}(u, v)$ 은 다음과 같이 결정된다.

$$\tilde{F}(u, v) = \begin{cases} F(u, v) & \text{for } 0 \leq u + v \leq L \\ w^\alpha(u, v) \cdot F(u, v) & \text{for enhancing} \\ G(u, v, \sigma) \cdot F(u, v) & \text{for smoothing} \end{cases} \quad (10)$$

여기서 L 은 저주파 대역의 크기이며, L 보다 작은 지역

의 주파수 성분들은 블록킹 현상을 방지하기 위하여 조절하지 않는다.

IV. 실험 및 결과

제안 방법의 성능을 검증하기 위하여 기존의 saliency 기반 비디오 코딩 방법에 의하여 전처리된 비디오, 제안하는 방법에 의하여 전처리된 비디오를 인코딩하였으며, 각 디코딩된 비디오의 객관적 화질 평가와 주관적 화질 평가로 나누어 검증하였다. 사용된 인코더는 H.264/AVC이다^[20].

1. Foveation filter 세팅

$N \times N$ DCT 블록 크기로 16을 채택하였고, 저주파 대역의 크기 L 은 $\lceil 16/3 \rceil$ 로 하였다. Normal 움직임 지역과 hysteretic 움직임 지역에서의 선명도 향상 강도 α 는 saliency의 수치가 평균보다 높을 때 0.9로 세팅하였고 그 외 지역은 0.95로 하였다. Fast 움직임 지역의 경우 σ 는 2로, 배경은 0.5로 세팅하였다. 제안 방법과 기존의 방법의 공정한 비교를 위하여 기존의 방법에서 smoothing 강도는 3단계로 나누어 처리하였다. 평균보다 높은 saliency가 있는 지역에는 smoothing을 수행하지 않으며, 평균보다 낮은 지역의 σ 는 0.3으로 하였다. 또한 배경의 경우, 제안 방법과 같은 0.5로 세팅하였다.

2. 객관적인 화질 평가

객관적인 화질 평가를 위하여 평가 방법으로 Cumulative Probability of Blur Detection (CPBD)를 사용하였다^[21]. CPBD는 인간이 얼마나 영상을 깨끗하게 인지하는지를 수치적으로 나타내는 방법으로 각 에지에서 흔들림(blurring)을 감지하는 확률을 측정하고, 계산된 수치가 1에 가까울수록 비디오 화질은 더 선명하다고 판단하는 화질 평가 방법이다. 객관적인 화질 평가를 위하여, 본 논문에서는 3가지의 디코딩된 비디오를 평가하였다. 3가지의 디코딩된 비디오는 기존의 foveation filtering 방법으로 처리된 비디오, 제

안하는 방법으로 처리된 비디오, 전처리되지 않은 비디오이다. <표 2>는 평균 CPBD의 수치를 보여준다. 표에서 보듯이 제안 방법의 CPBD 수치가 가장 높은 것을 볼 수 있고, 같은 bit rate에서 기존의 방법보다 0.02~0.07보다 높다는 것을 알 수 있다.

3. 주관적인 화질 평가

주관적인 화질 평가를 위하여 ITU-R BT.500-11에서 권고하는 MOS방법, 즉, stimulus non-categorical judgement 방법을 사용하였다^[22]. 이를 위하여 20명의 참가자를 대상으로 10개의 비디오를 제안방법과 기존의 방법의 비디오 화질을 테스트하였다. 평가는 low, medium, high bit rate로 디코딩된 비디오 화질로 수행되었다. 모든 비디오는 임의적으로 화면에 양옆으로 보이게 되고, 참가자들은 더 좋은 화질의 비디오를 선택하며, 그 비디오에 화질 점수를 준다. 화질 점수의 범위는 [0,4]이다. 화질 점수 1,2,3,4는 각각

“약간 좋음”, “좋음”, “더 좋음”, “훌륭함”이며, 0은 “차이 없음”을 뜻한다.

평가된 MOS를 분석하기 위하여 본 논문에서는 Difference Mean Opinion Score (DMOS)를 사용하였다. PQ_{ij} 와 CQ_{ij} 를 제안 방법과 기존 방법의 화질 점수라 하고, $i \in \{1, \dots, P\}$ 와 $j \in \{1, \dots, M\}$ 은 참가자 i 와 비디오 j 를 나타낸다. 본 실험에서는 $P = 20$, $M = 10$ 으로 하였다. 그 다음, 비디오 j 에 대하여 참가자 i 의 차이 점수를 $d_{ij} = PQ_{ij} - CQ_{ij}$ 로 계산하였다. 정규화된 차이 점수 z_{ij} 는 다음과 같다.

$$z_{ij} = \frac{d_{ij} - \mu_i}{\sigma_i} \quad (11)$$

여기서 μ_{ij} 와 σ_{ij} 는 테스트 비디오의 참가자 i 평균과 표준 편차이다.

표 2. 테스트 비디오에 대한 CPBD 점수
Table 2. The CPBD scores for the test video sequences

Sequences (CIF size)	Method	Bit rates(bps)			
		300K	500K	700K	1M
Paris	Uncompressed	0.68	0.69	0.69	0.70
	Conventional	0.70	0.72	0.73	0.74
	Proposed	0.74	0.76	0.77	0.78
Silent	Uncompressed	0.48	0.52	0.55	0.57
	Conventional	0.52	0.57	0.60	0.63
	Proposed	0.58	0.63	0.66	0.69
Ice	Uncompressed	0.59	0.63	0.66	0.67
	Conventional	0.60	0.65	0.67	0.69
	Proposed	0.67	0.72	0.74	0.75
Sequences (HD size)	Method	Bit rates(bps)			
		700K	1.2M	1.6M	2M
Pedestrian	Uncompressed	0.30	0.31	0.31	0.32
	Conventional	0.31	0.33	0.34	0.35
	Proposed	0.34	0.35	0.36	0.37
Night	Uncompressed	0.43	0.44	0.45	0.47
	Conventional	0.46	0.47	0.49	0.50
	Proposed	0.48	0.49	0.51	0.53

따라서 비디오 j 의 DMOS 는 다음과 같이 계산되며,

$$DMOS_j = \frac{1}{P} \sum_{i=1}^P (z_{ij} + \mu_{ij}) \quad (12)$$

스케일된 DMOS(SDMOS)는 다음과 같다.

$$SDMOS_j = \frac{DMOS_j}{Score\ range(=4)} \times 100(\%) \quad (13)$$

<표 3>은 DMOS와 SDMOS 점수를 보여준다. 표에서 보듯이, DMOS 점수는 모든 비디오에 대하여 0보다 크다. 이것은 기존의 방법에 비하여 제안 방법의 화질이 더 우수하다는 것을 의미한다. 또한 SDMOS로부터 화질의 향상정도가 12%~44%가 된다는 것을 알 수 있으며, 90% 신뢰도로 DMOS의 오류 비율은 전부 15% 이내라는 것을 알 수 있다.

<그림 8>과 <그림 9>는 제안 방법의 비디오 화질과 기존

방법의 비디오 화질을 보여준다. 두 그림에서 보듯이 배경의 smoothing 정도는 인간이 거의 인지 할 수 없기 때문에 영상의 전반적인 화질은 같다고 볼 수 있다. <그림 8>에서 제안 방법은 보행자 지역에서 saliency가 나타남을 볼 수 있으며, 빠르게 움직이는 택시를 fast 움직임 지역으로 분류함으로써 saliency가 나타나지 않음을 알 수 있다. 지역적인 비디오 화질 비교에서 제안 방법에 의한 영상이 보행자 지역에서 더 선명함을 알 수 있는 반면, smoothing 정도가 강한 택시 지역에서는 기존의 방법과 거의 차이를 알 수 없음을 알 수 있다. <그림 9>에서는 제안 방법이 계속해서 왼쪽 사람의 얼굴을 선명하게 하는 것을 볼 수 있으며, 움직임이 없을 때 smoothing을 수행하는 기존 방법의 단점을 보완한다. 따라서 기존 방법으로 비디오 처리했을 시에 나타나는 깜빡임(flickering) 현상을 제거하는 장점이 있음을 알 수 있다. 그러므로 제안하는 방법은 인간 시각 인지에 맞게 더 효율적으로 주어진 대역폭을 사용한다고 판단할 수 있다.

표 3. DMOS 점수와 그에 상응하는 신뢰구간
Table 3. The DMOS scores and corresponding confidence intervals

Sequences	Bit rates(bps)	DMOS (SDMOS)	DMOS confidence interval of 90%
Paris	300K	1.55 (39%)	[1.33, 1.77]
	500K	1.74 (44%)	[1.49, 1.99]
	1000K	1.76 (44%)	[1.51, 2.01]
Silent	300K	1.57 (39%)	[1.35, 1.79]
	500K	1.64 (41%)	[1.42, 1.86]
	1000K	1.68 (42%)	[1.45, 1.91]
Ice	300K	0.98 (25%)	[0.84, 1.12]
	500K	1.15 (29%)	[1.01, 1.31]
	1000K	1.17 (29%)	[1.01, 1.33]
Pedestrian	700K	0.48 (12%)	[0.41, 0.55]
	1200K	0.82 (21%)	[0.71, 0.93]
	2000K	0.91 (23%)	[0.78, 1.04]
Night	700K	0.78 (20%)	[0.67, 0.89]
	1200K	0.89 (22%)	[0.76, 1.02]
	2000K	0.90 (23%)	[0.77, 1.03]

그림 8. 디코딩된 'Night' 영상의 화질 비교, (a) 기존의 방법으로 디코딩된 영상 (b) 제안 방법으로 디코딩된 영상
 Fig. 8. Comparison of the decoded 'Night' sequence processed by the proposed method and the existing method, (a) Decoded sequences processed by the existing method (b) Decoded sequences processed by the proposed method

그림 9. 디코딩된 'Paris' 영상의 화질 비교, (a) 기존의 방법으로 디코딩된 영상 (b) 제안 방법으로 디코딩된 영상
 Fig. 9. Comparison of the decoded 'Paris' sequence processed by the proposed method and the existing method, (a) Decoded sequences processed by the existing method (b) Decoded sequences processed by the proposed method

V. 결론

본 논문에서는 인간 인지 기반 비디오 코딩을 위한 비디오 처리 방법을 제안하였다. 제안 방법은 인간 시각 인지의 효과들을 반영하는 방법으로써 움직임 패턴을 개발하였고, 개발된 움직임 패턴과 기존의 saliency 모델을 결합하여 새로운 인간 시각 인지 모델을 제안하였다. 또한, 인간 시각 인지 모델을 구현하기 위하여 개선된 foveation filter를 개발하였다. 개발된 foveation filter는 인간 시각 인지 모델에 따라 비디오의 신호를 smoothing 또는 enhancing하는 방법이다.

객관적인 화질 평가와 주관적인 화질 평가로 기존의 방법보다 제안 방법의 성능이 우수하다는 것을 검증하였으며, 주관적인 화질 평가에서 기존의 방법보다 12%~44%의 화질 향상이 있었음을 알 수 있었다. 그러므로 제안하는 비디오 처리 방법은 인간 인지에 더 알맞도록 주어진 대역폭을 사용하는 인간 인지 기반 비디오 코딩 방법을 제공한다.

참 고 문 헌

- [1] Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment.
- [2] K. Minoo and T. Q. Nguyen, "Perceptual video coding with H.264", IEEE Asilomar Conf. Signals, Syst. Comput., pp. 741-745, Nov. 2005.
- [3] N. Jacobson and T. Q. Nguyen, "Video processing with scale-aware saliency: application to frame rate up-conversion", Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), pp. 1313-1316, May 2011.
- [4] K. Shin, M. Stolte and S. C. Chong, "The effect of spatial attention on invisible stimuli", Attention, Perception, & Psychophysics, vol. 71, No. 7, pp. 1507-1513, 2009.
- [5] Y. Yeshurun, and M. Carrasco, "Attention improves or impairs visual performance by enhancing spatial resolution", Nature, Vol. 396, pp. 72-75, Nov. 1998.
- [6] Z. Wang, and L. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection", IEEE Trans. Image Processing, Vol. 12, No. 2, pp. 243-254, Feb. 2003.
- [7] D. Walther and C. Koch, "Modelling attention to salient proto-objects", Neural Networks, Vol. 19, pp. 1395-1407, 2006.
- [8] N. Jacobson and Y. Lee and V. Mahadevan and N. Vasconcelos and T. Q. Nguyen, "A novel approach to FRUC using discriminant saliency and frame segmentation", IEEE Trans. Image Processing, Vol. 19, No. 11, pp. 2924-2934, Nov. 2010.
- [9] J. Harel, and C. Koch, and P. Perona, "Graph-based visual saliency", Advances in Neural Information Processing Systems, Vol. 19, pp. 545-552, 2007.
- [10] Y. F. Ma and H. J. Zhang, "A model of motion attention for video skimming", ICIP, pp.129-132, Sep. 2002.
- [11] E. Shechtman and Y. Caspi and M. Irani, "Space-time super-resolution", IEEE Trans. Pattern Anal. and Mach. Intell. Vol. 27, No. 4, pp. 531-545, Apr. 2005.
- [12] D. Melcher, and M. C. Morrone, "Spatiotopic temporal integration of visual motion across saccadic eye movements", Nature Neuroscience, Vol. 6, No. 8, pp. 877-881, Aug. 2003.
- [13] 김원하, 오형석, 임태하, "Video Processing based on Motion Analysis", IPIU2012, 2월, 2012
- [14] N. Jacobson, and Y. Freund, and T. Q. Nguyen, "Occlusion boundary detection using an online learning framework", Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), pp. 913-916, May 2011.
- [15] John Watkinson, The Mpeg Handbook, Focal Press, 2004.
- [16] H. D. Cheng and J. Chen and J. Li, "Threshold selection based on fuzzy c-partition entropy approach", Pattern Recognition, Vol. 31, No. 7, pp. 857-870, 1998.
- [17] R. Gonzalez and R. E. Woods, Digital Image Processing, 2nd edition, Prentice Hall, 2002.
- [18] D. Martin and D. Tal and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics", ICCV, Vol. 2, pp. 416-423, 2001.
- [19] S. Aghagholzadeh, and O. K. Ersoy, "Transform Image Enhancement", Opt. Eng., Vol. 31, pp. 614-626, Mar. 1992.
- [20] H.264 Reference Software Version JM17.2[Online], Available: <http://iphome.hhi.de/suehring/tml>, 2010.
- [21] N. D. Narvekar and L. J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)", IEEE Trans. Image Processing, vol. PP, No. 99, pp. 1-7, 2011.
- [22] Methodology for the Subjective Assessment of the Quality of Television Pictures 2002, ITU-R Recommendation BT.500-11, 2002.

저 자 소 개

오 형 석

- 2005년 2월 : 경희대학교 전자공학과 공학사
- 2007년 2월 : 경희대학교 일반대학원 전자전파공학과 석사
- 2009년 8월 ~ 2010년 8월 : University of California San Diego (UCSD) 방문 연구원
- 2012년 2월 : 경희대학교 일반대학원 전자전파공학과 박사
- 2012년 3월 ~ 현재 : 경희대학교 학술연구교수
- 주관심분야 : 영상/동영상 신호처리, 컴퓨터 비전, 동영상 부호화

김 원 하

- 1985년 2월 : 연세대학교 전자공학과 공학사
- 1988년 5월 : University of Wisconsin-Madison 전기공학과 석사
- 1996년 1월 ~ 7월 : (미) Motorola 연구원
- 1997년 5월 : University of Wisconsin-Madison 전기공학과 박사
- 1997년 8월~2000년 2월 : (미) Los Alamos National Lab. 연구원
- 2000년 3월~2003년8월 : 명지대학교 정보통신공학과 조교수
- 2009년 8월 ~ 2010년 8월 : University of California San Diego (UCSD) 교환 교수
- 2003년 9월 ~ 현재 : 경희대학교 전자정보대학 전자전파공학과 교수
- 주관심분야 : 영상/동영상 신호처리, 컴퓨터 비전, 동영상 부호화