화상 통화시 화자의 얼굴화질을 강화하는 동영상 부호화 기법

이승철 °, 낭종호

°삼성전자 통신연구소, 서강대학교 컴퓨터학과

°schlee@samsung.com, jhnang@ccs.sogang.ac.kr

A Video Encoding Mechanism Improving the Quality of Speaker Face Region on Video Telephony

Seung-Cheol Lee°, Jong-Ho Nang

°Telecommunication R&D Center, Samsung Electronics, Department of Computer Science, Sogang Univ.

요 약

본 논문에서는 화상 통화를 위한 비디오 인코딩에서 화자의 얼굴 화질을 강화하여 인코딩 할 수 있는 동영상 인코딩 방법을 제안한다. 제안한 인코딩 방법에서는 이미지의 Cr 데이터 및 움직임벡터 정보를 이용하여 빠르게 화자 얼굴 영역을 검출하고, 이 영역에 대하여 선택적인 양자화를 통하여 상대적으로 많은 비트량을 할당하여 화자의 얼굴 화질을 상대적으로 강화한다. 이 방법을 H.263 인코더에 적용하는 경우 전체적으로 이런 방법을 적용하지 않았을 때와 비교하여 18% 정도의 추가적인 CPU 오버헤드가 필요하였지만, 얼굴 영역에 대하여서는 PSNR 3dB 정도의 화질이 개선될 수 있음을 실험을 통하여 증명하였다.

1. 서 론

무선 통신 기술의 발전으로 휴대전화가 3 세대로 진화함에 따라 전송 대역폭이 크게 증가하여 초기 CDMA 기술에서의 14.4kbps 에 불과했던 데이터 전송속도가 현재 CDMA-EVDO 및 WCDMA 기술을 이용하여 2Mbps 에 이르게 되었다[1, 2, 3]. 또한 이동통신용 프로세서 및 하드웨어의 성능이 향상되어 휴대전화에 100MIPS 를 상회하는 성능의 RISC chip 을 장착하여 H.263[4] codec 의 구현이 가능함에 따라, 기존의 음성통화 중심의 서비스 한계를 뛰어넘어 화상통신을 기반으로 하는 다양한 멀티미디어 서비스가 상용화 되었다. 화상통화의 경우 H.323[5] 및 H.324M[6] 형식을 통하여 RTP/RTCP 가 UDP 계층 위에서 동작하면서 채널당 64kbps 의 대역폭을 가지고 양방향으로 통신을 하고 있으며, 이중 H.263 인코더를 이용하여 인코딩된 데이터는 약 48kbps 를 점유하게 되고 나머지는 음성데이터가 사용한다[3]. 이러한 대역폭은 휴대전화의 이동성 및 사용자수 증가에 따른 QoS 를 고려하여 현실화 한 것으로 기존 통신기술에 비하여 나아진 환경임에는 분명하나 비디오 인코딩에 충분하지는 못하다.

일반적으로 협대역 환경에서 비트율 제어를 위하여 양자화값을 크게 적용하게 되는 데 이때, 블록화 현상과 같은 부작용으로 화질이 떨어지며, 특히 움직임이 많거나 장면이 변하게 되면 이 현상이 더욱 심화된다. 이러한 부작용을 피하고 화질을 유지하기 위해서는 프레임 생성 속도를 떨어뜨려야 하나, 이는 동영상의 움직임이 부자연스러워 지기 때문에 이를 이용한 화질 조절은 한계가 있다.

본 논문에서는 영상 내의 배경과 화자를 구분하여 이를 차별화한 양자화를 적용하여 중요하다고 생각되는 화자의 얼굴부분에 충분한 비트량을 할당하여 화질 열화를 방지하고, 이보다 비교적 적은 비트량으로 주변 배경의 영상을 인코딩 하도록 제안하였다. 인코딩 과정에서 얻어지는 정보를 적절히 이용하여 계산량을 크게 줄였으며, 휴대전화라는 계산능력이 떨어지는 하드웨어 기반에서 구현이 가능하였다.

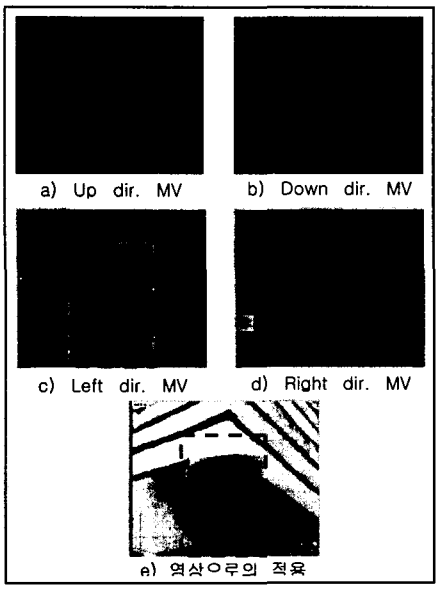
2. 제안하는 적응적 H.263 인코더

본 논문에서는 화상통신에서 화자의 영역을 검출하고 이 화자의 얼굴과 같이 의복 밖으로 드러난 신체의 화질 저하를 방지하기 위한 방법을 제안하고자 한다. 화상 통화에서는 화자의 대부분을 얼굴영역이 차지하게 되며 이를 위해 화자의 영역, 특히 얼굴 부위를 검출하고 이 검출된 영역에 대하여 그렇지 않은 영역과 차별적으로 양자화를 시도한다.

2.1 화자 영역 검출

움직이는 영상에서 개체는 배경에 비해서 일정한 방향과 크기로 움직이는 경향이 있다. 또한 배경을 비롯한 영상 전체는 한 프레임과 시간적으로 경과한 다음 프레임 간에는 하나의 방향성과 크기를 가지게 되므로, 하나의 움직임 벡터를 가질 수 있다. 영상 내 개체들은 이 영상 전체에 대한 벡터를 제외한 만큼의 움직임을 가지고 있고 이를 정량적으로 얻을 수 있다. 이 각각의 움직임 벡터들에 대하여 일정한 방향성과 크기에 대해서 기

하학적 분포를 얻을 수 있고, 이 매크로블록들의 연속적이고 서로 인접하는 영역을 구할 수가 있다. 이때 움직임 벡터는 크기 뿐만 아니라 방향성에 있어서 매우 많은 다양성을 갖게 되는 데, 이를 모두 계산하는 과정은 실시간성이 요구되는 화상통신에 적합하지 못하다. <그림 1>에서 a), b), c), d)는 각각 상, 하, 좌, 우 네 가지 방향에 대한 벡터들의 분포를 나타낸다. 이 분포에서 색이 밝을수록 큰 값이며 일정한 범위를 갖도록 제한한 것이다. 이때 연속적인 분포를 가지면서 가장 넓은 영역을 차지하는 것인 c)를 구한다. 이를 분석하기 전의 입력영상에 대응시킨 것이 e)이다. 그러나, 얼굴 이외의 다른 움직이는 큰 개체가 있다면 이 역시 얼굴 영역으로 판단될 수 있기 때문에 추가적인 색차 정보를 이용한 추가적인 처리가 요구된다.

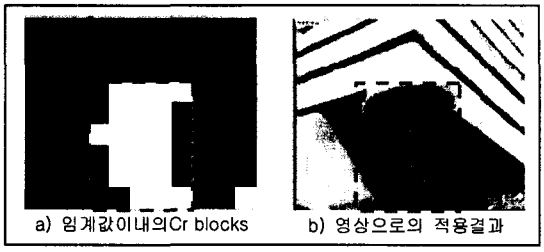


<그림 1> 움직임벡터에 기초한 영역

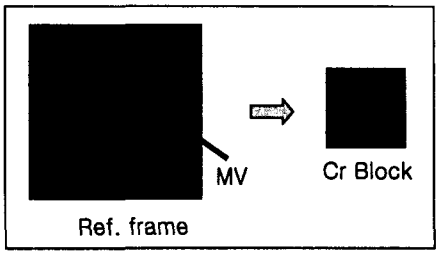
<그림 2>는 DCT Cr(chrominance red)블록의 DC 값을 분석하여 얼굴영역을 추정해 내고 있는 것을 보여준다. 각 블록별 DC 값을 얻어서, 특정 범위 이내의 값에 해당하는 블록들의 공간적인 연관성을 구하여 그 분포를 a)로 나타내었다. 이때 임계범위를 DC 값의 절대값에 대하여 1100 에서 1300 사이로 설정하였으며 이에 대한 실제 영상의 적용은 b)로 나타내었다.

화상통신시 화자의 얼굴영상은 넓은 부분을 차지하기 때문에 상기 두 가지, 즉 움직임벡터와 색차정보를 이용하는 방법에서 요구하는 조건을 만족하는 매크로블록들의 영역(16x16)으로도 충분히 얻을 수 있다.

한편, DCT 블록의 DC 값은 매 프레임마다 이전 프레임 값을 참조하여 갱신해 주어야 다음 프레임에서 이 값을 이용할 수 있다. 그리고 Cr DCT 블록은 그 크기가 16x16 영역에 영향을 주지만, 움직임 벡터는 0.5 단위까지 지정할 수 있는 데, 이 벡터의 크기가 16 의 배수가 아니라면(대부분의 경우에 해당됨), 이는 <그림 3>과 같이 움직임 벡터가 가리키는 영역은 4 개의 서로 다른 DCT 블록의 DC 값들로부터 영향을 받게 된다. 이에 영향을 주는 참조프레임의 각 영역의 면적에 대한 각 DC 값들의 가중치 평균값을 구하여 대략적인 DC 값을 추정할 수 있다.



<그림 2> Cr block 의 DC 값에 기초한 영역



<그림 3> DCT-domain 에서의 영상참조

2.2 영역 선택적 양자화

움직임벡터와 Cr 블록의 DC 값을 기초로 화자의 얼굴영역을 찾게 되면 이 영역에 대한 전체 비트량에 대한 할당 가중치를 높여주고 나머지 영역에 대한 가중치는 줄여준다. 전체 프레임을 기준으로 할 경우 가중치의 합은 1 이 된다. 본 논문에서 제안하는 방법은 움직임 벡터 정보를 이용하므로 처음에 오는 인트라 형식의 프레임은 일반적인 방법으로 인코딩하게 된다. 이때, 사용된 양자화값을 이용하여 양자화맵을 구성하고 그 다음 프레임부터 양자화맵에 기초하여 얼굴영역과 그 이외의 영역에 대한 가중치를 부여한 양자화를 실행하게 된다. 따라서 양자화가 상대적으로 크게 적용되는 얼굴 이외의 영역은 화질이 다소 떨어지고 블록화 현상이 발생할 수 있지만 얼굴영역은 그 만큼의 화질강화 효과를 얻을 수 있다.

한편 이와 같은 방법으로 양자화를 할 경우 일정 부분을 인코딩한 후 비트량을 체크하는 기준이 GOB 또는 프레임이 된다. 매크로블록마다 비트율 제어를 하는 것이 바람직 하나 얼굴영역/나머지영역에 대한 고려를 매크로블록 단위로 계산할 경우 많은 오버헤드가 수반될 수 있기 때문이다. 또한 얼굴영역의 비중이 일정비율을 유지하도록 가중치를 두는 것이 나머지영역의 화질을 지나치게 떨어뜨리지 않는 요인이 된다.

통신용 비디오 코덱으로 사용되는 H.263 의 경우 매크로 블록 1 개당 움직임벡터와 Cr DCT 블록의 DC 값이 1 개이므로 영역의 경계는 매크로블록 단위가 되고 비트율 제어의 단위 역시 마찬가지 이므로 제안하는 양자화

방법을 구현하기 위한 추가적인 연산은 필요하지 않게 된다.

3. 실험을 통한 성능비교

본 논문에서 제시하는 방법이 실제 컴퓨터에서 구현되어 그 결과가 타당한 지를 검증하였다. 여기에는 사람이 분리해낸 얼굴 영역에 대하여 기존의 일반적인 H.263 인코더와 제안된 방법에 의해 구현된 인코더 사이의 화질의 차이를 알아보고, 부가적인 오버헤드에 의한 실행속도의 차이를 비교해 본다. 또한 실제로 비트율 제어가 잘 되어 비트량의 차이가 없는 지에 대한 확인도 하였다.

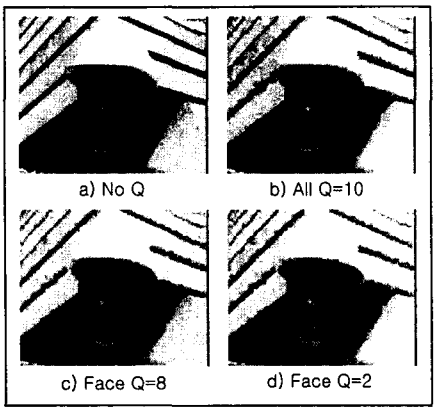
먼저 화질의 차이는 실험의 편의상 PC 상에서 작업을 하여 결과를 도출하였으며, 실행 성능의 검증은 현재 상용화 진행 중인 삼성전자 무선단말기를 이용하였으며 이의 사양은 CPU ARM920t 135Mhz, System 33.6Mhz 의 제원을 가지고 있다[7].

<표 1> 실제 얼굴영역에 대한 기존 방법과의 PSNR 을 통한 화질비교 (단위:dB)

양자화	Y-frame	Cr-frame	Cb-frame
No Quant	42.31	42.13	42.21
All Qp=10	36.5	35.8	36.2
Face Qp=8	37.34	36.61	36.7
Face Qp=2	39.76	39.14	39.19

<표 1>은 화상통신 환경과 비슷한 샘플 영상인 foreman raw image 들을 기존 및 제안한 방법으로 인코딩 후 다시 디코딩 한 뒤, 10 frame 을 무작위로 선택하여 얼굴영역 화질의 평균값을 구하여 비교한 것이다. 그 결과를 <그림 4>에 보였으며, a)는 양자화 과정을 거치지 않도록 실험하여 PSNR(peak signal to noise ratio) 42dB 정도를, b)는 중간 정도의 화질이라 할 수 있는 Qp=10 을 영상 전체에 적용하여 양자화를 하였을 때에 36.5dB 를 얻었다. 여기에 대하여 제안한 방법을 사용하여 얼굴영역의 Qp=8, Qp=2 인 경우에 대하여 각각 대략 37, 39dB 정도를 얻었고, 이를 c), d)에 시각적으로 나타내었다. 이는 a)와 같이 전혀 양자화를 하지 않아서 통신용으로는 사용할 수 없는 영상의 화질에는 미치지 못하지만 여기에 꽤 근접했음을 보여준다. 보통 수준의 통신상태에서 전송이 가능한 Qp=10 인 상태인 b)에 비하여 3dB 정도의 얼굴영역의 화질에 있어서 이득을 보았으며, 비트율 제어에 의해 배경의 화질이 약간 떨어지는 단점이 있다. 그러나, <그림 4>의 b), c), d)의 배경은 시각적으로 크게 차이가 나지 않지만, 얼굴영역은 차이가 크게 느껴진다.

성능에 있어서 기존 방법에 의한 인코더의 평균 비디오프레임 생성 속도는 QCIF 기준 22frame/s 였으나, 본 논문에서 제안하는 방법을 적용시킬 경우, 약 18%의 속도가 감소되는 18frame/s 를 얻었다. 실제에 있어서 통신 대역폭을 감안하여 화질과 프레임 속도 사이의 적절한 기준은 대략 10frame/s 이내가 되기 때문에, 이 정도의 오버헤드는 무선단말기 상에서의 구현에 있어서 큰 영향을 주지 않는 다고 볼 수 있다.



<그림 4> 다양한 양자화 방법에 따른 실험결과

4. 결 론

본 논문에서는 화상통신을 위한 H.263 인코딩 과정에서 얻어지는 움직임벡터 및 색차 DCT 블록의 일부 정보를 이용하여, 얼굴로 추정되는 영역을 구하고 이 부분의 화질저하를 방지하는 방법을 제시하였다. 이는 기존에 알려진 복잡한 개체인식 및 얼굴인식 알고리즘 보다 간단한 수식으로서 제한된 계산자원을 활용하면서도 실시간 특성을 요구하는 무선통신 단말기의 제약을 극복하고 얼굴 영역을 찾을 수 있었다. 정확성에 있어서, H.263 스트림에서 얼굴영역의 화질이 기존 방법에 의한 스트림보다 나아진 것을 실험을 통해 확인하였으며, 비트량을 고정시키기 위해 주변 영역의 화질이 열화되는 점과 인코딩 속도의 감소의 단점이 확인되었다.

이로서 실제 무선화상통신을 하는 사용자의 체감 화질은 개선될 것으로 보인다. 계산과정에서 얻어지는 움직임벡터의 분포표 및 Cr DCT 블록의 DC 분포표에서 유효한 영역을 효과적으로 얻어내는 방법에 대해서는 앞으로의 과제가 될 것이며, 이 부분이 더욱 개선된다면 화자의 얼굴영역이 더욱 정확히 찾아 질 것이므로 나머지 영역의 화질 저하를 좀 더 방지할 수 있으리라 판단된다.

참고문헌

[1] Qualcomm Corp. http://www.qualcomm.com
[2] SK Telecom. http://www.sktelecom.com
[3] KTF. http://www.ktf.com
[4] ITU-T Recommendation H.263, Video coding for low bit rate communication
[5] ITU-T Recommendation H.323, Visual telephone systems and equipment for local area networks which provide a non-guaranteed quality of service
[6] ITU-T Recommendation H.324, Terminal for low bit-rate multimedia communication
[7] ARM Ltd. http://arm.com