동영상 콘텐츠 생성을 위한 음악과 사진 분석

정 명 범[†] · 고 일 주[††]

요 약

데이터 전송 기술의 발달과 디지털 기기의 다양한 보급으로 소비자들은 디지털 콘텐츠를 생산하는 주체로 변화하였다. 생성되는 다양한 콘텐츠 중 사용자는 동영상 제작에 큰 관심을 보였으며, 그 중 음악과 사진으로 동영상을 제작하는 방법은 사용자에게 보다 손쉽게 동영상을 만들 수 있도록 제공 되었다. 그러나 현재의 방법은 사진과 사진 간의 연관성이 결여되었을 뿐 아니라 음악의 리듬과 관계없이 일정 시간 간격에 따라 사진이 변화한다. 본 논문에서는 음악 분석을 통하여 음악 리듬에 따라 사진이 변화하고, 사진 분석을 통하여 사진 간의 연관성을 나타낼 수 있는 동영상 제작 방법을 제안한다. 음악 분석은 RMS를 이용하여 리듬이 강한 부분을 찾았으며, 사진 분석은 구조 단순도와 얼굴 영역 추출을 이용하여 인물 사진과 풍경사진으로 분류하였다. 사진 분석은 86.4%의 성공률을 보였으며, 이를 이용하여 음악 리듬에 맞은 사진 변화 위치와 사진 간의 연관성을 가진 순서 배치를 할 수 있었다. 따라서 음악 분석과 사진 분석을 이용한 자연스럽고 효과적인 동영상을 제작 할 수 있다.

키워드 : 콘텐츠 제작 기술, 사진 분석, 음악 분석, 디지털 신호 처리

Analysis of Music and Photo for User Creative Movie

Myoung-Bum Chung[†] · Il-Ju Ko[††]

ABSTRACT

Consumers changed to the subject to produce a digital contents as data transmission technique is advanced and a digital machine is diffused variously. Users are interested greatly in a user creative movie (UCM) production among various online contents. The UCM production method which uses the music and picture is the method that users make the UCM more easily. However, the UCM production service has the problem that any association does not exist in the music and picture and that the picture changes according to fixed time interval without the relation at a music rhythm. To solve this problem, we propose the UCM production method which uses a music analysis and picture analysis in the paper. A music analysis finds a picture change time according to the rhythm and a picture analysis finds the association of the picture. A music analysis finds strong parts of the sound which uses Root-Mean-Square (RMS). And a picture analysis classifies the picture as a scenery picture and people picture which uses structure simplicity of the picture(SSP) and face region detection. A picture analysis got correct result of 86.4% in the experiment and we can finds the association at each picture and arranges the sequence which the picture appears. Therefore, if we use a music and picture analysis at the UCM production, users may make natural and efficient movie.

Key Words : Contents product method, Photo analysis, Music analysis, Digital Signal Processing

1. 서 론

최근 데이터 전송 기술의 발달로 인터넷 환경이 좋아지고, 전송 속도가 향상되었다. 이와 더불어 카메라가 달린 핸드폰, 디지털 카메라 등의 디지털 기기 보급은 온라인에서 소비자 역할에만 머물렀던 사용자들을 멀티미디어 콘텐츠 생산의 적극적인 주체로 변화 시켰다. 그 예로 사용자는 자신이 직접 찍은 사진을 온라인의 게시판이나 블로그 등에 올려 다른 사용자들과 함께 공유하며, 사진과 음악을 이용해 동영상을 만들어 평가를 받거나, 상업적인 목적에 이용하기도 한다. 이렇게 사용자들이 직접 제작하여 부가가치를 창출해 내는 콘텐츠를 일컬어 '사용자 제작 콘텐츠(User Created Contents : UCC)' 또는 '사용자 창작 콘텐츠(User Generated Contents)'라고 한다[1].

UCC는 콘텐츠 매체에 따라 다음과 같이 다섯 가지로 분류된다. 복합 미디어로 구성된 User Packaged Contents (UPC)를 포함하여 텍스트, 이미지, 오디오, 동영상 등이다[2]. 그 중 동영상은 기존의 텍스트나 이미지로 표현할 수 없었던 사용자의 욕구를 채워줄 수 있으며, 그에 따라 동영상에 대한 사람들의 관심은 점차 증대되었다[3, 4]. YouTube(http://www.

※ 본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음
† 종신회원 : 숭실대학교 대학원 미디어학과 박사과정
†† 종신회원 : 숭실대학교 미디어학부 조교수
논문접수 : 2007년 3월 14일, 심사완료 : 2007년 5월 10일

youtube.com)와 같은 포털 서비스에서는 비전문가도 쉽게 동영상을 제작할 수 있게 UCC 툴을 제공하며, 사람들이 동영상을 검색 할 수 있게 동영상 검색 서비스를 제공하고 있다 [5].

이러한 동영상 콘텐츠를 제작하는 방법은 다음과 같다. 첫 번째는 웹캠이나, 디지털 캠코더를 이용하여 동영상을 제작하는 것으로 직접 영상을 찍어 편집하여 올리는 것이다. 이것은 실제 움직이는 영상을 보여주기 때문에 사실감을 줄 수 있지만 '영상 편집'이라는 기술적인 문제 때문에 비전문가가 사용하기에는 다소 어려움이 있어 실제로는 편집 없이 그대로 보여주는 것이 대부분이다. 두 번째는 온라인이나, 개인 컴퓨터에 있는 사진들을 사용자가 선택하고, 음악에 적절히 배열하여 합성하는 것으로 비전문가도 쉽게 동영상을 제작할 수 있다. 이것은 온라인 디지털 액자나, 모바일 폰에 전송하기 위한 동영상 제작에 많이 사용되고 있다. 그러나 이 방법은 사용자가 사진을 하나하나 나타날 위치에 지정해야 하며, 지정한 배열에 따라 사진들을 보여줄 뿐 사진 간의 연관성이 결여 될 가능성이 크다. 게다가 사진들의 배열 또한 음악의 상태에 관계없이 시간 간격에 따라서만 사진을 보여주는 단점을 가지고 있다.

본 논문에서는 앞에서 언급한 두 번째 방법의 단점을 보완하는 방법으로 음악과 사진 분석을 이용한 동영상 콘텐츠를 생성하는 방법을 제안 한다. 먼저 음악 분석은 음악의 리듬이 강하게 들리는 부분을 찾아 그 곳에 사진들이 나타나게 하기 위한 것이다. 음악은 리듬을 가지고 있으며 리듬에 의해 강한 소리(Max Sound : MS)와 약한 소리(Min Sound : NS)를 번갈아가며 나타낸다[6][7]. 일반적으로 사람들은 MS 부분에서 어떠한 동작이 일어나거나, 움직임이 있기를 기대한다. 한 예로 캐릭터 애니메이션을 보면 캐릭터의 움직임을 음악의 MS 부분에서 행동이 변화하게 함을 볼 수 있다 [8]. 본 논문에서는 음악의 MS 부분을 찾아내기 위해 디지털 신호 처리의 Pulse code modulation (PCM) 데이터를 이용한 Root-Mean-Square (RMS) 값을 사용하였으며, RMS 값이 작은 값에서 일정 값 이상 커질 때를 MS 부분으로 추적하였다[9].

다음으로 사진 분석은 나타나는 사진 간의 연관성을 얻는다. 사진은 구조 단순도와 얼굴영역 추출을 사용하여 인물사진과 풍경 사진으로 분류된다. 구조 단순도는 본 논문에서 사진을 효과적으로 분류하기 위해 제안하는 알고리즘이다. 구조 단순도는 '사람들이 풍경사진을 찍을 때 장소, 구도, 색상 등을 생각하여 촬영한다.' 는 것을 착안한 방법으로서 사진 구도상의 특징점을 찾아내 수치화 한다. 일반적으로 구조 단순도가 낮은 값 일 때 인물 사진, 높은 값 일 때 풍경 사진일 가능성이 크다. 얼굴 영역 추출은 얼굴 인식에서 일반적으로 사용하는 Haar 분류기를 이용한 CBCH로부터 사진에 사람이 있고, 없음을 판단한다[10] [11]. 따라서 구조 단순도와 얼굴 영역 추출의 비율적 조합에 의해 인물사진과 풍경 사진을 보다 효과적으로 분류할 수 있다. 결론적으로 우리는 앞에서 언급한 음악과 사진 분석의 결과로부터 음악 리듬에 맞추어 사진들이 나타나고, 사진 간의 연관성을 갖는 동영상을 자동으로 만들 수 있다.

논문의 구성은 다음과 같다. 2장에서는 음악 분석에 사용사용했던 RMS를 구하는 방법과 사진 분석의 얼굴 영역 검출에 사용했던 CBCH에 관한 관련 연구를 기술한다. 3장에서는 본 논문에서 사용한 음악 분석 방법을 설명하며, 4장에서는 제안하는 구조 단순도 알고리즘을 설명하고 사진 분류를 위한 얼굴 영역 검출과 구조 단순도의 조합 비율에 대해 언급을 한다. 5장에서는 음악 분석과 사진 분석의 실험 및 결과를 통해 제안한 방법의 유용성을 보이고, 6장에서는 결론을 제시한다.

2. 관련 연구

음악 분석은 PCM 데이터로부터 얻은 RMS를 사용한다. 표준 CD 음질은 1초에 44100Hz로 소리를 샘플링하게 되므로 Shannon의 샘플링 이론에 따라 디지털로 분석 가능한 진동수는 $f_N = f_s/2$에 해당되는 22050 Hz가 된다[12]. 그리고 대부분의 논문에서는 Peak와 RMS를 구하기 위한 방법으로 0.05초 (20ms)라는 시간으로 구간을 나누었다. 이는 실시간으로 화면을 갱신하며 분석하기 위해서이다[13]. 따라서 PCM 데이터를 이용한 RMS의 계산 방법은 식 (1)과 같다.

$$S_t = \frac{(PCM_{(t*1764)})^2 + (PCM_{(t*1764)+1})^2 + \ldots + (PCM_{(t*1764)+1763})^2}{1764} \quad (1)$$

RMS_t는 t시간에 해당하는 RMS 값이며, 1764라는 숫자는 20ms 동안에 나타나는 PCM 데이터의 횟수이다. 따라서 RMS_t는 각각의 20ms에 해당하는 PCM 데이터들의 값들을 제곱한 평균으로부터 구할 수 있다.

사진 분석은 영상 처리 기술 중 얼굴 검출 기술을 이용한다. 얼굴 검출은 영상의 모든 영역을 비교하여 얼굴인지 아닌지를 판단하는 것으로 영상의 종류를 판별하는 분류기를 학습시켜 원하는 패턴인지 아닌지를 구별한다. 예를 들어 얼굴 검출에서는 분류기에 여러 얼굴 영상을 보여줘서 얼굴의 특징을 학습시킨 뒤, 임의의 영상이 얼굴인지 아닌지를 분류기가 판단하게 한다. 이러한 방법에는 신경망 회로(Neural Network), SVM(Support Vector Machine), PCA(Principle Component Analysis), LDA(Linear Discriminant Analysis)등 많은 알고리즘이 있으며 그 중 CBCH(Cascade of boosted classifiers working with haar-like features)를 이용한 객체 검출 알고리즘은 빠른 속도와 높은 정확도를 갖는다.

CBCH를 이용한 검출 방법은 Haar 분류기를 조합한 알고리즘이다. Haar 분류기의 연산이 단순하여 속도가 빠르지만 연산이 단순한 만큼 얼굴 검출률은 높지 않다. 그러나 이러한 Haar분류기를 100개 혹은 1000개 이상을 적절히 조합함으로써 분류기의 성능을 높일 수 있다는 것이 CBCH의 핵심이다. 얼굴 검출에 적절한 Haar 분류기의 조합을 찾기 위해서는 아다부트스(Adaboost) 알고리즘을 사용한다. 아다부스트는 얼굴 영상에서 가능한 모든 형태의 Haar 분류기에 대해 얼굴 판별 능력이 뛰어난 순서대로 Haar 분류기를 추

출해준다. 그리고 추출된 분류기들은 임의의 영상에 대해서 각각 얼굴 영역인지 아닌지를 판별하며 결과 값의 비율에 따라서 얼굴 영상인지 아닌지 판별하게 된다. 이때 1000개의 Haar 분류기를 사용한다고 했을 때도 한번에 1000개 모두를 비교하는 것이 아니라 처음엔 1개, 그 다음 10개, 25개, 50개 등 그 개수를 증가시키면서 차례로 비교하는 방법을 사용한다. 이렇게 단계를 나누어서 비교하는 것을 cascade라고 하며 얼굴 검출 속도를 가속화시킬 수 있다.

CBCH는 인텔(Intel)사가 제작한 Open CV의 다양한 영상처리 기술 중 얼굴 인식에 사용되고 있다. Open CV는 Open Source Computer Vision Labrary의 약자로 영상처리를 위한 저수준의 함수를 표준 Dynamic Link Library(DLL) 또는 Static Library형으로 제공한다. Open CV는 객체, 얼굴, 행동 인식, 독순(입술 읽기), 모션 추적 등의 응용 프로그램에서 사용되고 있다. Open CV에서는 얼굴 인식 기술인 CBCH를 구현하기 위해 주요 함수 2개로 얼굴 검출을 구현하였다. 우선 cvLoad( ) 함수를 이용하여 정면 얼굴 검출에 대한 CBCH 파일인 haarcascade_frontalface_default.xml을 읽어온다. 다음으로 CvHaarClassifierCascade *cascade와 cvHaarDetectObject( ) 함수를 이용하여 실제 얼굴 검출을 수행한 후, 2개의 함수를 이용하여 입력 영상으로부터 얼굴 검출을 할 수 있다.

3. RMS를 이용한 음악 분석

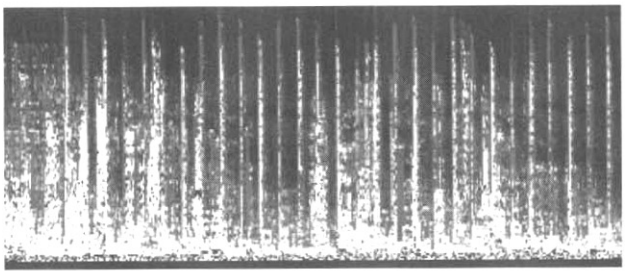
일반적으로 사람들은 강한 소리(MS) 부분에서 동작이 일어나거나 움직임이 있기를 기대하며, 그와 같은 예는 춤이나, 무용에서 잘 나타난다. 따라서 동영상 제작을 위한 음악 분석은 음악의 리듬이 강하게 들리는 부분을 찾아 그 부분에서 사진들이 자연스럽게 바뀌게 하기 위한 것이다.

본 논문에서는 MS 부분을 찾기 위해 PCM 데이터를 이용하여 RMS 값을 구한 후, 시간축에 따른 RMS의 변화량을 측정하였다. RMS의 값은 2 장에서 언급한 것으로 PCM 데이터들의 제곱의 합에 평균을 구한 것이다. 하나의 PCM 데이터는 최소 -32768에서 최대 32767까지의 값을 가지므로, 실제 RMS 값을 구하여 비교하기에는 범위가 다소 크다는 문제가 있다. 따라서 논문에서는 RMS의 데이터 값을 전체 RMS의 최대값에 의해 RMS의 범위를 0에서 100으로 정규화를 시킨 후 계산하였다.

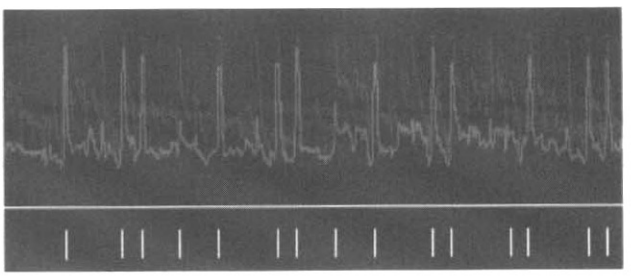
MS 부분은 음량의 변화가 크게 변화하는 곳을 말한다. 따라서 MS 부분은 RMS의 값이 일정 수치보다 낮은 값에서 순간 높은 값으로 변화하는 부분이라 할 수 있으며, 여기서는 낮은 값의 기준을 전체의 1/4에 해당하는 25 이하로, 높은 값의 기준을 전체의 1/2에 해당하는 50 이상으로 하여 측정하였다. 이때 사람의 귀는 작은 소리에서 큰소리로 변화하는 것에는 민감한 반면에 큰 소리에서 작은 소리로 변화하거나 큰 소리에서 큰 소리로의 변화에는 민감하지 않다. mp3의 손실 압축 방법 또한 이러한 사람의 귀의 특성을 이용한 것이다 [14] [15]. 본 논문에서도 사람의 귀의 특성을 고려하여 MS 부분 직후의 값은 측정하지 않게 하였다. (그림 1)은 음악 분석의 RMS 정보와 그때 나타나는 MS 부분을 측정한 예이다. (그림 1-a)는 실제 음악 파형을 나타내며, (그림 1-b)에 흰 선 위의 그림은 파형 정보로부터 얻은 RMS를 나타낸다. 그리고 (그림 1-b)에 아래의 흰색선은 RMS로부터 얻은 MS 부분을 나타낸다.

(그림 1-a), (그림 1-b)에서와 같이 RMS를 이용한 음악 분석을 통해 MS 부분을 측정할 수 있다. 그러나 이러한 MS 값은 음악의 종류에 따라 나타나는 개수가 다르다. Hip-hop, Rock, Dance 같은 곡은 드럼이나 타악기에 의해 강약이 일정 구간 내에서 두드러지게 나타난다. 따라서 MS의 개수 또한 많이 측정될 수 있다. (그림 2)는 Chris Brown의 Run it 이라는 Hip-hop 음악을 분석한 것이며, 아래의 흰색 선이 MS 부분을 측정한 것으로 MS 부분이 많이 나타남을 알 수 있다.

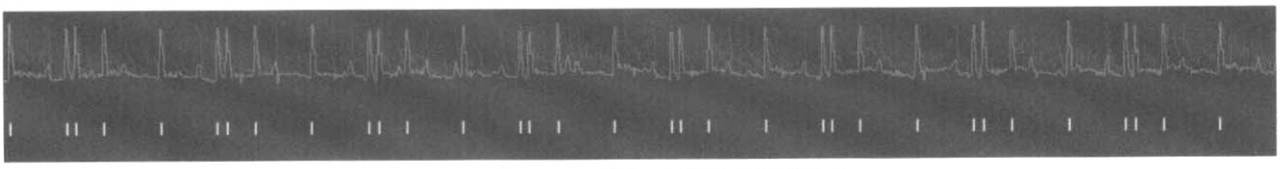
반대로 Ballad나 R&B와 같은 타악기의 소리가 크게 강조되지 않는 곡은 강약이 두드러지게 나타나지 않는다. 즉, MS의 개수가 일정 구간 내에서 적게 나타난다. (그림 3)은



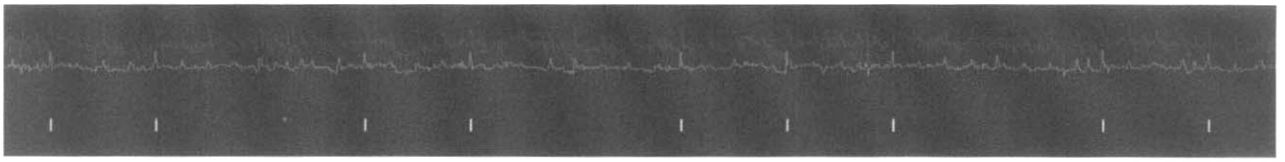
(그림 1-a) 실제 음악 파형



(그림 1-b) RMS 파형 값과 MS부분 측정 예



(그림 2) Hip-hop에서 MS 부분을 측정한 예



(그림 3) Ballad에서 MS 부분을 측정한 예

Elton John의 Something About The Way You Look Tonight 이라는 Ballad 음악을 분석한 것이다.

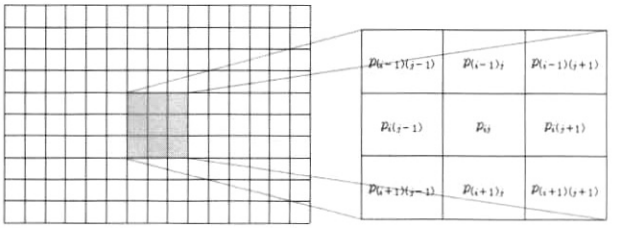
위와 같이 음악의 장르에 따라 강약의 변화에 대한 시간적 차이가 크다. 같은 구간 내에서 리듬이 강조되는 곡은 MS의 개수가 많이 나타나며, 그와 달리 리듬이 강조되지 않은 곡은 MS의 개수가 상대적으로 적게 나타난다. 따라서 MS의 개수에 따른 사진의 변화 위치를 지정하는 방법이 필요하다.

논문에서는 사진의 변화 위치를 지정하기 위해 노래 전체 구간을 사진의 개수만큼 나누어 각 위치에 먼저 배치하였다. 그 후 각 구간에 대해 MS의 개수가 많은 경우 그 구간에서 MS 값이 가장 큰 위치에 사진이 변화하게 한다. 반대로 구간에 대한 MS의 개수가 적거나 없는 경우 가장 가까운 위치의 MS가 나타나는 위치에서 사진이 변화하게 한다. MS 개수의 많고 적음의 기준은 현재 대부분의 곡에서 쓰이고 있는 4/4박자를 기준으로 하여 각 구간별 8 회 이상 나오는 경우를 많음으로 하였다. 이는 리듬의 기준이 되는 드럼이 일반적으로 한마디에 8회 나오는 것을 바탕으로 한 계산이다.

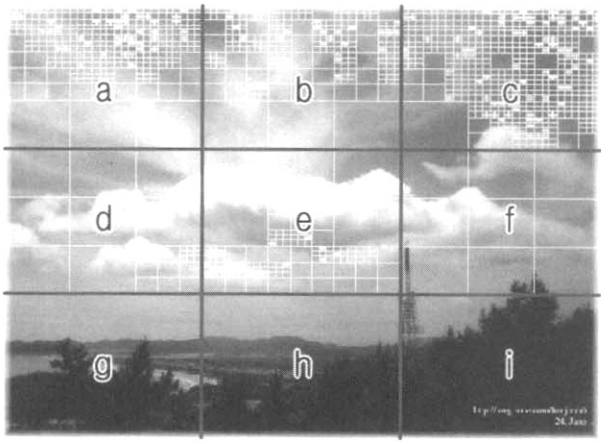
4. 구조 단순도와 얼굴영역추출을 이용한 사진 분석

사진 분석은 동영상에서 나타나는 사진의 순서에 대한 연관성을 주기 위한 것으로 입력 받은 사진의 내용을 분석하여 재배열한다. 사진 분석은 구조 단순도와 얼굴영역 추출을 사용하여 인물 사진과 풍경 사진으로 분류한다. 구조 단순도는 본 논문에서 사진을 효과적으로 분류하기 위해 제안하는 알고리즘으로 '사람들이 풍경사진을 찍을 때 장소, 구도, 색상 등을 고려한다.'는 것을 착안한 방법이다. 인물 사진은 장소나 구도에 신경을 쓰지 않고 인물을 중심으로 찍는 반면, 풍경 사진은 구도, 색상 등의 배치에 중심을 둔다. 특히 사진의 상·하에 대한 색상의 구도 차이가 많이 나타나는데 이러한 차이 값을 구조 단순도라 하며, 색상을 중심으로 9등분으로 나누어 그 값을 구할 수 있다. 따라서 풍경 사진에 가까울수록 상·하에 대한 색상 차이가 많으므로 구조 단순도 값이 높게 나타나며, 인물 사진에 가까울수록 색상 차이가 적으므로 구조 단순도 값이 낮게 나타난다. 구조 단순도는 입력 받은 사진을 9등분으로 나누고, 그 안에서 색상 복잡도에 의해 나누어진 횟수를 구하여 복잡도를 계산함으로 얻을 수 있다.

색상 복잡도를 구하는 방법은 다음과 같다. 사진을 9등분으로 나눈 각각의 영역을 테두리 1픽셀을 제외한 모든 픽셀에 대하여 식(2)와 같이 가운데 픽셀(p_{ij}) H(Hue) 값과 주변 픽셀들 H값과의 차이 값 평균을 구한다. (그림 4)는 색상 복잡도를 구하기 위한 픽셀을 나타낸 것이다.



(그림 4) 색상복잡도를 구하기 위한 픽셀간의 차이값 계산



(그림 5) 색상복잡도를 이용한 사진의 각 영역분할 횟수

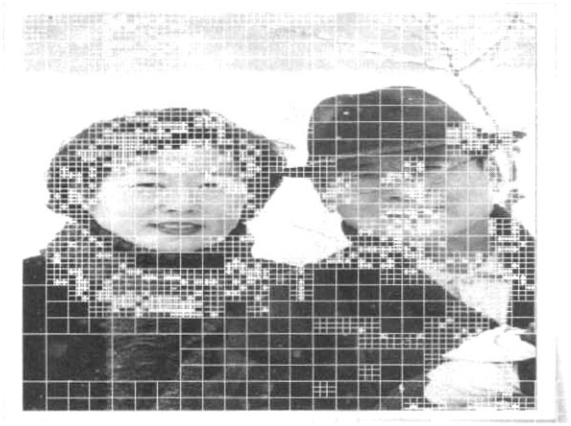
$$\sum_{i=2}^{m-1}\sum_{j=2}^{n-1}\frac{1}{8}\left(|p_{ij}-p_{(i-1)(j-1)}|+|p_{ij}-p_{(i-1)j}|+|p_{ij}-p_{(i-1)(j+1)}|+|p_{ij}-p_{i(j-1)}|+|p_{ij}-p_{i(j+1)}|+|p_{ij}-p_{(i+1)(j-1)}|+|p_{ij}-p_{(i+1)j}|+|p_{ij}-p_{(i+1)(j+1)}|\right) \quad (2)$$

이때 색상 복잡도는 모든 차이 값들의 합을 그 영역의 픽셀 값으로 나누어 구한다. 분할된 각 영역은 그 영역의 색상 복잡도를 계산하여 값이 일정 이상을 가지는 경우 다시 9등분으로 나누어 재귀적(Recursive)으로 색상 복잡도를 계산하며 분할한다. 색상 복잡도를 이용하여 (그림 5)와 같이 각 영역의 분할 횟수를 얻을 수 있다. (그림 5)의 a, b, c, d, e, f, g, h, i는 각 영역이 분할 된 횟수를 의미한다.

각각의 분할 횟수로부터 식(3)과 (4)를 이용하여 사진의 각 영역에 대한 복잡도를 계산한다.

$$Pc_a=\left|a-\frac{d+g}{2}\right|+\left|a-\frac{e+h}{2}\right|+\left|a-\frac{f+i}{2}\right| \quad (3)$$

식 (3)은 a 영역의 d, e, f, g, h, i에 대한 복잡도 값이다. 식 (3)을 이용하여 나머지 b, c에 대한 복잡도를 구해내며,



(그림 6) 인물 사진

(그림 7) 풍경 사진

g, h, i 영역에 대한 복잡도는 식 (4)와 같이 구한다.

$$Pc_g = \left|g - \frac{a+d}{2}\right| + \left|g - \frac{b+e}{2}\right| + \left|g - \frac{c+f}{2}\right| \quad (4)$$

각 영역별 복잡도인 Pc_a, Pc_b, Pc_c, Pc_g, Pc_h, Pc_i 로부터 모든 값을 더하여 평균을 구한 후 a~i 영역 중 분할 횟수가 가장 많은 값으로 다시 나누어 100을 곱한 값이 사진의 복잡도이며, 구조 단순도(Ps)는 수식 (5)와 같이 얻을 수 있다. $P_{\max}$는 9개의 영역 중 최대 분할 값이다.

$$Ps = 100 - \left\{\left(\frac{Pc_a + Pc_b + Pc_c + Pc_g + Pc_h + Pc_i}{6 \times P_{\max}}\right) \times 100\right\} \quad (5)$$

다음 (그림 6)과 (그림 7)은 구조 단순도를 구하기 위해 사진을 색상 복잡도에 따라 재귀적으로 나눈 것이다. (그림 6)은 인물 사진에 대한 색상 복잡도를 나타내며, 위, 아래의 색상 값이 모두 복잡하여 구조 단순도를 계산하면 낮게 나타남을 알 수 있다. (그림 7)은 풍경사진에 대한 색상 복잡도를 나타낸 것이다. 인물 사진과는 대조적으로 위와 아래에 대한 색상 값의 복잡도가 차이가 나며, 구조 단순도 값은 높게 나타난다.

사진 분석은 앞서 구한 구조 단순도와 CBCH를 이용한 얼굴 영역 추출의 비율적 조합으로 사진을 분류한다. 구조 단순도만을 이용하여 사진을 분류하는 경우 색상의 구성 비율을 가지고 분류하기 때문에 얼굴색이 없는 사진에서도 인물 사진이라는 오류가 나타날 수 있다. 반대로 얼굴 영역 추출만 이용하여 사진을 분류하는 경우 풍경 사진에서도 인물을 찾아내려는 오류가 발생한다. 즉, 구조 단순도와 CBCH의 비중을 조절하여 효과적인 사진 분류를 할 수 있다. 본 논문에서는 구조 단순도와 얼굴 영역 추출 값을 조합하여 이용하였다. 얼굴 영역이 추출되면서 구조 단순도가 낮은 값을 나타낼 때 인물 사진으로 분류하고, 얼굴 영역이 없으면서 구조 단순도가 높은 값을 나타낼 때 풍경 사진으로 분류한다. 조합 비율은 구조단순도 70%, 얼굴 영역 검출 30%로 하였을 때 가장 높은 정확성을 보였다.

5. 실험 및 결과

실험은 사진 분석을 통한 인물 사진과 풍경사진을 분류하는 실험과 사진 분석으로부터 재배치된 사진을 음악에 맞추어 동영상이 재작되는 실험으로 하였다.

구조 단순도와 얼굴 영역 검출을 이용한 사진 분류 실험은 다음과 같다. 자료 수집은 디지털 카메라 사진을 인터넷으로부터 무작위로 구하였으며, 인물 사진과 풍경 사진 250장씩 총 500장을 사용하였다. 구조 단순도의 유효성을 검증하기 위해 얼굴 영역 검출만을 이용한 사진 분류와 얼굴 영역 검출과 구조 단순도를 함께 사용한 사진 분류 두 가지로 실험 하였다. 실험에 사용한 얼굴 영역 추출 알고리즘은 다음과 같다.

얼굴 영역 검출은 로그-보색 검출과 Hue/Red 검출을 이용하였다. 로그-보색 검출 방법은 Fleck의 스킨 필터를 이용한 방법으로 식 (6), (7)을 통하여 R, G, B값을 로그 보색 칼라 표현 값인 I, R_g, B_y로 변환시킨다.

$$L(x) = 105 \times \log_{10}(x+1+n) \quad (6)$$

$$I = L(G),\ R_g = L(R) - L(G),\ B_y = L(B) - \frac{L(G)+L(R)}{2} \quad (7)$$

이때 n값은 [0, 1] 사이에 존재하는 난잡음(random noise)을 나타내며, 이렇게 얻어진 값은 식 (8)을 사용하여 색상 값(H)과 채도 값(S)을 구한다.

$$H = \tan^{-1}(R_g/B_y),\ S = \sqrt{R_g^2 + B_y^2} \quad (8)$$

식 (8)에서 얻은 H, S 값은 사람의 피부색에 해당하는 색상과 채도 값을 비교하고 그에 만족하는 모든 화소 값을 마킹함으로써 입력 이미지의 얼굴 영역을 검출할 수 있다. (사람의 피부 색깔 : 색상=[110, 150], 채도=[20, 60] 이거나 색상=[130, 170], 채도=[30, 130])

그리고 Hue/Red 검출 방법은 피부색을 이루는 RGB 컬러 값을 조사한 결과, 피부색과 상관관계가 높고, G와 B보

다 분포 범위가 작은 R을 피부색 검출 인자로 사용한다. RGB 영상은 피부색 처리에서 조명 변화의 영향을 줄이기 위해 밝기 정보를 컬러 정보에서 쉽게 분리할 수 있고, 또 다양한 피부색을 효율적으로 검출하기 위하여 색의 순도를 컬러 정보에서 쉽게 분리할 수 있는 HSI 컬러공간으로 변환하여 활용한다. 즉 HSI 컬러공간에서 밝기 표현 요소로 조명의 변화에 민감한 I 요소를 피부색 검출요소에서 제외한다. 그리고 HSI 컬러 공간의 색상을 표현하는 요소인 S와 H중 다양한 피부색을 검출하기 위하여 색상의 순도인 S요소를 제외한 H요소를 피부색을 검출 하는데 이용한다. 수식 (9)는 RGB 컬러공간을 HSI 컬러공간으로 변화하는 변환식 중 H를 구하는 식이다.

$$H=\cos^{-1}\left[\frac{\frac{1}{2}[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right] \quad (9)$$

Hue/Red 검출 방법은 식 (9)에서 얻은 H와 R의 비에 의하여 피부색을 검출할 수 있다.

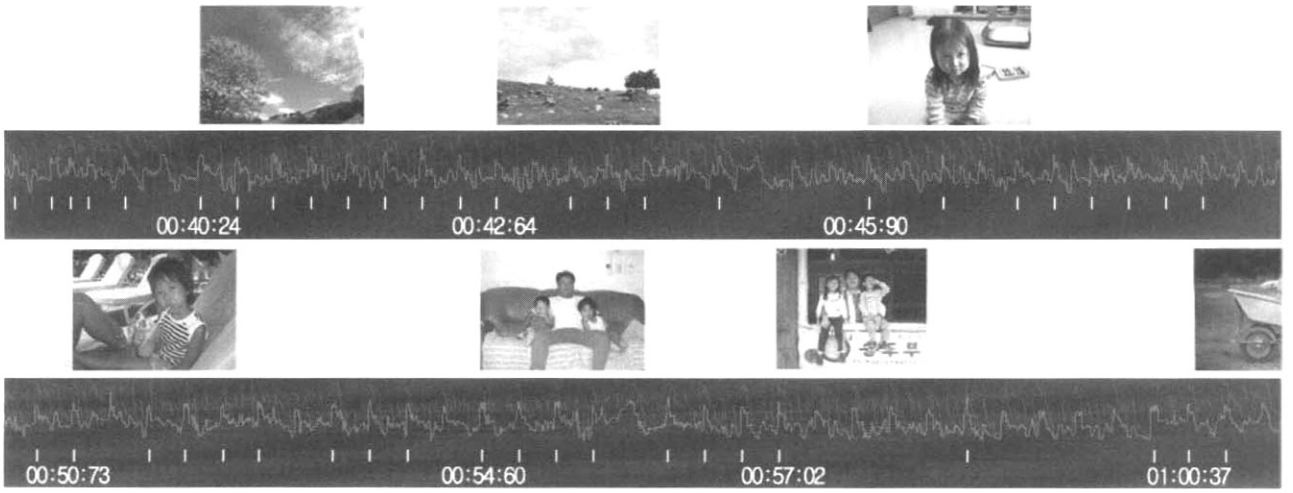
<표 1>은 실험에서 사람이 있는 사진을 인물 사진으로 분류하고, 사람이 없는 사진은 풍경 사진으로 분류한 결과 값을 나타낸다. 인물 사진만 있는 250장을 실험한 결과 얼굴 영역 검출만을 이용한 실험에서는 192장(76.8%)을 인물 사진으로 분류하였고, 얼굴 영역 검출과 구조 단순도를 함께 사용한 실험에서는 219장(87.6%)을 인물 사진으로 분류하였다. 풍경 사진만 있는 250장을 실험한 결과로는 얼굴 영역 검출을 이용한 실험에서는 203장(81.2%)을 풍경 사진으로 분류하였고, 얼굴 영역 검출과 구조 단순도를 함께 사용한 실험에서는 213(85.2%)장을 풍경 사진으로 분류하였다. <표 1>의 정답 합계와 정인식율을 보면 얼굴 영역 검출만을 이용한 실험은 395장(79%)의 정답을 찾은 데 비해 얼굴 영역 검출과 구조 단순도를 함께 이용한 실험에서는 432장(86.4%)의 정답이라는 보다 나은 결과를 얻을 수 있었다.

풍경 사진에 비해 인물 사진이 정답률이 낮은 것은 얼굴 영역 검출 기술의 오류인 것으로 예상 된다. 얼굴 영역 검출 기술은 사진의 얼굴 색상과 유사한 색상을 찾으려 하고, 유사한 색상이나 형태가 있는 경우 얼굴 영역으로 간주한다. 인물이 정면으로 보고 있는 사진의 경우 얼굴 영역 검출은 눈과 입을 찾아, 얼굴이 있다고 판단 할 수 있다. 그러나 사진의 얼굴이 옆모습인 경우나, 얼굴의 일부분이 다른 사물에 의해서 가려진 경우 오류가 나타나기 쉽다.

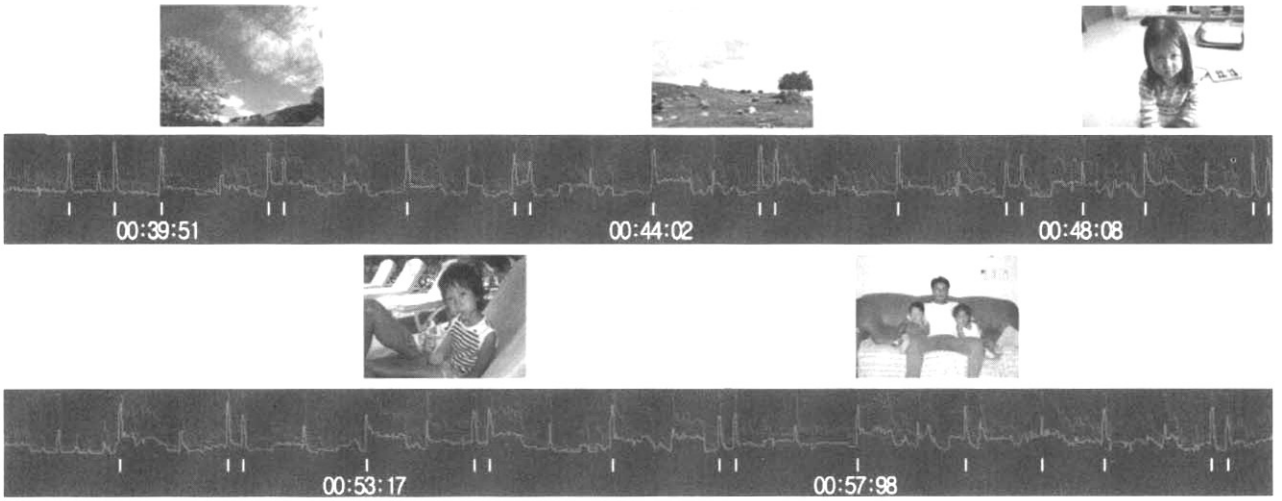
사진 분석을 통해 분류한 사진들을 3장에서 언급한 방법으로 배치하여 동영상을 제작해 보았다. 이때 음악은 MS 개수가 많이 나타나는 곡 Los Del Rio가 부른 "Macarena"와 MS 개수가 적게 나타나는 곡 Usher가 부른 "You got it bad"를 이용하였다. 사진은 인물 사진과 풍경 사진을 섞어 50장을 한 디렉토리에 넣어 분석 후 배열되게 하였다. 제작된 결과는 다음 (그림 8), (그림 9)에서 볼 수 있다. (그림 8)은 MS 개수가 많이 나타나는 곡 "Macarena"에 대한 것으로 MS 값이 높은 부분을 찾아 사진이 각각 변화됨을 볼 수 있다. 또한 사진 분석에 의해 동영상의 앞부분은 풍경

〈표 1〉 인물 사진과 풍경 사진에 대한 실험 결과

		정 답	오 답	정답 합계	오답 합계	정인식율
얼굴 영역 검출	인물 사진	192	58	395	105	79.0%
	풍경 사진	219	31			
얼굴 영역 검출 + 구조 단순도	인물 사진	203	47	432	68	86.4%
	풍경 사진	213	37			



(그림 8) MS 값이 많은 음악 "Macarena"를 이용한 영상 제작 예



(그림 9) MS 값이 적은 음악 "You got it bad"를 이용한 영상 제작 예

사진들이 먼저 나타나고, 뒷부분은 인물 사진들이 나타나 사진관의 연관성을 보임을 알 수 있었다. (그림 9)는 MS 개수가 적게 나타나는 곡 "B"를 이용한 것으로 음악 분석을 통해 "You got it bad" 곡의 리듬에 맞추어 (그림 8)에서와는 다른 사진 변화 시간을 볼 수 있었다.

즉, 실험에서는 논문에서 제시한 알고리즘을 이용하여 사용자가 원하는 사진이 있는 폴더를 지정하고, 풍경이 먼저 나올지, 인물이 먼저 나올지, 번갈아 가며 나올지를 지정함으로써 사진 간의 나타나는 순서를 사용자의 선택에 맞게 자동으로 순서를 생성해 준다. 또한 사용자가 선택한 음악을 분석함으로써 리듬에 의해 강한 소리를 찾아 그 부분에 맞추어 사진이 변화되게 한다. 빠른 노래에서는 그 리듬에 맞추어 강한 소리 부분이 자주 나타나는 만큼 여러 장의 사진이 빨리 나타나게 하고, 느린 노래에서는 강한 소리 부분이 많이 나타나지 않기 때문에 사진들의 변화가 천천히 나타나게 할 수 있다. 이는 사람들이 춤을 출 때 보이거나, 리듬에 따라 박자를 맞추는 현상과 같으며, 사람들이 일반적으로 느끼는 강한 소리 부분에서 어떠한 동작이 일어나거나, 움직임이 있기를 기대하는 것에 부합하는 동작으로서 사용자들이 UCM 콘텐츠를 제작할 때 보다 자연스러운 콘텐츠를 생성할 수 있음을 알 수 있다.

6. 결 론

본 논문은 동영상 콘텐츠 생성을 위한 음악과 사진 분석 방법을 제안하였다. 음악 분석으로는 RMS를 이용하여 노래의 강한 부분을 찾아내었으며, 사진 분석은 구조 단순도와 얼굴 영역 검출을 이용하여 사진 간의 연관성을 찾아내어 동영상 제작 시 보다 자연스러운 영상을 만들어낼 수 있게 하였다. 이러한 방법은 현재의 UCC 콘텐츠 생성에 많은 도움이 될 것이며, 사용자에게는 보다 편리하면서도 효과적인 영상 제작 툴을 제공할 수 있다. 또한 현재 사용되고 있는 디지털 액자나 모바일 폰에 전송하는 영상 제작 툴에 바로 적용이 가능하며, 나아가 PMP나 mp3 플레이어에서의 이미지와 음악을 이용한 영상 제작에도 활용이 될 수 있다.

본 논문에서 제안한 것은 음악 분석에 있어 음악의 여러 가지 특성 중 RMS만을 이용한 분석이다. 음악은 RMS 외에 Peak, 파형, 스펙트로그램 등의 특성들이 있다. 따라서 FFT (Fast Fourier Transform)를 이용한 파형 분석이나, 스펙트로그램, Peak등을 이용한 분석법을 사용하여 좀 더 정확한 MS 부분을 찾아낼 수 있을 것이다.

사진 분석은 인물 사진과 풍경 사진만을 분류하였다. 사진의 특성은 인물이 있는 것 외에도 사진이 전체적으로 이루고 있는 색상, 구도 등 다양한 특성을 가지고 있다. 따라서 그러한 다양한 특성 값을 찾아낸다면 두 가지 분류가 아닌 여러 가지 분류 체계를 가지고 사진의 연관성을 만들어 낼 수 있다. 즉, 구체화된 사진의 배열을 함으로써 동영상 제작에 있어 보다 자연스러움을 얻어 낼 수 있을 것이다. 따라서 본 연구에서 다루지 않았던 음악의 여러 가지 특성을 이용한 강약 분석과 사진이 가지고 있는 다른 특성들을 이용한 다양한 사진 분류로부터 보다 자연스러운 동영상을 제작되게 하는 것이 향후의 연구 과제이다.

참 고 문 헌

[1] Wikipedia, "*User-generated content*," http://en.wikipedia.org/wiki/User-generated_content

[2] 김문형, 남제호, 홍진우, "*UCC의 동향 및 전망*," 정보통신연구진흥원, ITFIND 주간기술동향, 제1262호, 2006. 9.

[3] Enid Burns, "*Nealy 50MM Americans Create Web Content*," ClickZ Network, ClickZ News, May 30, 2006.

[4] The Guardian, "*A Bigger bang*," The guardian Weekend, Nov. 2006.

[5] EnVible(Learners Video Network), http://www.envible.com/

[6] Haruto Takeda, Takuya Nishimoto, Shigeki Sagayama, "*Rhythm and Tempo Recognition of Music Performance*

from a probabilistic Approach," ISMIR 2004, pp.357-364, Oct. 2004.

[7] N. Whiteley, A. T. Cemgil, and S. J. Godsill. "*Bayesian modelling of temporal structure in musical audio*," ISMIR 2006, pp.29-34, Victoria, 2006.

[8] Takaaki Shiratori, Atsushi Nakazawa, Katsushi Ikeuchi, "*Dancing-to-Music Character Animation*," In Computer Graphics Forum, Vol.25, No.3, May. 2006.

[9] Jon C. Schmidt, Janet C. Rutledge, "*Multichannel Dynamic Range Compression For Music Signals*," Acoustics, Speech, and Signal Processing 1996, Vol.2, pp.1013-1016, Atlanta, May. 1996.

[10] Rainer Lienhart, Jochen Maydt, "*An Extended Set of Haar-like Features for Rapid Object Detection*," IEEE ICIP 2002, Vol.1, pp.900-903, Sep. 2002.

[11] A. Mohan, C. Papageorgiou, T. Poggio, "*Example-based object detection in images by components*," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.23, No.4, pp 349-361, Apr. 2001.

[12] Mohamed E. El-Hawary, "*Principles of Electric Machines with Power Electronic Applications*," 2nd Ed., 496page, Wiley-IEEE Press, Jun. 2002.

[13] J. Abel, D. Berners, "*On Peak-Detecting and RMS Feedback and Feedforward Compressors*," Audio Engineering Society, ISSU 5914, Britain, 2003.

[14] Karl-Heinz, Brandenburg, "*MP3 and AAC Explained*," AES, 17th Interenational Conference, Florence, Italy, Aug. 1999.

[15] S. Kiranyaz, A.F. Qureshi, M.Gabbouj, "*A fuzzy approach towards perceptual classification and segmentation of MP3/AAC audio*," International Symposium on Control, Communications and Signal Processing, pp.727-730, Hammamet, Tunisia, Mar. 2004.

정 명 범

e-mail : nzin@ssu.ac.kr

2004년 숭실대학교 미디어학부 (학사)

2006년 숭실대학교 대학원 미디어학과 (공학석사)

2006년~현 재 숭실대학교 대학원 미디어학과 박사과정

관심분야 : 감성인식, 콘텐츠공학, 멀티미디어 정보검색 등

고 일 주

e-mail : andy@ssu.ac.kr

1992년 숭실대학교 전산학과 (학사)

1994년 숭실대학교 대학원 전산학과 (공학석사)

1997년 숭실대학교 대학원 전산학과 (공학박사)

2003년~현 재 숭실대학교 미디어학부 조교수

관심분야 : 감성인식, 콘텐츠공학, 멀티미디어 정보검색 등