XCRAB : 내용 및 주석 기반의 멀티미디어 인덱싱과 검색 시스템

이 수 철[†] · 노 승 민[†] · 황 인 준[††]

요 약

최근들어 오디오, 비디오와 이미지 같은 다양한 디지털 멀티미디어 데이터의 인덱싱, 브라우징과 질의를 위한 새로운 형태의 시스템이 개발되었다. 이러한 시스템은 각 미디어 스트림을 실제 물리적 이벤트에 따라서 작은 유닛단위로 나누고, 물리적 이벤트들을 검색을 위해서 효율적으로 인덱스화 시킨다. 본 논문에서는 오디오-비주얼 데이터의 분석과 세그멘테이션을 위해서 각 데이터가 가지고 있는 오디오, 이미지, 비디오 특징을 이용하는 새로운 방법을 사용한다. 이것은 이미지나 비디오만을 분석했던 이전의 방법들을 문제점을 해결 할 수 있다. 본 논문에서는 이와 같은 방법을 이용하여 XCRAB이라고 불리는 웹 기반 멀티미디어 검색 시스템을 구현하였고, 성능평가를 위해서 여러가지 질의의 조합을 이용하여 실험을 하였다.

XCRAB : A Content and Annotation-based Multimedia Indexing and Retrieval System

Soo Cheol Lee[†] · Seung Min Rho[†] · Een Jun Hwang[††]

ABSTRACT

During recent years, a new framework, which aims to bring a unified and global approach in indexing, browsing and querying various digital multimedia data such as audio, video and image has been developed. This new system partitions each media stream into smaller units based on actual physical events. These physical events within each media stream can then be effectively indexed for retrieval. In this paper, we present a new approach that exploits audio, image and video features to segment and analyze the audio-visual data. Integration of audio and visual analysis can overcome the weakness of previous approach that was based on the image or video analysis only. We implement a web-based multimedia data retrieval system called XCRAB and report on its experiment result.

키워드 : 멀티미디어 데이터베이스(Multiemdia Database), 멀티미디어 정보 검색(Multimedia Information Retrieval), 인덱싱(Indexing)

1. 서 론

최근들어 저장 및 웹 기술의 발전에 따라서 디지털 데이터의 양이 기하 급수적으로 증가하고 있다. 대부분의 데이터들은 오디오, 이미지, 비디오와 같은 멀티미디어 데이터들을[1, 2] 포함하고 있고, 이러한 데이터들을 이용하기 위해서 대량의 정보들이 저장되어 있는 저장공간에서 효율적으로 검색할 수 있는 방법의 개발이 필요하다. 하지만 대용량의 데이터베이스로부터 멀티미디어 데이터를 검색하는 문제는 많은 제약 사항을 안고 있다. 이것은 데이터베이스 안에 존재하는 멀티미디어 데이터에 대한 보다 효율적이고 정확한 인덱싱 기술의 개발이 어렵다는 것에 기인한다. 기존의 멀디미디어 데이터 분석은 사용자가 멀티미디어 데이터를 직접 분석함으로써 적절한 내용을 텍스트 형태로 입력하는 방식[3]이었다. 이와 같은 방법은 좋은 성능을 가질 수 있는 시스템을 구성할 수 있지만, 관리자에게 많은 시간과 노력을 요구한다. 또한, 관리자에 의한 개인적 주관이 데이터에 포함될 수 있으므로 다양한 시각을 갖는 사용자의 욕구를 충족시켜 주기에는 한계가 있다. 그러므로 다양한 사용자들의 질의를 충족 시켜 줄 수 있도록 여러 특징을 사용하는 내용기반 검색(Content-based Retrieval)시스템[2, 4, 5]이 적합하다. 특히 비디오의 경우 정지영상과는 다르게 전체 비디오로부터 인덱싱에 사용될 적절한 대표 프레임들을 추출해 내는 추가적인 작업이 필요하다. 다양한 형태의 질의를 만족시켜 주기 위해 정확한 대표 프레임의 추출은 필수적이다. 대표 프레임 추출 후에 검색 단계가 개발되

※ 본 연구는 한국과학재단 목적기초연구 R05-2002-000-01224-0(2004) 지원으로 수행되었음
† 준 회 원 : 아주대학교 정보통신 전문대학원
†† 종신회원 : 아주대학교 정보통신 전문대학원 부교수
논문접수 : 2004년 5월 14일, 심사완료 : 2004년 8월 6일

지 않았다면 사용자는 비디오 장면중 현재 프레임과 동일한 장면, 혹은 특정 인물이 출연하는 장면을 모두 검색 하고자 하는 경우, 비디오의 전체 대표 프레임들을 모두 검색해야만 한다.

본 논문에서는 기존의 멀티미디어 검색시스템에서 사용하는 단편적인 검색방법과는 달리 비디오 데이터에 포함되어 있는 오디오와 이미지 정보를 이용하는 시스템을 구현하였고, 각 미디어 형을 위한 인덱싱 프레임 웍을 설계하였다. 구현 시스템은 오디오 분석을 위하여 크기(Volume)과 피치(Pitch) 또는 에너지(Energy)등 소리의 특성을 추출하고, 오디오 신호의 구간을 분석하여 스펙트로그램 형태로 표현한다. 이미지 분석은 기존의 2차원 분석방법이 아닌 3차원 분석방법을 이용한다. 3차원 분석방법은 기존의 방법이 인식하지 못하는 공간관계인 포함(Inside), 앞쪽(In front of), 바깥쪽(Outside) 등과 같은 관계들을 인식하기 때문에 정확하다. 본 논문에서 구현한 시스템과 프레임 웍은 MPEG-7 시스템의 기본적인 요구사항을 구체화 하였다. 본 논문의 구성은 다음과 같다. 2장에서는 멀티미디어 검색에 관련된 연구들을 살펴보고, 3장에서는 구현 시스템인 XCRAB의 인덱싱 프레임 웍에 대해서 기술하고 4장에서는 시스템 구조와 장면 결정과정에 대해서 설명한다. 5장에서는 구현 시스템의 인터페이스와 몇 가지 실험결과에 대해서 기술한다. 마지막으로 6장에서 결론을 맺는다.

2. 관련 연구

본 논문과 관련된 연구 분야로서 비디오를 표현하는 방법에 관한 비디오 모델 연구와 원하는 이미지, 오디오와 비디오를 검색하는 주석 기반 멀티미디어 검색, 내용 기반 멀티미디어 검색등의 멀티미디어 검색 연구를 들 수 있다. 지난 10여년동안 많은 관련 연구들이 수행되어 왔는데, 이 장에서는 그 중에서 대표적인 연구들을 소개한다.

먼저 이미지 검색에 관련된 대표적인 시스템과 연구들을 살펴보면 다음과 같다. QBIC[4]은 내용기반 검색을 이용한 대표적인 시스템으로 이미지를 효율적으로 관리, 조직하고 탐색하는 도구이다. IBM에서 개발되었으며, 데이터베이스에 저장된 이미지에 대해 시각적인 내용으로 질의를 할 수 있다. 이미지에 포함된 객체는 다르지만 색상이 유사한 경우 더 정확한 질의를 하기 위해 키워드나 텍스트를 사용한다. 또 다른 내용기반 검색 시스템들은 보다 정확한 이미지 표현 방법을 사용하는데 이것으로는 Virage[6]와 Chabot[7]이 있다. 하지만 대부분의 검색 시스템들은 이미지 객체간의 공간 정보에 대해서는 전혀 고려를 하고 있지 않다.

2D-String을 이용한 공간관계 표현기법[8, 9]은 x축과 y축에 따라서 이미지 객체를 표현하는 것으로 이미지에 있는 객체간의 방향(Direction)관계를 스트링형태로 표현하고, 2D-H, 2D-PIR과 같은 확장된 형태의 객체 표현법이 있다. 2D-H string과 2D string은 단지 방향 관계만을 표현하지만 2D-PIR string은 이미지 객체간의 방향과 위상 관계를 표현함으로써 다른 표현 기법보다 효율성 면에서 뛰어나다.

다음으로 오디오 검색에 관련된 연구로서, Ghias[10]는 마이크로 폰을 통하여 받은 사용자의 허밍에서 음높이 변화(Pitch contour)를 감지하여 UDR(Up, Down, Repeat) 스트링으로 표현하고 오디오 데이터베이스의 컨텐츠와 비교해서 유사한 멜로디를 찾아내는 시스템[10, 11]을 소개하고 있다. 웹기반 오디오 검색 시스템인 MELody index[12] 역시 마이크로 폰을 이용해서 사용자의 질의를 받고 질의 멜로디와 오디오 데이터베이스 내의 멜로디를 비교해서 유사한 정도에 따라 후보 멜로디의 리스트를 보여 주고 재생해볼 수 있도록 하였다. 오디오 검색에서는 멜로디의 음향, 간격, 리듬을 이용했고 멜로디 UDR 스트링 매칭을 위해서 다이나믹 프로그래밍 알고리즘을 이용한 유사 검색 기법을 사용하였다. Themefinder[13]는 16세기의 서양 클래식 음악, 포크송들을 대상으로 Humdrum 명령을 사용하여 웹상에서 사용자가 원하는 곡의 테마를 찾을 수 있는 시스템이다.

비디오 검색에 관련된 연구로서, Mackay 등이 개발한 주석시스템인 EVA[14]는 사용자가 주석을 달고 이를 이용하여 검색할 수 있다. 그러나 이 시스템에서는 사용자간의 주석 공유는 고려하지 않았다. Hirata와 Kato가 만든 내용 기반 이미지 검색 시스템인 QVE[15]는 이미지로부터 추출한 외곽선 데이터를 검색에 이용하는데, 실제 질의 처리 과정에서 유사성을 검사하기 위해서 데이터베이스에 저장된 각각의 이미지에 대한 이동이나 스케일링, 회전등의 기하학적인 변환을 일일이 고려해야 한다는 단점이 있다. Adali와 Subrahmanian[16] 등은 비디오 데이터에 대한 정형적 모델와 그러한 데이터를 효율적으로 저장하기 위한 공간 데이터 구조를 제공하고, 여러 종류의 비디오 질의를 처리하기 위한 알고리즘을 제시하였다. 또한 이러한 개념에 기반하여 AVIS(Advanced Video Information System)라는 프로토타입을 개발하였다. SMOOTH Video DB[17]는 Klagenfurt 대학에서 제안한 비디오 인덱싱 모델인 VIDEX를 기반으로 하여 인덱스 데이터베이스를 구축하였고, 자바 기반의 주석, 질의 및 브라우징을 할 수 있는 클라이언트와 UDP와 RTP를 지원하는 비디오 서버로 구성되어 있다. SMOOTH는 이벤트, 개체, 사람, 위치와 같은 high-level의 의미 정보들과 히스토그램과 같은 물리적인 low-level 정보들을 저장하여 텍스트 기반의 질의를 지원한다. Vane[18]은 Tcl/Tk를 이용한 유연한 질의 및 주석 인터페이스와 SGML DTD를 이용한 데이터 모델을 구현하였다. 이때 사용된 비디오 데이터는 교육용 비디오와 뉴스이며, 이들 비디오 데이터로부

터 DTD를 생성하거나 기존의 DTD를 이용하여 주석 및 질의를 하게 된다.

3. XCRAB 인덱싱 프레임 웍

멀티미디어 데이터베이스 시스템 설계시 가장 중요한 것은 사용자 인터페이스를 이용한 멀티미디어 데이터의 저장과 검색이다. 본 장에서는 멀티미디어 데이터의 인덱싱 기법을 살펴본다.

3.1 오디오 특징 분석

대부분의 오디오 신호처리 분석기법들은 인간이 소리를 인지하는 방법과 매우 유사하며, 크기(Volume)과 피치(Pitch) 또는 에너지(Energy) 등 소리의 특성들을 추출하여 이를 분석한다. 또한, 오디오 신호의 구간을 분석하여 2차원 평면상에 시간축과 주파수축에 대하여 해당 주파수의 크기를 흑백이나 칼라로 매핑시켜 표현하는 형태인 스펙트로그램(Spectrogram) 역시 많이 사용되는 오디오 신호처리 분석기법들 중 하나이다. (그림 1)은 웨이브형태의 파형과 그에 해당하는 스펙트로그램의 예를 보여준다. 본 절에서는 다음의 5가지 특성들을 이용하여 오디오 데이터를 분석하는 방법을 소개한다.

3.1.1 Short-time Average Energy

단시간 평균 에너지 함수(Short-time average energy function)는 오디오 신호처리 분석기법중 가장 많이 쓰이는 기법으로, 이때 평균 에너지라 함은 오디오 신호의 세기(Loudness)를 나타낸다. 이 함수는 주로 음성(Voice)과 잡음(Noise) 신호들을 구분하는데 사용되며 이는 다음의 수식으로부터 얻어진다.

$$E_m = \frac{1}{N}\sum_{n=0}^{N-1} x(n)^2 \times h(m-n) \tag{1}$$

$x(n)$은 입력 신호를 m은 샘플링하게 될 샘플의 개수를 의미한다. 또한, 신호처리 분석을 하기법으로 많이 사용되고 있는 FFT(Fast Fourier Transform) 분석 기법은 연속된 신호에서 임의의 일정 구간을 가져와서 분석을 하게 되는데, 이때 주로 이용되는 방법이 윈도우 함수(Window Function)이다. $h(m)$은 윈도우 함수를 나타내며, 다음은 크기 N의 사각형 윈도우(Square Window)를 보여준다.

$$h(m) = \begin{cases} 1 & \text{for } 0 \le m \le N-1 \\ 0 & otherwise \end{cases} \tag{2}$$

하지만, 이러한 사각형 윈도우를 사용하게 되면 신호가 절단되어 불연속적인 부분이 존재하게 되며 이러한 불연속점은 누설오차인 리키지(Leakage)라는 것을 발생시킨다. 오디오나 음성 신호로부터 발생되는 리키지 문제는 대부분 양끝에서의 불연속에 기인하는 것으로 FFT 기법을 사용시 양끝단의 불연속 부분을 없애기 위해서 양끝단의 값을 0에 가까운 값으로 만들어주는 윈도우를 사용한다. 이들 윈도우 중 가장 많이 사용되는 해밍 윈도우(Hamming Window)는 다음과 같다.

$$h(m) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi m}{N-1}\right) & \text{for } 0 \le m \le N-1 \\ 0 & otherwise \end{cases} \tag{3}$$

3.1.2 Average Zero Crossing Rate

영교차율(ZCR)은 오디오 신호가 영점 축을 교차하는 회수를 나타내며, 이는 다음의 수식으로부터 얻어진다.

$$ZCR = \frac{\sum_{n=1}^{N} |\, sgn\, x(n) - sgn\, x(n-1)}{2N}, \tag{4}$$

$sgn\, x(n)$이 '0'보다 큰 경우(양수)에는 '1'값을 가지고, 음수인 경우에는 '-1'값을 가진다. 식 (4)에서 얻어지는 평균 영교차율은 주로 음성 신호와 비음성(무성) 신호를 구별하는데 사용되는데, 그 이유는 대부분의 무성 신호가 음성 신호에 비해 더 높은 영교차율 값을 가지기 때문이다. 즉, 영교차율 값이 높다는 것은 오디오 신호가 무성(Unvoiced) 신호를 가진다는 뜻이고, 그렇지 않은 경우에는 영점 축을 교차한 회수가 적은 음성 신호를 의미한다.

3.1.3 Energy Distribution

에너지 분포(Energy Distribution)는 소리의 세기를 나타내는 에너지가 가지는 높고 낮은 주파수 특성들의 분포를 말하는 것으로, 대역폭(Bandwidth)과 함께 음악과 음성 신호를 구별하는데 유용하게 쓰인다. 예를 들어, 음악은 주로 음성 신호에 비해 높은 주파수 대역을 가지고 있기 때문에 주파수 대역의 높고 낮음을 측정하는 것은 매우 중요하다. 이러한 주파수 대역의 높고 낮음은 특정 애플리케이션에 좌우되는 경우가 많은데 이는 대부분 음성 신호의 주파수가 7Hz를 거의 넘지 않기 때문이며, 7Hz를 기준으로 주파수 대역의 높고 낮음을 결정할 수 있게 된다.

3.1.4 Bandwidth

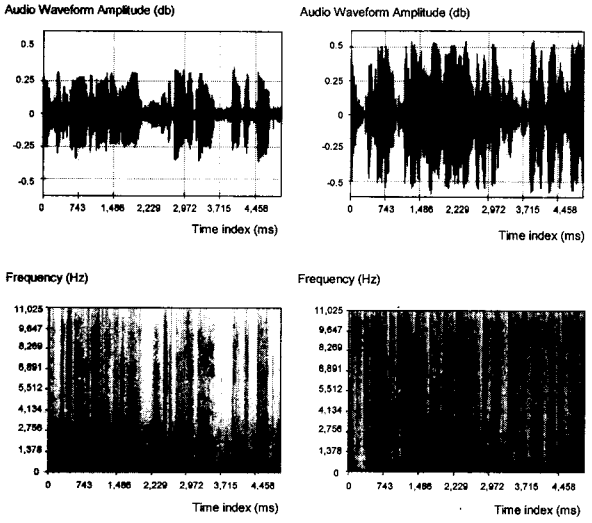
대역폭은 오디오 신호의 최고 주파수와 최저 주파수의 차이를 의미하며, 주파수 대역폭이라고도 말한다. 대역폭을 계산하는 가장 간단한 방법으로는 최고 주파수에서 최저 주파수를 빼는 방법이 있으며, 이때 최저 주파수는 silence 레벨인 3dB(데시벨)을 넘는 주파수로 정의된다. 대부분의 음악은 음성 신호에 비해 더 높은 대역폭을 가지고 있기 때문에, 대역폭은 에너지 분포와 함께 오디오 신호로부터 음악과 음성 신호를 구분하기 위해 많이 사용된다.

3.1.5 Harmonicity

조화도(Harmonicity)는 가장 낮은 파형을 가지는 기본주파수(Fandamental Frequency)와 주기적으로 반복 파형을 가지는 정수배의 주파수인 배음(음의 배합)을 가지는 오디오 신호를 의미한다. 예를 들어, 기본주파수가 220Hz인 경우 440Hz, 660Hz, 880Hz와 같이 기본주파수의 정수배인 배음을 가진다. 이러한 음의 조화도는 기본주파수를 계산함으로써 얻어낼 수 있으며, 기본주파수는 다음의 수식으로부터 얻어진다.

$$F_n = |FFT(x(m) \times w(n-m))|, \quad (5)$$

$x(m)$은 입력 신호를 n은 샘플링하게 될 샘플의 개수를 의미하고, $w(n)$은 단시간 평균 에너지 함수와 마찬가지로 해밍 윈도우 함수를 나타낸다. 대부분의 악기는 기본주파수에 여러 하모닉스(Harmonics : 고조파)를 포함하므로써 사운드의 특성을 나타내고 있는데, 이러한 특성은 악기를 주로 사용하는 사운드인 음악의 경우에 많이 나타난다.

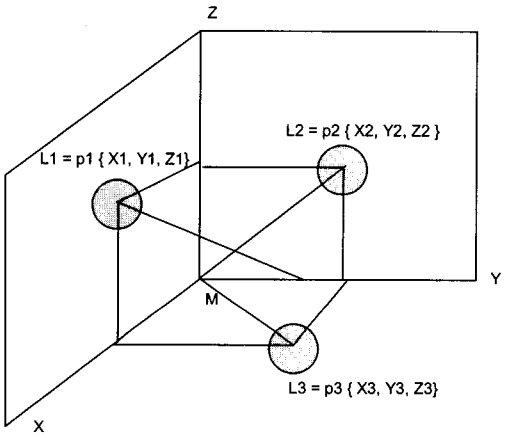


(그림 1) 오디오 파형과 스펙트로그램

3.2 이미지 인덱스 기법

멀티미디어 데이터베이스에서 효율적인 이미지 검색과 저장을 위해서 가장 중요한 것은 인덱싱이다. 이미지는 이미지 객체에 대한 일반적인 정보와 이러한 정보에 관련된 모양(Shape), 공간위치(Spatial location)와 같은 두 가지 정보를 포함하고 있다. 이미지 데이터베이스 내의 이미지들을 효율적으로 관리하기 위해서 각 이미지에 공간정보를 내포(Embed)시킨다. 본 논문에서 이미지 객체의 공간정보는 x축, y축, z축에 따라서 이미지 객체를 표현하는 3D 방식을 사용하고, 멀티미디어 데이터베이스에 질의를 하기위해서 사용자는 자신이 관심 있는 이미지 객체를 선택하는 ROI (Region of Interest)[19]를 지정한다. 이렇게 지정된 객체들은 2D-string 형태로 표현되는 것이 아니라 3차원 형태로 표현된다.

데이터베이스에 저장된 모든 이미지들은 각각이 독특한 특성을 가진 객체들로 구성되어 있고, 이러한 객체들 간에는 다양한 공간 관계가 존재할 수 있다. 공간 관계는 상대좌표와 절대좌표로 표현할 수 있고, 3차원 공간에서의 경우 객체 O의 공간 위치 좌표 P0 = (X0, Y0, Z0)로 나타낼 수 있다. 따라서 만약 한 이미지들에 n개의 객체가 존재하는 경우 전체 이미지는 P = {P1, P2, …, Pn}의 위치 좌표의 집합으로 표현될 수 있다. 각 위치 좌표에 해당하는 객체는 의미적 정보를 가지고 있기 때문에, 이러한 정보를 주석 처리할 수 있다. 이와 같은 위치 좌표를 공간 위치 점(Spatial location point)[19]이라고 부른다. 단순화를 위해서 이미지 객체는 하나의 공간 위치 좌표로 표현된다. 이미지의 공간 위치 점들 간의 공간관계를 표현하기 위해서, 이미지를 같은 크기의 3차원 공간으로 분할한다. 그림 2는 공간 위치 점이 서로 다른 공간에 위치한 세 개의 오브젝트와 중심점 M과의 공간 위치 좌표와 예를 보여준다.



(그림 2) 공간 위치 좌표

공간 위치 좌표를 기반으로, 본 논문에서는 이미지 객체들의 위치에 따른 이미지 위치 연산자를 표 1과 같이 정의하였다. n개의 객체들로 구성된 이미지가 있다고 가정할 때, 그래프를 이용해서 객체간의 공간 관계를 정의할 수가 있는데, 이것을 공간 그래프(Spatial graph)라고 한다.

(그림 3)은 원본 이미지와 그것의 공간 그래프를 나타낸다.

[정의 1] 공간 그래프는 그래프내의 각 엣지에 레이블이 지정되어 있는 비방향성(Undirectional) 그래프로 두 개의 원소로 이루어진 순서쌍(V, E)이다.

- V = {L1, L2, L3, …, Ln}으로 구성된 노드의 집합으로 이미지 객체를 표현한다.
- E = {e1, e2, e3, …, en}으로 구성된 엣지의 집합으로 두

개의 노드 L1, L2를 연결하고, 노드간의 공간관계를 표시한다.

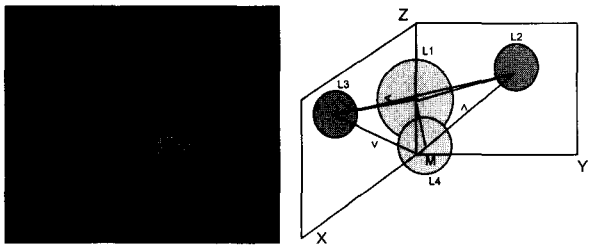
〈표 1〉 이미지 위치 대수

Notation	Operator	Meaning
X < Y	Lupper	X는 Y의 왼쪽 위에 위치
X ∧ Y	Llower	X는 Y의 왼쪽 아래에 위치
X > Y	Rupper	X는 Y의 오른쪽 위에 위치
X ∨ Y	Rlower	X는 Y의 오른쪽 아래에 위치
X ∪ Y	Upper	X는 Y의 위에 위치
X ∩ Y	Below	X는 Y의 아래에 위치
X] Y	Right	X는 Y의 오른쪽에 위치
X [Y	Left	X는 Y의 왼쪽에 위치
X / Y	Center	X 또는 Y는 M의 중앙에 위치
X % Y	Overlap	X와 Y는 중첩 되어있음
X ⊗ Y	Inside	X는 Y의 안에 위치
X ⊕ Y	Outside	X는 Y의 밖에 있음
X • Y	In front of	X는 Y의 앞에 위치

[정의 1]에 따라서 (그림 3)의 공간 관계를 표현하면 다음과 같다.

$V = \{L1, L2, L3, L4\}$

$Rel = \{L1 \cup M\}, \{L1 > L2\}, \{L1 \wedge L3\}, \{L1 \bullet L4\},$
$\{L2 \wedge M\}, \{L2 \wedge L1\}, \{L2 \wedge L3\}, \{L2 \wedge L4\},$
$\{L3 \vee M\}, \{L3 < L1\}, \{L3 < L2\}, \{L3 \vee L4\},$
$\{L4 \bullet L1\}, \{L4 > L2\}, \{L4 < L3\}$



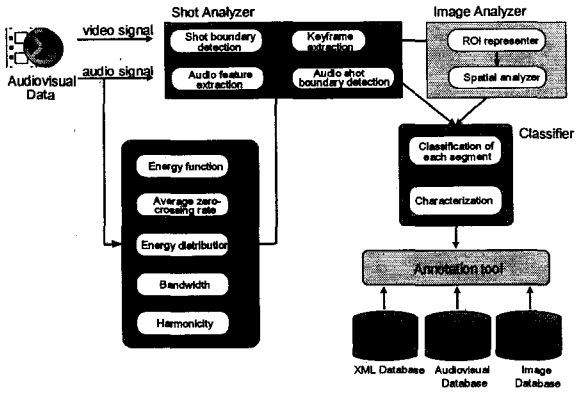
(그림 3) 원본 이미지와 공간 그래프

4. XCRAB 시스템

본 장에서는 논문에서 제안하는 주석p 및 내용 기반 멀티미디어 검색 시스템인 XCRAB에 대해서 설명한다. XCRAB 시스템은 이식성과 플랫폼 독립성을 위해 자바로 구현하였으며, 일부 기능은 애플릿으로 구현되어 어떠한 시스템에서라도 브라우저를 통해 이용할 수 있다. 비디오 컨텐츠는 MPEG을 사용하였고 분석된 키 프레임들은 JPEG 형태의 이미지로 저장되어 이미지 분석기를 통하여 데이터베이스에 저장된다. 오디오 컨텐츠는 비디오 컨텐츠로부터 WAVE 형태의 파일로 저장되어 샷 분석기를 통하여 오디오 음향 특징 정보를 추출하였다.

4.1 시스템 구조

(그림 4)는 XCRAB 시스템의 개략적인 구조와 멀티미디어 데이터의 분석 과정과 질의의 흐름을 순서대로 보여준다. XCRAB 시스템은 크게 샷 분석기(Shot Analyzer), 이미지 분석기(Image Analyzer), 분류기(Classifier), 그리고 주석 도구(Annotation Tool)로 구성되어 있다.



(그림 4) XCRAB 시스템 구조

본 논문에서는 시각적인 정보인(Visual Information) 색상 히스토그램 같은 이미지 특성만을 이용하는 기존의 시스템과는 달리 음향/영상 정보를(Audio/Visual Information) 같이 이용하여 비디오 데이터를 분석하는 기법을 제안하고 있다. 비디오 데이터의 분석 과정은 다음과 같다. 우선 비디오 데이터를 샷 경계 검출(Shot Boundary Detection) 알고리즘을 이용하여 샷의 경계 검출과, 키 프레임(Key Frames) 정보를 추출한다. 추출된 키 프레임들은 각각 정지 화상(Still Image)인 JPEG 포맷으로 저장된 후, 사용자에 의해 이미지 내의 관심 개체들에 대한 공간관계 분석을 하게된다. 또한, 오디오 데이터는 비디오 데이터로부터 추출되어 WAVE 포맷으로 저장된 후, 앞서 3장에서 언급한 분석 방법을 이용하여 오디오 특징들을 분석하고 샷의 경계 검출을 하게 된다. 이렇게 검출된 오디오 및 비디오 샷들은 분류기를 통하여 의미있는 장면(Scene)을 결정하고, 주석 처리된 후 데이터베이스에 저장된다.

시스템의 각 주요 구성요소들을 살펴보면 다음과 같다.

4.1.1 샷 분석기(Shot Analyzer)

샷 분석기는 크게 비디오와 오디오 분석기의 두 모듈로 구성되어 있다. 비디오 분석기는 샷 경계 검출 알고리즘 중 색상 히스토그램과 움직임 벡터를 이용하여 하나의 비디오 데이터에 대한 모든 키 프레임과 샷 경계를 검출한다. 샷

경계 검출 알고리즘을 이용하여 분할된 샷들 중에는 장면의 전환이 빈번하게 이루어져 실제로는 하나의 샷임에도 불구하고 여러 개의 샷으로 분석되는 경우가 발생한다. 이와 같이 부정확한 샷의 검출을 줄이고자, 프레임의 개수가 10개 이하인 샷들은 그 이전 샷과 병합하여(Merge) 새로운 샷을 생성하게 된다. 오디오 분석기는 오디오 데이터를 3.1절에서 언급했던 여러 오디오 특성들을 분석하여 특징 백터들을 만들고, 이들 백터들을 이용하여 다음과 같이 6개의 오디오 타입으로 분류한다 : Silence, Pure Speech, Music, Speech with Music, Environmental Sound, and Speech with Environmental Sound. 각 오디오 샷들은 하나의 오디오 타입을 가지게 되며, Silence 샷들은 의미 있는 오디오 샷으로 판단하지 않는다. 따라서, 우선적으로 각각의 오디오 프레임에 대해서 ZCR과 Energy값을 구하여, 이들 값들이 각각 해당 임계치 값인 50을 넘거나 3(dB : decibel)보다 적은 경우에 이를 Silence Frame이라고 정의한다. 하나의 오디오 샷내에 Silence Frame의 비율이 70%를 넘게되는 경우 이를 Silence Shot이라고 정의하고, 이를 의미없는 샷이라고 판단하여 비디오 샷 검출과 마찬가지로 이전 샷과의 병합을 통하여 새로운 오디오 샷을 생성하게 된다. 이렇게 분석된 오디오와 비디오 샷들은 이미지 분석기를 통하여 재 분석된 후 분류기를 거쳐 의미있는 장면을 결정하게 된다.

4.1.2 이미지 분석기(Image Analyzer)

비디오 분석기를 통하여 추출되어 저장된 키 프레임에 해당되는 이미지들로부터 사용자가 정의한 관심영역과 그들 간의 공간관계들을 분석하는 것이 이미지 분석기의 기능이다. (그림 4)에서 보는것과 같이, 사용자는 관심영역 표현기(ROI Rrepresenter)를 통해서 해당 이미지내의 관심있는 개체들을 선택하게 된다. 이렇게 선택된 개체들간의 공간관계는 3.2절에서 언급했던 공간 관계 그래프와 위치 연산에 의해서 분석된다.

4.1.3 분류기(Classifier)

샷 분석기와 이미지 분석기를 통하여 분석된 오디오와 비디오 샷들은 장면 결정 알고리즘을 이용하여 비슷한 시간에 발생한 오디오/비디오 샷들을 비교 분석하여 의미있는 장면들로 분류된다.

4.1.4 주석 도구(Annotation Tool)

비디오 샷 분석기와 이미지 분석기를 통해서 추출된 2차원, 3차원상의 공간관계 분석정보를 팔레트(palette)를 이용하여 백그라운드나 특정 객체의 색상을 구체적으로 선택하거나 키워드를 이용하여 주석처리를 한다. 비디오 역시 주석 처리 이전에 분석기를 통하여 검출된 샷 정보와 이들 샷들을 보면서 특정 비디오의 시간 위치에 대한 이벤트와 풍부한 정보들에 대해 서술할 수 있다.

4.2 장면 결정(Scene Determination)

비디오 및 오디오 샷 분석기와 이미지 분석기를 통해 추출된 비디오와 오디오 샷들은 우선 후보 샷이 되며, 각 후보 비디오 샷과 후보 오디오 샷들간의 장면 전환 시점을 기준으로 병합(Merging)과 조정(Adjusting)을 통해 새로운 후보 샷을 정하고, 이들 후보 샷들내의 대표 오디오 타입에 따라 의미있는 장면을 결정하게 된다[20].

CSv_i ($i = 1,\cdots, n$), CSa_j ($j = 1,\cdots, m$) = Candidate shot boundaries extracted by the video and audio shots
$t(CS)$ = Starting time of a candidate shot boundary
$F(t(CSa))$ = Audio features applied at time $t(CSa)$ {*Silence, Speech, Music, Speech with Music, Environmental Sound, and Speech with Envirnomental Sound*}

Step 1 : *If* ($t(CSv_i) = t(CSa_j)$) *then*
Candidate scene boundary is detected
go to step 3
Else
go to step 2

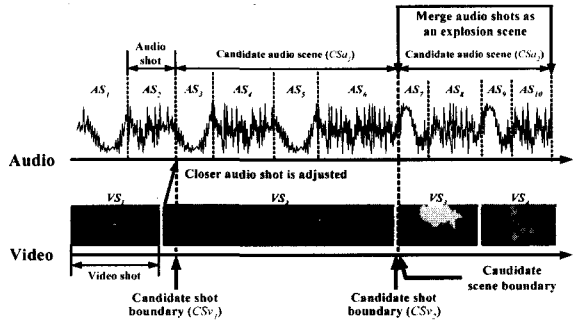
Step 2 : $diff1 = Diff(t(CSv_i, CSa_j))$, $diff2 = Diff(t(CSv_i, CSa_{j+1}))$
Candidate shot boundary is adjusted to $t(CSa_j)$
(*when*, $diff1 < diff2$)
Candidate shot boundary is adjusted to $t(CSa_{j+1})$
(*when*, $diff1 = diff2$)
go to step 3

Step 3 : $Max1 = Max(F(t(CSa)))$ *between* $t(CSv_i)$ *and* $t(CSv_{i+1})$
$Max2 = Max(F(t(CSa)))$ *between* $t(CSv_{i+1})$ *and* $t(CSv_{i+2})$
If ($dist(Max1, Max2) = Tf$) *then*
Merge the consecutive video shots (CSv_i *and* CSv_{i+1}) *and adjust a candidate shot boun dary to* $t(CSv_{i+1})$
Else
Scene boundary is determined

(그림 5) 장면 결정 알고리즘

(그림 5)와 (그림 6)은 장면 결정 알고리즘과 그 과정을 보여주고 있다. 후보 샷들로 이루어진 집합을 후보 장면이라고 정의하며, 해당 후보 장면들을 검출하는 과정은 다음과 같다. 우선, 첫 번째 단계에서 오디오 후보 샷(CSa_j)과 비디오 후보 샷(CSv_i)이 발생한 시점이 같다면, 후보 장면이라고 잠정적으로 결정하게 된다. 그렇지 않은 경우에는 두 번째 단계로 넘어가서 비디오 후보 샷이 발생했던 시점을 기준으로 비슷한 시점에 발생한 두 개의 이웃한 오디오 후보 샷(CSa_j, CSa_{j+1})과의 시간차를 비교하여, 그 중 가까운 시간차를 가지는 오디오 후보 샷의 발생 시점으로 시간을 조정한다. 마지막 단계에서는 두 번째 단계에서 조정된 비디오 후보 샷에 해당되는 오디오 후보 샷들을 비교하게 된다. 해당 비디오 후보 샷 내의 오디오 후보 샷들 중 가장 높은 비율을 가지는 오디오 후보 샷을 대표 후보 샷으로 지정하고, 이웃한 비디오 후보 샷들(CSv_i & CSv_{i+1}, CSv_{i+1} & CSv_{i+2})

간의 대표 후보 샷을 비교하여 임계값(T_f)보다 같거나 작은 경우에는 병합과 조정을 통해 새로운 후보 샷을 결정하고, 그렇지 않은 경우에 이를 의미있는 장면이라고 결정하게 된다.

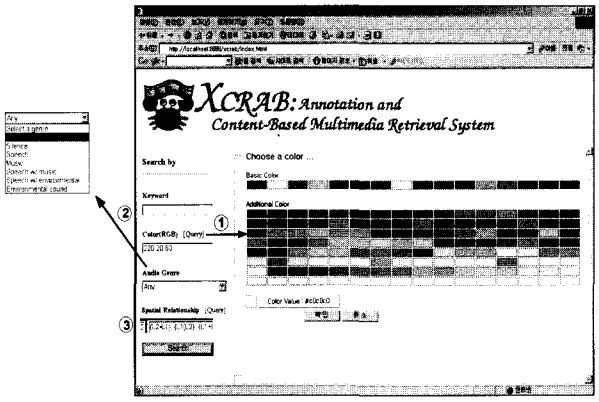


(그림 6) 장면 결정 과정

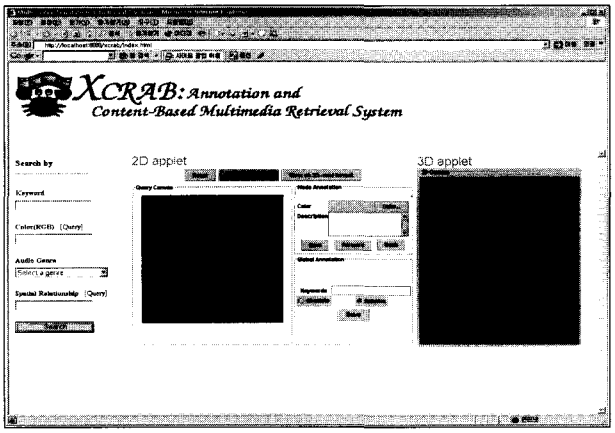
(그림 6)은 (그림 5)에서 설명하고 있는 장면 결정 알고리즘을 통하여 각 후보 샷들이 어떻게 병합 및 조정이 되며, 장면을 결정하는지에 대해 자세히 보여주고 있다. AS는 오디오 샷을 VS는 비디오 샷을 나타내며, CSa와 CSv는 각각 오디오 후보 샷들과 비디오 후보 샷들을 의미한다. 비디오 샷인 VS_2와 비슷한 시점에 발생한 오디오 샷들인 AS_2와 AS_3의 시간차를 비교한 뒤, 좀 더 가까운 시점에 발생한 오디오 샷인 AS_3의 발생 시점으로 시간을 조절하고 이를 첫 번째 비디오 후보 샷(CSv_1)으로 정의한다. 이와 같은 방법으로 나머지 후보 샷들을 정의하며, 해당 후보 샷들내의 오디오 샷들 중에서 가장 높은 비율을 가지는 대표 샷을 지정한다. 해당 후보 샷들내에 대표 오디오 샷들이 결정되면, 이웃한 후보 샷들간의 대표 오디오 샷들을 비교하여 유사도 값이 일정 임계값을 넘는 경우에는 의미 있는 장면이라고 결정하게 된다. 예를 들면, 비디오 후보 샷인 CSv_2내의 오디오 대표 샷은 AS_8과 AS_{10}이 해당 샷 내에서 차지하는 비율이 다른 오디오 샷들에 비해 크기 때문에, 이를 병합하여 대표 샷으로 지정하게 된다. 이와 같은 방법으로 결정된 CSv_1내의 오디오 대표 샷과의 비교를 통해 CSv_2를 하나의 의미있는 장면으로 결정하게 된다.

5. 구 현

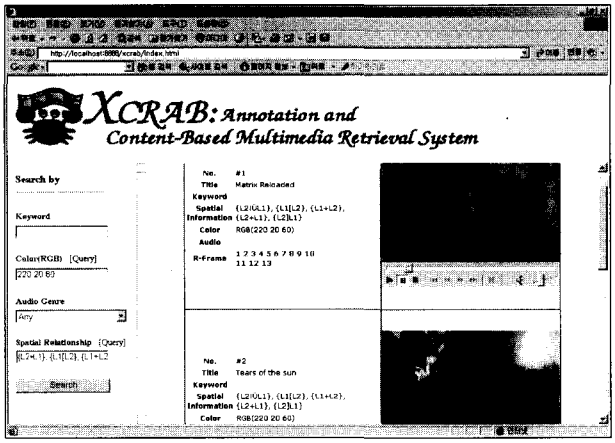
(그림 7)은 XCRAB 시스템의 사용자 질의 인터페이스를 보여준다. 인터페이스는 질의 방법에 따라 키워드, 색상, 오디오, 공간 관계 등으로 나누어 진다. 색상에 의한 질의는 (그림 7)①에서 보여지듯이, 팔레트로부터 해당 색상을 선택하면 RGB 값으로 변환되어 입력된다. 오디오의 경우는 6개의 오디오 장르로부터 원하는 오디오 타입을 선택하여 질의를 할 수 있다. (그림 7)③은 공간 관계 질의를 나타내며, 캔버스상에서 개체들간의 공간 관계의 설정은 (그림 8)과 같다.



(그림 7) 사용자 질의 인터페이스



(그림 8) 공간 관계 질의 인터페이스



(그림 9) 검색 결과 화면

(그림 8)은 스케치 모드의 질의 캔버스(Query Canvas)를 이용하여 질의 개체를 생성하고 이들 간의 공간 관계를 2차원 및 3차원으로 분석하는 인테페이스를 보여준다. 우선, 2D Applet 질의 캔버스에서 공간위치 점인 R을 기준으로 하여

개체들간의 공간 관계를 표현한다. 그런 후에 3D Applet 질의 캔버스상에서 개체들간의 3차원 공간 관계를 재설정한 후 저장하면 (그림 7)③과 같이 위치 연산식으로 표현된다.

(그림 9)는 (그림 7)과 (그림 8)에서의 질의를 기반으로 검색한 결과 화면을 보여준다. 우선 색상의 RGB 값인 "220 20 60"과 두 개체들간의 공간 관계를 질의 조건으로 주었으며, 오디오의 장르는 "Any"로 하였다. 화면상의 공간 관계를 위치 연산식으로 표현하면 다음과 같다.

$$\{L2 \bullet L1\} \ \& \ \{L1 \ [\ L2\} \ \& \ \{L1 \oplus L2\} \ \& \ \{L2 \oplus L1\} \ \& \ \{ \ L2 \ ] \ L1\}$$

5.1 실 험

실험을 위해서 구현 시스템을 윈도우 2000환경에서 실행하였고, 비디오 데이터는 데이터의 손실과 전송률을 향상시키기 위해서 RAID 저장 시스템에 저장하였다. 실험에서 사용하는 비디오 데이터는 영화 스포츠, 뉴스 클립을 포함하고 있는 비디오 세그먼트로 구성되어 있다. 구현 시스템인 XCRAB에서 비디오 객체의 오디오, 공간 정보와 주석의 여러 가지 조합으로 질의를 생성하였다. <표 2>는 트리 앞에 있는 빨간색 집이 있는 비디오 샷을 찾기 위한 질의에 대한 결과를 보여주는 것으로, 데이터베이스에는 다음과 같은 질의가 전달된다.

Q1 : "Find the video shots with a red house in front of the tree"

질의 1은 데이터베이스 에서 모든 집과 관련된 샷들을 검색하는 것으로 나무 앞이라는 공간 조건과 빨간색 이라는 색상 조건를 포함하고 있다. 데이터베이스 내에서 질의에 정확하게 부합하는 비디오 샷은 5개가 있다. 표에서 질의 1은 키워드만을 이용하고, 2-3은 키워드와 색상 정보의 조합으로 질의를 하였다. 질의 4-5는 키워드와 색상정보외에 공간 정보를 질의에 조합하여 사용하였다. 키워드 "House and Tree"와 공간 정보인 "In front of"를 조합한 질의가 가장 정확도가 높음을 알 수 있다.

<표 2> 단순 질의(Q1) 결과

#	Spatial Constraint	Color	Keyword	Retrieved	Relevant	Precision
1	" - "	-	Tree	215	5 House	2.3%
2	" - "	Red		47	5	10.6%
3	" - "	Red	House and Tree	21	5	23.8%
4	Left or Right	Red	House and Tree	13	5	38.5%
5	In front of	Red	House and Tree	5	5	100%

다른 실험으로 다음과 같이 오디오, 공간 정보, 색상, 키워드를 포함하는 좀 더 복잡한 질의를 생각해보자.

Q2 : "Find the video shots have a man and woman singing in the rain and wear a red hat"

질의 Q2는 빗속에서 노래하고, 빨간색 모자를 쓴 모든 남자와 여자를 포함하는 비디오 샷을 검색하는 질의이고, 이것을 만족하는 비디오 샷은 24개가 있다. 표에서 보면 알 수 있듯이 키워드와 공간 정보를 이용한 질의는 정확도에 있어서 중간정도의 성능을 보인다. 하지만 공간 정보와 오디오 정보를 조합하여 질의를 하게 되면 보다 좋은 성능을 보임을 알 수 있다.

<표 3> 복합 질의(Q2) 결과

#	Audio	Spatial Constraint	Color	Keyword	Retrieved	Relevant	Precision
1	" - "	" - "	" - "	Man	786	24	3%
2	" - "	" - "	Red	Woman	358	24	6%
3	" - "	" - "	Red	Man and Woman	142	24	17%
4	" - "	" - "	Red	Man and Woman and Hat	85	24	28.2%
5	" - "	Left or Right	Red	Man and Woman and Hat	57	24	42.1%
6	" - "	Overlap or Left or Right	Red	Man and Woman and Hat	43	24	56%
7	Singing	Left or Right	Red	Man and Woman and Hat	18	24	75%
8	Singing	Overlap or Left or Right	Red	Man and Woman and Hat	24	24	100%

6. 결 론

본 논문에서는 멀티미디어 데이터의 인덱싱 프레임 웍과 이것을 기반으로한 검색 시스템인 XCRAB을 구현하였다. 구현 시스템은 기존의 검색 시스템과는 달리 여러 형태의 미디어 형을 이용하여 검색을 수행하고, 주석과 내용

을 이용한 복합질의를 지원하기 때문에 검색의 정확도가 높다.

실험 결과에서 알 수 있듯이 정지 영상의 검색에서는 공간 정보와 색상이나 키워드 같은 특징을 복합적으로 이용하여 질의를 처리할 경우 정확도가 높아짐을 알 수 있었다. 비디오 검색에서는 오디오 정보를 포함한 질의가 다른 질의보다 정확도가 좋음을 알 수 있었다. 본 논문에서 제안한 인덱싱 프레임 웍은 질의의 성능을 향상 시켜주고 여러가지 멀티미디어 형에 대해서 유연하게 사용할 수 있음을 알 수 있다.

멀티미디어 데이터베이스의 인덱싱 역시 검색의 효율을 결정짓는 중요한 요소이기 때문에 향우에는 구현된 시스템에 다차원 인덱싱 기법을 적용하여 대용량의 멀티미디어 데이터에 대해서도 효율적인 검색이 가능하도록 할 것이다.

참 고 문 헌

[1] Dongge Li, Ishwar K. Sethi, Nevenka Dimitrova, Thomas McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters*, Vol.22, No.5, pp.533-544, 2001.

[2] M. Flickner et al., "Query by Image and Video Content : The QBIC System," *Computer*, Vol.28, No.9, pp.23-32, 1995.

[3] B. Y. Ricardo and R. N. Berthier, Modern Information Retrieval, *ACM press*, 1999.

[4] W. Niblack, et al., "The QBIC project : Query images by content using color, texture and shape," SPIE V 1908, 1993.

[5] J. R. Smith and S.-F. Chang, "VisualSEEk : a fully automated content-based image query system," *ACM Multimedia*, Boston, May, 1996.

[6] N. Kosugi, Y. Nishihara and T. Sakata, "A Practical Query-By-Humming System for a Large Music Database," *Proc. of ACM Multimedia 2000 Conference*, November, 2000.

[7] V. E. Ogle and M. Stonebraker, "Chabot : Retrieval from a Relational Database of Images," *IEEE Computer*, Vol.28, No.9, September, 1995.

[8] Lee, S. Y. and F. J. Hsu, "Spatial reasoning and similarity retrieval of images using 2D-C String knowledge representation," *Pattern Recognition*, Vol.25-3, pp.305-318, 1992.

[9] M. Nabil, A. H. H. Ngu and J. Shepherd, "Picture Similairty Retrieval Using the 2D Projection Interval Representation," *IEEE Trans. Knowledge and Data Eng.*, Vol.8, No.4, pp. 533-539, Aug., 1996.

[10] A. Ghias, J. Logan, D. Chamberlin and B. Smith, "Query by humming - musical information retrieval in an audio database," *Proc. of ACM Multimedia Conference*, San Francisco, 1995.

[11] S. Rho and E. Hwang, "FMF(Fast Melody Finder) : A Web-based Music Retrieval System," *Lecture Notes in Computer Science, Springer-Verlag*, Vol.2771, pp.179-192, 2003.

[12] R. J. McNab, L. A. Smith, D. Bainbridge and I. H. Witten, "The New Zealand digital library MELody inDEX," *D-Lib Magazine*, May, 1997.

[13] Huron, D., C. S. Sapp and B. Aarden, Themefinder, 2000. http;//www.themefinder.org.

[14] W. E. Mackay, G. Davenport, "Virtual video editing in interactive multimedia applications," *Communications on ACM*, 32, pp.802-810, 1989.

[15] K. Hirata and T. Kato, "Query by visual example-content based image retrieval," *Advances in Database Technology* (EDBT '92), pp.56-71, 1992.

[16] S. Adali., et al., "The Advanced Video Information System : data structures and query processing," *ACM Multimedia Systems*, Vol.4, No.4, pp.172-186, 1996.

[17] H. kosch, R. Tusch, et al., "The SMOOTH Video DB - Demonstration of an integrated generic indexing approach," *ACM Multimedia Conference*, Los Angeles, USA, pp.495-496, October-November, 2000.

[18] M. Carrer, L. Ligresti and T. D. C. Little, "A Tcl/Tk-Based Video Annotation Engine," *Proc. USENIX, Fifth Annual Tcl/Tk Workshop*, Summer, 1997.

[19] S. Lee and E. Hwang, "Spatial Similarity and Annotation-Based Image Retrieval System," *IEEE Fourth International Symposium on Multimedia Software Engineering*, Newport Beach, CA, December, 2002.

[20] S. Rho and E. Hwang, "Video Scene Determination using Audiovisual Data Analysis," *Proc. of the 24th International Conference on Distributed Computing Systems (ICDCS'04) Workshops - Multimedia Network Systems and Applications* (MNSA '04), Tokyo, Japan, pp.124-129, March, 2004.

이 수 철

e-mail : juin@ajou.ac.kr

1998년 한남대학교 컴퓨터 공학과(학사)

1998년~2000년 아주대학교 정보통신 전문대학원(석사)

2000년~2002년 아주대학교 정보통신 전문대학원 박사수료

2003년~현재 아주대학교 정보통신 전문대학원 박사과정

관심분야 : 데이터베이스, 멀티미디어 시스템, 정보 통합, XML 응용, 유비쿼터스 컴퓨팅

노 승 민

e-mail : anycall@ajou.ac.kr

2001년 아주대학교 컴퓨터 공학과(학사)

2001년~2003년 아주대학교 정보통신 전문대학원(석사)

2003년~현재 아주대학교 정보통신 전문대학원 박사과정

관심분야 : 데이터베이스, 멀티미디어 시스템, XML 응용, XML 보안

황 인 준

e-mail : ehwang@ajou.ac.kr

1988년 서울대학교 컴퓨터공학과(학사)

1990년 서울대학교 컴퓨터공학과(석사)

1998년 Univ. of Maryland at College Park 전산학과(박사)

1998년~1998년 Hughes Research Lab. 연구교수

1998년~1999년 Bowie State Univ., Assistant Professor

1999년~2002년 아주대학교 정보통신전문대학원 조교수

2003년~현재 아주대학교 정보통신 전문대학원 부교수

관심분야 : 데이터베이스, 멀티미디어 시스템, 정보 통합, 전자상거래, XML 응용, 유비쿼터스 컴퓨팅