하이브리드 인식을 이용한 불법 콘텐츠 추적시스템 설계 및 구현

김원겸*, 박경수*, 김상진*, 유원영**
*(사)한국저작권단체연합회 저작권보호센터
**한국전자통신연구원
e-mail:{wgkim, kspark, sjkim}@cpcmail.or.kr, zero2@etri.re.kr

Design and Implementation of Illegal Content Tracking System Using Hybrid Content Recognition

Won-gyum Kim*, Kyungsoo Park*, Sangjin Kim*, Won Young Yu**
*Copyright Protection Center
**Contents Research Division, ETRI

요 약

본 논문에서는 멀티미디어 데이터에 대한 내용기반 인식 기법을 이용하여 인터넷에 불법으로 배포되어 있는 콘텐츠를 추적하는 기법을 소개한다. 내용기반 인식 기법은 콘텐츠의 원신호에서 내용기반 해쉬나 혹은 축약된 형태의 특징벡터를 추출하여 콘텐츠를 인식하는 기술로 저작권보호 분야에서 불법 저작물을 필터링하는데 많이 활용되고 있다. 불법 콘텐츠 추적시스템은 인터넷에서 광범위하게 유포되어 있는 저작물을 검색하여 그 내용을 기반으로 인식하여 불법 여부를 판단한 후 삭제메일이나 재전송 중지 등의 후속 조치를 자동으로 수행하는 저작권보호 시스템이다. 본 논문에서는 오디오, 비디오, 어문, 게임 콘텐츠에 대해 내용을 기반으로 인식을 수행하고 불법 여부를 판단하여 재전송 중지 조치를 취하는 능동적 저작물 추적 시스템을 제안한다. 제안된 시스템에서는 검색모듈에 의해 수집된 다양한 저작물에 대해 저작물별 독립적으로 인식 기능을 수행하는 기능을 제공한다.

1. 서론

최근 디지털 기술과 인터넷 환경의 급속한 발전으로 인해 디지털 콘텐츠의 제작과 판매가 활발해지고 있다. 그러나 손실 없이 대량 복제가 가능한 디지털 콘텐츠의 특성과 사용자들의 유료 콘텐츠 사용에 대한 인식 부족으로 디지털 콘텐츠의 지적재산권 침해가 빈번히 발생하여, 콘텐츠 산업 발전을 저해하는 심각한 문제로 대두되고 있다.

디지털 콘텐츠에 대한 저작권을 보호하기 위한 주요 기술로는 DRM(Digital Rights Management), 워터마킹(watermarking), 디지털핑거프린팅(digital fingerprinting), 콘텐츠인식 기술 등이 있다. 일반적인 DRM 시스템은 사용자가 콘텐츠를 요청하면 암호화된 콘텐츠를 보내주고, 실행 시 지불 시스템과 연결하여 지불이 완료되면 암호를 풀 수 있는 키를 다시 보내주어 사용자가 콘텐츠를 사용할 수 있게 하는 구조로 되어 있다. 암호화에 근간을 두고 있는DRM은 콘텐츠에 대하여 지불을 하지 않은 일반 사용자들의 접근자체를 봉쇄하지만, 암호화가 풀린 이후의 상황에서는 DRM기술만으로는 콘텐츠의 저작권보호가 사실상 어렵게 된다. 실제로 음악파일이나 동영상 같은 멀티미디어 콘텐츠의 경우 복호화 된 상태나 혹은 콘텐츠의 재생 시 다시 캡춰(capture)되어 암호화되지 않고 배포되는 경우가 대부분이다.

최근에는 이러한 불법 배포 콘텐츠의 재전송을 방지하기 위해 불법 콘텐츠 추적 기술이 많이 활용되고 있다. 불법 콘텐츠 추적 기술은 웹크롤러(web crawler)나 검색로봇 등으로 콘텐츠를 검색 및 다운로드, 게시증거 수집을 수행하고 다운로드 한 콘텐츠에 대해 워터마크나 핑거프린트를 추출하거나 특징기반 인식을 통해 불법 여부를 판단한 후 삭제메일이나 재전송 금지 등의 후속 조치 등을 자동으로 수행하는 저작권보호 기술이다.

저작권정보를 삽입하는 워터마킹이나 구매자 정보가 삽입되는 핑거프린팅의 경우 콘텐츠가 배포되기 전에 삽입되어야 하기 때문에 현재로써는 내용기반 인식 기술을 이용한 불법 콘텐츠 추적이 더 많이 활용되고 있다.

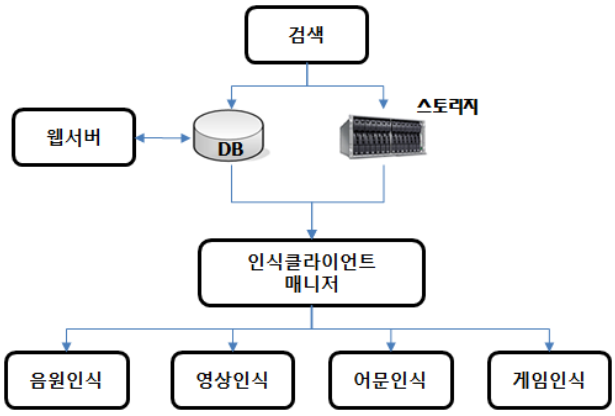
본 논문에서는 오디오, 영상, 어문, 게임 콘텐츠에 대한 내용 기반 인식 기술을 이용한 불법 콘텐츠 추적 시스템을 제안한다. 추적 시스템은 다양한 콘텐츠에 대한 검색 및 인식 기술이 통합되어야 하기 때문에 효율적인 운영을 위해서 각 시스템에 대한 연동 설계가 중요하다. 본 논문의 구성은 다음과 같다. 2장에서는 불법 콘텐츠 추적 시스템의 개요 및 구성에 대해 설명한다. 다음으로는 다양한 콘텐츠에 대한 인식 방법에 대해 고찰하고 결론 및 향후 연구 방향을 제시한다.

2. 불법 콘텐츠 추적 시스템

최근 많이 활용되고 있는 불법 콘텐츠 추적시스템은 불법 콘텐츠가 주로 유통되고 있는 웹하드나 P2P, 블로그, 카페 등을 자동으로 검색하여 불법 콘텐츠가 존재할 경우

그 게시정보를 수집하고 삭제요청 메일 등의 저작권 보호 조치를 자동으로 수행하는 종합시스템이다. 콘텐츠의 유통 상황을 실시간으로 파악할 수 있고 또한 불법 유통에 즉각적으로 대응할 수 있어 최근 많이 활용되고 있다.

본 논문에서 제안하는 불법 콘텐츠 추적 시스템은 크게 콘텐츠 검색모듈, 인식모듈, DB와 저장소, 삭제메일 등의 후속조치 및 각종 통계 자료를 제공하는 웹서버로 구성된다. (그림 1)은 본 논문에서 제안하는 시스템의 전체 구성도이다.



(그림 1) 불법 콘텐츠 추적 시스템 구성

검색시스템은 웹하드나 P2P등의 OSP[1]로부터 콘텐츠를 검색하여 게시정보를 수집하고 콘텐츠를 다운로드하여 게시정보는 DB에, 콘텐츠는 스토리지에 저장한다. 인식클라이언트매니저는 다운로드된 콘텐츠를 검증하고 특징점을 추출하여 각 인식시스템에 전달하고 그 결과를 DB에 저장한다.

제안한 시스템에서는 검색시스템과 인식시스템이 비동기적으로 동작한다. 기존의 불법 콘텐츠 추적시스템에서는 검색모듈이 콘텐츠를 다운로드 한 후 특징점을 추출하여 인식하는 기능까지 수행하였는데, 이런 경우 검색시스템에 부하가 많이 걸려 검색의 효율성이 낮았다. 또한 검색시스템이 다운일 경우에는 인식시스템도 활용할 수 없는 상황이 된다. 이를 보완하기 위해서 검색과 인식을 분리하고 특징점 추출 기능을 전담하는 인식클라이언트 매니저를 배치함으로써 추적 시스템 전체의 효율성을 증대시켰다.

2.1 검색시스템

검색모듈은 불법으로 유포되어진 것으로 추정되는 저작물을 검색하기 위해 자동로그인, 게시물이동, 업로드명 등의 게시정보 수집, 게시화면 캡춰 및 저장, 인식을 위해 저작물을 자동으로 다운로드 하는 기능을 포함한다. 한국의 인터넷 환경에서는 다운로드 속도가 우수하기 때문에 웹하드나 P2P의 특수유형 OSP에 고용량의 저작물이 불법으로 유통되고 있는 실정이다. 제안하는 검색 모듈은 이런 특수유형의 OSP에서 저작물을 자동으로 검색할 수 있는 기능을 제공한다.

웹하드나 P2P의 특수유형 OSP는 회원가입이 필수이고 로그인 후 유료로 저작물을 다운로드해야 하기 때문에 일반 검색 크롤러(crawler)를 사용한 검색만으로는 부족하다. 또한 내용기반 인식을 통해 정확한 불법 저작물을 추적하기 위해서는 게시번호나 제목, 게시자, 게시캡춰화면 등의 게시정보를 수집하고 검색된 저작물을 다운로드 해야 한다. 더욱이 웹하드나 P2P 들은 그 서비스 형태를 자주 바꾸기 때문에 이런 변경된 인터페이스에도 유연하게 수정되어 검색이 가능하도록 해야 한다. 따라서 일반 크롤러와는 달리 이런 기능을 지원하도록 개발하여야 한다.

제안된 검색모듈은 개발 및 수정의 용이성을 지원하기 위해 자동화 스크립트 언어(automation script language)를 사용해 개발되었다[1]. 자동화 스크립트 언어는 간단한 반복 작업을 쉽게 구현할 수 있도록 고안된 기초적인 프로그래밍 언어로 비주얼베이직 스크립트(VB Script)언어와 유사하다. 검색을 위한 제어로직은 스크립트 언어로 개발되었으며, 자동로그인, 게시정보 추출, 콘텐츠 다운로드 등의 OSP를 제어하는 기능은 C/C++로 개발하여 라이브러리화 하였다. 스크립트에서는 개발된 라이브러리를 COM 형태로 등록하여 사용하였다.

2.2. 하이브리드 인식 시스템

국내의 불법 콘텐츠는 주로 웹하드나 P2P를 통해 90% 이상이 유통되고 있으며, 다운로드 속도가 빨라 주로 대용량, 고화질의 콘텐츠가 다수이다. 또한 같은 내용의 콘텐츠라도 여러 사람의 업로더(uploader)에 의해 다양한 파일 포맷으로 유통되며, 각각의 업로더들은 같은 파일을 일정 주기로 반복적으로 업로드 하는 유통 행태를 보인다. 불법 콘텐츠 추적시스템이 효율성을 갖기 위해서는 이러한 다양한 포맷의 대용량 콘텐츠를 빠르게 인식하여야 한다. 본 논문에서 제안하는 하이브리드 인식 시스템은 이러한 유통 환경에 따라 기존의 대표적인 파일 인식방법인 해쉬(hash) 기반 인식과 내용기반(content-based) 인식 방법을 서로 보완하여 사용한다. 해쉬기반 인식은 반복되어 업로드 되는 콘텐츠에 대해 빠른 인식 시간을 제공할 수 있고, 내용기반 인식은 다양한 파일포맷에 대해 높은 인식률을 제공할 수 있다.

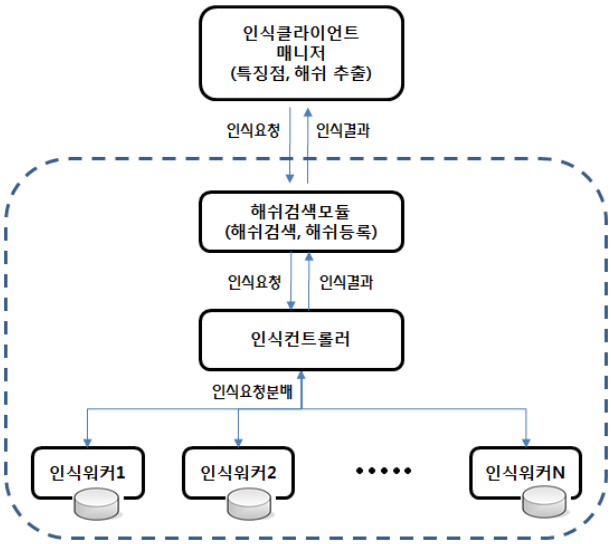
제안한 시스템에서는 해쉬 충돌의 위험이 적은 SHA1 알고리즘을 사용하여 파일의 해쉬를 추출하였다. 해쉬기반 인식은 검색시간이 매우 빠르다는 장점이 있으나 아주 작은 부분의 변형이 있을 경우 인식이 불가능하다.

내용기반 인식 시스템은 검색시스템으로부터 다운로드한 저작물에 대해 그 저작물의 내용을 기반으로 저작물을 인식(identification)하는 시스템이다. 내용기반 인식 방법은 콘텐츠의 고유 성분을 특징벡터로 추출하여 DB화 한 시스템이다. 임의의 콘텐츠에 대해 인식할 경우 같은 방법으로 특징을 추출하여 이미 구축되어진 DB에서 비슷한

1) OSP(Online Service Provider) : 인터넷 상에서의 콘텐츠 서비스 공급자로 웹하드, P2P 등은 특수유형 OSP로 분류된다.

특징벡터를 검색하여 인식한다. 내용기반 인식은 해쉬(hash) 인식 방법에 비해 여러 가지 파일 포맷이나 왜곡에 대해 강인한 장점이 있어 다양한 파일 형태의 콘텐츠가 존재하는 불법 콘텐츠 유통 환경에서 기존의 해쉬기반 방법보다 우수한 인식률을 보이지만 검색시간이 길다는 단점이 있다.

본 논문에서는 이런 장단점을 고려하여 먼저, 검색된 저작물을 해쉬기반으로 인식하고 인식이 실패하면 내용기반으로 검색한다. 내용기반 검색이 성공하였을 경우, 추출된 해쉬값을 다시 등록하여 추후에 같은 파일이 검색되었을 경우에는 해쉬기반으로 인식되도록 하였다, 이로 인해 인식시스템 전반적으로 해쉬 또는 특징점 한 가지만 사용하는 단일 시스템에 비해 검색시간과 인식률을 개선하였다.



(그림 2) 인식 시스템 구성

(그림 2)는 제안된 인식시스템의 시스템 구성도이다. 인식시스템에는 해쉬검색을 수행하는 해쉬검색모듈과 내용기반 인식을 수행하는 컨트롤러 및 다수의 워커들이 존재한다. 내용기반 인식의 검색 시간을 줄이기 위해 실제 검색을 수행하는 다수의 워커를 배치하고 각 워커들에 추출요청을 분해하고 결과를 머지(merge)하는 컨트롤러를 둔다. 각 워커는 고유의 특징점 DB를 갖는다.

제안한 시스템에서는 추적하고자 하는 저작물의 해쉬값과 특징점을 추출하여 미리 등록시켜야 한다. 이 때 각 저작물은 고유 코드체계를 사용한 ID를 갖는다. 다음으로는 각 저작물에 대한 특징벡터에 대해 간략히 소개한다.

2.2.1 음원저작물인식

일반적으로 음원저작물을 인식하기 위한 특징벡터로는 Mel-Frequence Cepstrum Coefficients(MFCC), Spectral Flatness Measure(SFM) 등의 주파수 영역 특성을 사용한다. 본 시스템에서도 주파수영역에서의 정규화된 서브밴드 중심점(Normalized Spectral Subband Centroids:SSC)을 특징으로 사용한다[4][9]. 많은 비교 테스트에 의해 SSC 특징은 MFCC나 tonality 같은 다른 특징들과 같이 오디오 인식 분야에서 널리 사용되고 있다.

2.2.2 영상저작물인식

영상저작물을 인식하는 방법은 여러 가지가 있으나 크게 오디오신호를 이용한 방법과 비디오신호를 이용한 방법으로 나눌 수 있다. 오디오신호를 이용한 방법은 위에 설명한 음원저작물인식방법과 같으며 비디오신호를 이용한 방식에는 이미지인식 방법을 기반으로 다시 여러 가지가 존재한다[6][7]. 본 논문에서는 비디오신호에서의 장면전환(scene change)속성[8]과 계층적 대칭차(HSD:hierarchical symmetric difference)속성 및 오디오신호를 계층적으로 구성하여 영상인식시스템을 구성하였다.

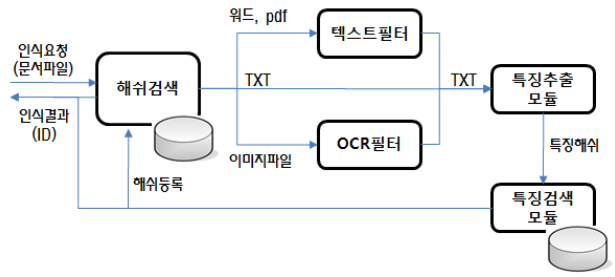
2.2.3 어문저작물인식

일반적으로 어문을 인식하기 위한 특징으로는 유사문서 검색에 사용되는 방법의 일종으로 어문에 존재하는 명사들의 상대적 위치 정보나 빈도수 등이 많이 사용된다. 하지만 이런 일반적인 유사문서 검색 방법을 이용할 경우 특징을 추출하는 시간과 대용량의 어문DB에서 문서들 각각을 검색하는 시간이 많이 소요된다. 불법 콘텐츠 추적시스템에서는 빠른 인식시간이 요구되기 때문에 이러한 방법은 부적절하다. 또한 유사문서 검색에서는 검색하고자 하는 문서에 대해, DB에 존재하는 유사한 문서들을 그 유사도에 따라 추천해 주는 방식이지만 본 시스템에서는 인식하고자 하는 쿼리 문서의 ID만이 요구되므로 그 차이가 있다. 즉, 제안된 시스템에서는 인식하고자 하는 문서에 대해 DB에 존재여부와, 존재한다면 그 ID만을 빠른 시간 안에 출력해야 한다. 이를 위해 본 시스템에서는 특정길이 이상의 명사가 존재하는 문장(혹은 문단)의 해쉬값을 특징으로 사용한다. 따라서 임의의 어문저작물에 대한 특징벡터는 다수의 해쉬값이며 이는 어문에 따라 각각 다르다. 추출하는 명사의 특정길이는 실험적으로 정하였다.

추가적으로 어문저작물은 저작권이 존재하는 책 등의 어문이 텍스트나 스캔본의 이미지 형태로 불법 유통되는 경우가 대부분이다. 스캔이외에도 간혹 워드파일이나 pdf 형태로 유통되는 경우가 많은데 어문저작물 인식을 위해서는 일단 텍스트 형태로 추출해야 한다. 제안된 어문인식 시스템에서는 이러한 전처리를 위하여 다양한 워드 문서에서 텍스트를 추출할 수 있는 텍스트필터와 스캔된 이미지 문서에서 텍스트를 추출할 수 있는 OCR필터를 연동하였다. 어문인식시스템에 대한 구조는 (그림 3)과 같다.

2.2.4 게임저작물인식

불법으로 유통되는 게임저작물은 대부분 설치패키지 형태로 압축파일에 묶여져 있거나 iso, nrg 등의 이미지파일 형태가 많다. 기본적인 게임저작물 인식 방법은 일반 소프트웨어를 인식하는 방법과 같이 해쉬기반 인식을 활용한다. 즉, 게시된 불법 파일에 대한 해쉬값을 추출하여 인식



(그림 3) 어문인식시스템 구성

시스템에 등록하는 방식이다. 하지만 이 방법은 게시된 파일의 압축포맷이나 이미지 포맷이 변경되면 인식이 불가능하여 다시 재등록해야 하는 단점이 있다.

제안된 시스템에서는 게시된 파일에서 압축이나 이미지 포맷을 해제한 후 설치패키지의 특정파일에 대해서만 해쉬값을 추출하는 방법을 사용한다. 특정파일은 exe, dll 혹은 그 게임패키지에만 존재하는 특정 파일들이다. 따라서 하나의 게임패키지에 여러 개의 해쉬값이 특징으로 존재할 수 있다. 임의의 패키지에 대해 인식을 수행할 때에는 exe, dll, 혹은 특정 확장자를 갖는 파일들에서 해쉬값을 추출한 후 인식시스템의 해쉬DB에서 검색을 수행하여 매핑되는 저작물ID를 인식결과로 리턴한다.

3. 구현

제안된 시스템은 오라클DB와 MS윈도우즈 서버를 이용해 구현되었다. 음원인식시스템은 3분 길이의 가요 기준으로 10만곡에 대한 해쉬 및 특징DB를 구축하였고 워커1개당 2만곡 정도의 특징DB를 분리하여 저장하였다. 영상인식시스템은 1시간 길이 6천편의 영상에 대해 해쉬와 특징을 추출하여 DB를 구축하였다. 음원인식시스템과 마찬가지로 워커1개당 2천편의 영화를 구축하여 총3대의 워커로 구성하였다. 어문인식시스템은 5,000편이상의 어문을, 게임인식시스템은 500개 이상의 게임패키지에 대한 특징DB를 구축하였다.

<표 1> 인식방법에 따른 평균인식 시간(초)

콘텐츠 / 인식방법	음원	영상	어문
내용기반인식	1.2	5.6	1.2
하이브리드인식	0.5	2.2	0.8

<표 1>은 제안한 하이브리드 인식방법과 내용기반 인식 방법에 대해 콘텐츠별 평균인식시간을 나타낸 것이다. 인식시간을 계산한 테스트셋은 검색시스템이 실제의 웹하드에서 수집한 저작물을 이용하였고, 크기는 음원이 1,000곡, 영상이 1시간 기준 500편, 어문은 A4 2페이지 분량 1,000편이다. 게임저작물 인식은 기본적으로 해쉬를 사용하기 때문에 실험에서 제외하였다.

제안한 하이브리드 인식 시스템의 평균인식속도가 기존의 내용기반 인식 방법보다 40 - 50% 정도 개선되었음을 알 수 있다. 이는 반복되어 업로드 되는 저작물이 많아 상당부분 해쉬값을 통해 인식이 되었기 때문이다.

4. 결론

본 논문에서는 웹하드나 P2P에서 대량으로 유통되고 있는 콘텐츠를 검색하고 인식하는 불법 콘텐츠 추적 시스템의 설계와 구현에 대해 고찰하였다. 현재 국내의 웹하드나 P2P에서는 그 다운로드 속도가 우수하여 대용량의 파일들이 반복적으로 유통되고 있는 실정이다. 이러한 콘텐츠를 효율적으로 검색하고 인식하기 위하여 본 논문에서는 자동화 스크립트 언어를 이용한 검색시스템과 해쉬 및 내용 기반 특징을 동시에 활용하는 하이브리드 인식시스템을 이용하는 불법 콘텐츠 추적 시스템을 제안하고 구현하였다. 제안된 시스템은 빠른 인식 시간과 높은 인식율을 제공하여 다량의 불법 콘텐츠를 검색하는 데 효율적인 결과를 보이고 있다.

참고문헌

[1] Autoit, http://www.autoscript.com

[2] 김원겸, 이선화, 장호욱, “불법 복제 콘텐츠 추적을 위한 핑거프린팅 기술 동향,” *전자통신동향분석 제18권 제4호*, pp.82-94, 2003

[3] 정혜원, 이준석, 서영호, “불법콘텐츠 추적 기술 연구 동향,” *전자통신동향분석 제20권 제4호*, pp.120-128, 2005

[4] Jin S. Seo, Minho Jin, Sunil Lee, Dalwon Jang, Seungjae Lee, and C. D. Yoo, "Audio Fingerprinting Based on Normalized Spectral Subband Moments," *IEEE Signal Processing Letters, Vol. 13, No. 4*, 2006

[5] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of algorithms for audio fingerprinting," *in Proc. IEEE Workshop Multimedia Signal Processing*, pp. 169-173, 2002

[6] A. Hampapur, K.H. Hyun, and R. Bolle, "Comparison of Sequence Matching Techniques for Video Copy Detection," *Proc. Storage and Retrieval for Media Databases*, 2002

[7] S.I.Lee and C. D. Yoo, "Robust Video Fingerprinting for Content-Based Video Identification," *IEEE Trans. on Circults and systems for Video Technology, Vol. 18, No. 7*, pp.983-988, 2008

[8] N. Shivakumar, "Detecting Digital Copyright Violation on the Internet," doctoral dissertation, Stanford Univ. 1999

[9] K.K. Paliwal, "Spectral subband centroid features for speech recognition," *in Proc. IEEE ICASSP*, pp.617-620, 1998