(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2008년05월08일
(11) 등록번호 10-0828166
(24) 등록일자 2008년04월30일

(51) Int. Cl.
H04N 5/93 (2006.01) *H04N 7/24* (2006.01)
(21) 출원번호 10-2007-0057478
(22) 출원일자 2007년06월12일
심사청구일자 2007년06월12일
(56) 선행기술조사문헌
KR 1020070016008 A
(뒷면에 계속)

(73) 특허권자
고려대학교 산학협력단
서울 성북구 안암동5가1 고려대학교 내
한국방송공사
서울 영등포구 여의도동 18번지
(72) 발명자
고한석
서울 용산구 서빙고동 신동아아파트 16동 703호
정석영
서울 서대문구 연희1동 444-16 B101호
(뒷면에 계속)
(74) 대리인
현종철

전체 청구항 수 : 총 16 항 심사관 : 유병철

(54) 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출방법, 메타데이터를 이용한 동영상 탐색 방법 및 이를기록한 기록매체

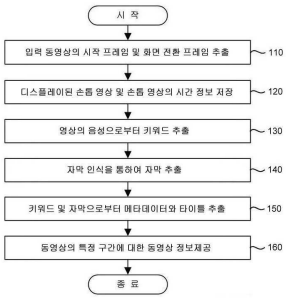
(57) 요 약

동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법, 메타데이터를 이용한 동영상 탐색 방법 및 이를 기록한 기록매체가 개시된다.

본 발명에 따른 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법은, 상기 메타데이터를 포함하는 동영상을 입력하고, 상기 입력된 동영상의 시작 프레임 및 화면 전환 프레임을 추출하는 단계, 상기 추출된 시작 프레임 및 화면 전환 프레임을 손톱 영상으로 디스플레이하고, 상기 디스플레이된 손톱 영상 및 상기 손톱 영상의 시간 정보를 저장하는 단계, 상기 입력된 동영상에 포함된 음성의 음소에 따라 화자의 음성을 인식하고, 상기 인식된 음성 데이터를 문자 데이터로 변환하고, 상기 변환된 문자데이터로부터 키워드를 추출하는 단계, 상기 입력된 동영상으로부터 자막을 검출하고, 상기 검출된 자막으로부터 자막 인식을 통하여 자막을 추출하는 단계, 사용자가 상기 디스플레이된 동영상의 시작 프레임 및 화면 전환 프레임 중 시작 샷과 끝 샷을 지정하면 상기 지정된 시작 샷, 끝 샷 및 상기 시작 샷과 끝 샷 사이의 손톱 영상에 포함된 키워드 및 자막으로부터 메타데이터와 타이틀을 추출하는 단계, 및 상기 추출된 메타데이터, 상기 시작 샷의 시간 정보, 상기 끝 샷의 시간 정보 및 상기 타이틀을 표시하는 단계를 포함한다.

본 발명에 의하면, 음성 인식 결과와 동영상 자체의 개방형 자막 인식 결과인 방송 콘텐츠에 포함되어 있는 메타데이터 정보를 자동으로 추출하고, 새로운 방송 자료들의 수작업에 의한 작업시간을 단축할 수 있으며, 과거의 방대한 양의 방송 데이터들에 대한 콘텐츠 관리 및 색인 작업을 자동으로 실행하여 방송 제작자들에게 양질의 콘텐츠를 제작할 수 있게 함으로써, 과거의 자료를 이용한 콘텐츠 제작에 소요되는 시간과 비용을 절약할 수 있도록 할 수 있을 뿐만 아니라 현재 광범위하게 적용되고 있는 인터넷 검색 사이트와 연동하여 실시간으로 제작되는 콘텐츠에 대한 멀티미디어 자료들의 검색을 가능하게 하고, 과거의 자료들에 대한 열람을 용이하게 함으로써, 사용자의 콘텐츠 사용의 편의성을 제공하는 효과가 있다.

대표도 - 도1



(72) 발명자

박수인

서울 강남구 도곡동 516-1 현대맨션 205호

윤종성

서울 성북구 종암1동 24-20번지 301호

김동준

서울 관악구 봉천6동 1692번지 8호

박성춘

서울 영등포구 당산동 376 강변래미안아파트 304동 1602호

하명환

서울 영등포구 당산동 121-19 계명아파트 602호

김건희

서울 도봉구 방학2동 604-26

(56) 선행기술조사문헌

KR 100679049 B1

JP2002199348 A

KR1020060087144 A

KR1020060089922 A

KR1020070042000 A

특허청구의 범위

청구항 1

메타데이터를 포함하는 동영상을 입력하고, 상기 입력된 동영상의 시작 프레임 및 화면 전환 프레임을 추출하는 단계;

상기 추출된 시작 프레임 및 화면 전환 프레임을 손톱 영상으로 디스플레이하고, 상기 디스플레이된 손톱 영상 및 상기 손톱 영상의 시간 정보를 저장하는 단계;

상기 입력된 동영상에 포함된 음성의 음소에 따라 화자의 음성을 인식하고, 상기 인식된 음성 데이터를 문자 데이터로 변환하고, 상기 변환된 문자데이터로부터 키워드를 추출하는 단계;

상기 입력된 동영상으로부터 자막 인식을 통하여 자막을 추출하는 단계;

사용자가 상기 디스플레이된 동영상의 시작 프레임 및 화면 전환 프레임 중 시작 샷과 끝 샷을 지정하면 상기 지정된 시작 샷, 끝 샷 및 상기 시작 샷과 끝 샷 사이의 손톱 영상에 포함된 키워드 및 자막으로부터 메타데이터와 타이틀을 추출하는 단계; 및

상기 추출된 메타데이터, 상기 시작 샷의 시간 정보, 상기 끝 샷의 시간 정보 및 상기 타이틀을 표시하는 단계를 포함하는 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법.

청구항 2

제 1 항에 있어서,

상기 화자의 음성은,

각각의 문장 단위로 인식되는 것을 특징으로 하는 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법.

청구항 3

제 1 항에 있어서,

상기 메타데이터와 타이틀을 추출하는 단계는,

상기 키워드와 자막이 일치하는 단어가 존재할 경우 상기 일치하는 단어에 가중치를 부여하고, 상기 가중치가 부여된 단어를 메타데이터로 우선적으로 추출하는 단계를 포함하는 것을 특징으로 하는 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법.

청구항 4

제 1 항에 있어서,

상기 타이틀은

상기 시작 샷의 자막으로 설정되는 것을 특징으로 하는 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법.

청구항 5

제 1 항에 있어서,

상기 자막 인식을 통하여 자막을 추출하는 단계는,

상기 입력된 동영상의 프레임으로부터 자막 후보 영역을 검출하고, 상기 검출된 자막 후보 영역을 검증하는 단계; 및

상기 검증된 자막 후보 영역으로부터 자막 인식을 수행하는 단계를 포함하는 것을 특징으로 하는 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법.

청구항 6

제 5 항에 있어서,

상기 자막 후보 영역은,

상기 입력된 동영상에 가우시안 필러링 후 컬러도메인에서의 컬러 미분 연산자에 따른 미분연산값을 이용하여 검출된 것을 특징으로 하는 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법.

청구항 7

제 5 항에 있어서,

상기 자막 인식을 수행하는 단계는,

상기 자막 후보 영역을 제1차 동적 문턱값 및 제2차 동적 문턱값을 이용하여 이진화를 수행하는 단계; 및

상기 이진화된 자막 후보 영역에 존재하는 자막을 글자 단위 분리 및 자소 분리를 수행한 후, 각각 자소 분리된 자소를 인식하고, 상기 인식된 자소에 따라 글자 단위로 자막을 인식하는 단계를 포함하는 것을 특징으로 하는 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법.

청구항 8

제 1 항에 있어서,

상기 메타데이터와 타이틀을 추출하는 단계는,

상기 자막과 키워드를 융합하여 메타데이터를 추출하되, 상기 자막이 문장형으로 추출될 경우 상기 자막 중 명사인 단어를 자막으로 추출하는 하는 단계를 포함하는 것을 특징으로 하는 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법.

청구항 9

제 1 항에 있어서,

상기 변환된 문자데이터로부터 키워드를 추출하는 단계는,

단어에 대한 중요도를 판단하는 핵심어 후보 단어를 포함하는 데이터베이스를 구성하는 단계; 및

상기 동영상에서의 상기 변환된 문자 데이터의 빈도 수와 상기 구성된 데이터베이스에서의 상기 변환된 문자 데이터의 빈도 수에 따라 상기 키워드를 추출하는 단계를 포함하는 것을 특징으로 하는 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법.

청구항 10

제 1 항에 있어서,

상기 동영상은 MPEG-2 표준에 따른 동영상을 포함하는 것을 특징으로 하는 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법.

청구항 11

동영상을 입력하고, 상기 입력된 동영상의 시작 프레임 및 화면 전환 프레임을 추출하는 하여 상기 시작 프레임 및 화면 전환 프레임을 손톱 영상으로 디스플레이하고, 상기 디스플레이된 손톱 영상 및 상기 손톱 영상의 시간 정보를 저장하는 단계;

상기 입력된 동영상에 포함된 음성의 음소에 따라 화자의 음성을 인식하고, 상기 인식된 음성 데이터를 문자 데이터로 변환하고, 상기 변환된 문자데이터로부터 키워드를 추출하는 단계;

상기 입력된 동영상으로부터 자막을 검출하고, 상기 검출된 자막으로부터 자막 인식을 통하여 자막을 추출하는 단계;

상기 디스플레이된 동영상의 시작 프레임 및 화면 전환 프레임 중 시작 샷과 끝 샷을 지정하여 상기 지정된 시작 샷과 끝 샷 사이의 지정 동영상에 포함된 키워드 및 자막으로부터 메타데이터와 타이틀을 추출하고, 상기 추출된 메타데이터, 상기 시작 샷의 시간 정보, 상기 끝 샷의 시간 정보 및 상기 타이틀을 포함하는 동영상 데이

터를 디스플레이하는 단계;

상기 디스플레이된 동영상 데이터를 XML 문서로 저장하는 단계; 및

특정 사용자가 탐색하고자 하는 검색어를 입력하면, 상기 XML 문서로 저장된 동영상의 저장 시간, 타이틀, 재생 길이 및 메타데이터를 출력하는 단계를 포함하는 메타데이터를 이용한 동영상 탐색 방법.

청구항 12

제 11 항에 있어서,

상기 자막 인식을 통하여 자막을 추출하는 단계는,

상기 동영상의 프레임 중 특정 영역을 자막 후보 영역으로 설정하는 단계;

상기 설정된 자막 후보 영역에 대하여 영상 프레임에 포함된 문자를 검출하는 단계; 및

검출된 문자로부터 자막을 추출하고, 상기 추출된 자막으로부터 명사인 단어를 추출하는 단계를 포함하는 것을 특징으로 하는 메타데이터를 이용한 동영상 탐색 방법.

청구항 13

제 11 항에 있어서,

상기 변환된 문자데이터로부터 키워드를 추출하는 단계는,

단어에 대한 중요도를 판단하는 핵심어 후보 단어를 포함하는 데이터베이스를 구성하는 단계; 및

상기 동영상에서의 상기 변환된 문자 데이터의 빈도 수와 상기 구성된 데이터베이스에서의 상기 변환된 문자 데이터의 빈도 수에 따라 상기 키워드를 추출하는 단계를 포함하는 것을 특징으로 하는 메타데이터를 이용한 동영상 탐색 방법.

청구항 14

제 11 항에 있어서,

상기 XML 문서는 MPEG-7의 표준에 따른 문서를 포함하는 것을 특징으로 하는 메타데이터를 이용한 동영상 탐색 방법.

청구항 15

제 11 항에 있어서,

상기 타이틀은,

상기 시작 샷의 자막으로 설정되는 것을 특징으로 하는 메타데이터를 이용한 동영상 탐색 방법.

청구항 16

제 1 항 내지 제 15 항 중 어느 한 항의 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

명 세 서

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

<10> 본 발명은 데이터 추출에 관한 것으로서, 더욱 상세하게는 일반 동영상 혹은 실시간으로 방송되는 영상을 정해진 구간 혹은 주제별 프레임 단위로부터의 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법, 메타데이터를 이용한 동영상 탐색 방법 및 이를 기록한 기록매체에 관한 것이다.

<11> 현대의 정보의 홍수 속에서 가독성과 속도성이 높은 텍스트 위주의 자료들이 멀티미디어 데이터로 변화되어 가는 추세에 있다. 이에 따라서 대용량의 멀티미디어 데이터의 처리를 위하여 많은 시간과 노력이 필요하게 되었다.

<12> 더욱이, 눈과 귀가 동시에 필요한 멀티미디어 정보 처리에 부담을 느끼는 사용자들이 증가하면서 적절한 요약 정보의 필요성이 대두 되었으며, 이를 위한 요약 정보 필요성이 대두 되었으며, 이를 위한 요약 정보 제공 기술들이 제안되었다.

<13> 예를 들어, 디지털 비디오 콘텐츠(Contents)는 전체를 재생하여 시청할 수도 있지만 경우에 따라서 비디오 전체를 보지 않고도 그 내용을 이해할 수 있도록 요약된 형태의 하이라이트 동영상이 프로그램 공급자에 의하거나 혹은 사용자 시스템 자체에서 자동 생성하는 형태로 제공되기도 한다.

<14> 하이라이트 동영상은 저장된 비디오 중에서 부분만을 재생하기 위한 것으로, 해당 비디오 스트림을 대표하는 특징이 있다. 하이라이트 동영상은 비디오 스트림(stream)의 특정 구간을 따라 저장 또는 재생하기 위한 것으로 제한된 시간 동안 디지털 비디오 녹화기에 저장된 여러 개의 비디오 중에서 하나를 사용자가 선택하여 보기를 원할 때, 사용자는 각 비디오 스트림의 하이라이트 동영상만을 재생하여 원하는 비디오 내용을 검색하는데 걸리는 시간을 절약할 수 있다. 또한 하이라이트는 저장된 비디오 스트림의 요약 정보 이외에도 디지털 비디오 저장 장치에서 사용자가 녹화할 비디오를 선택하는데 필요한 프로그램 가이드 장치 등에서 사용할 수 있는 프리뷰(preview) 기능을 제공할 수 있다.

<15> 그러나, 이는 사용자에게 비디오의 내용을 대표할 만한 의미가 있는 부분을 따로 추출해 내야 하므로 하이라이트를 생성할 구간의 설정은 상당히 까다로운 작업이다.

<16> 종래의 대한민국 특허공보 제0404322호에 의하면, 이는 폐쇄 자막과 함께 제공되는 뉴스 비디오를 뉴스 기사 단위로 분할하고 분할된 뉴스 기사 단위별로 뉴스 비디오를 압축하는 방법으로서, 폐쇄 자막 데이터를 기반으로 음성인식기법을 사용한다. 상기 특허 발명에서 사용된 자막방송 기반의 폐쇄 자막은 음성인식기의 인식률 저하를 방지하기 위한 방안으로 사용되었으나, 이를 위해서는 자막방송용 속기록이 1차적으로 준비되어야 하기 때문에 일반적으로 녹화되거나 자막방송이 생성되거나 저장되지 않은 방송 동영상에 대해서는 바로 적용하기에는 문제점이 있다.

<17> 또한, 대한민국 특허공보 제0540735호에 의하면, 이는 프레임별 자막 추출 및 인식을 통하여 각 프레임 자막의 인식 결과를 그 프레임(frame)과 인덱싱(Indexing) 시켜 영상 프레임을 검색할 수 있게 하는 방법이다. 이는 자막이 포함된 영상의 위치를 찾도록 자막과 그 자막이 포함된 영상 프레임을 연결한 것으로 영상에서 자막을 검출하여 인식하지만 그 대상을 하나의 영상 프레임의 인덱싱에 국한하였고, 그로 인해 전체적인 기사 내용과 관련 없는 자막이 포함된 경우일지라도 자막이 일치하면 검색의 대상에 포함되는 경우가 발생할 수 있다.

<18> 그리고, 1999년 IEEE 학회 논문인 "Automated generation of News content hierarchy by integrating audio, video and text information, 3025~3028p, Qian Huang외 4인"에 의하면, 상기 논문에서는 동영상에서 폐쇄 자막 검출, 얼굴 인식, 앵커샷 검출, 음성 인식을 통하여 뉴스에 대한 검색 정보를 제공하는 방법을 제안하고 있으나, 이는 폐쇄 자막에 한정되었다.

<19> 상기와 같이 종래의 요약 정보 제공 방법은, 비디오의 내용을 대표할 만한 의미가 있는 부분을 따로 추출해 내야 하므로 하이라이트를 생성할 구간의 설정 영상에서 자막을 검출하여 인식하지만 그 대상을 하나의 영상 프레임의 인덱싱에 국한되는 문제가 있고, 전체적인 기사 내용과 관련 없는 자막이 포함된 경우일지라도 자막이 일치하면 검색의 대상에 포함되는 경우가 발생하여 잘못된 요약 정보를 제공할 수 있으며, 폐쇄 자막을 통한 요약 정보 제공 방법은 자막방송용 속기록이 1차적으로 준비되어야 하기 때문에 일반적으로 녹화되거나 자막방송이 생성되거나 저장되지 않은 방송 동영상에 대해서는 바로 적용하기에는 문제점이 있다.

발명이 이루고자 하는 기술적 과제

<20> 따라서, 본 발명이 이루고자 하는 첫 번째 기술적 과제는, 음성 인식의 결과뿐만 아니라 폐쇄 자막의 유무에 상관없이 동영상의 영상정보로부터 자막을 인식하여 음성 인식과 함께 핵심어를 추출할 수 있는 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법을 제공하는 것이다.

<21> 또한, 본 발명이 이루고자 하는 두 번째 기술적 과제는, 상기 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법을 적용한 메타데이터를 이용한 동영상 탐색 방법을 제공하는 것이다.

발명의 구성 및 작용

<22> 상기 첫 번째 기술적 과제를 달성하기 위하여 본 발명은,

<23> 메타데이터를 포함하는 동영상을 입력하고, 상기 입력된 동영상의 시작 프레임 및 화면 전환 프레임을 추출하는 단계, 상기 추출된 시작 프레임 및 화면 전환 프레임을 손톱 영상으로 디스플레이하고, 상기 디스플레이된 손톱 영상 및 상기 손톱 영상의 시간 정보를 저장하는 단계, 상기 입력된 동영상에 포함된 음성의 음소에 따라 화자의 음성을 인식하고, 상기 인식된 음성 데이터를 문자 데이터로 변환하고, 상기 변환된 문자데이터로부터 키워드를 추출하는 단계, 상기 입력된 동영상으로부터 자막을 검출하고, 상기 검출된 자막으로부터 자막 인식을 통하여 자막을 추출하는 단계, 사용자가 상기 디스플레이된 동영상의 시작 프레임 및 화면 전환 프레임 중 시작 샷과 끝 샷을 지정하면 상기 지정된 시작 샷, 끝 샷 및 상기 시작 샷과 끝 샷 사이의 손톱 영상에 포함된 키워드 및 자막으로부터 메타데이터와 타이틀을 추출하는 단계, 및 상기 추출된 메타데이터, 상기 시작 샷의 시간 정보, 상기 끝 샷의 시간 정보 및 상기 타이틀을 표시하는 단계를 포함하는 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법을 제공한다.

<24> 상기 두 번째 기술적 과제를 달성하기 위하여 본 발명은,

<25> 동영상을 입력하고, 상기 입력된 동영상의 시작 프레임 및 화면 전환 프레임을 추출하는 하여 상기 시작 프레임 및 화면 전환 프레임을 손톱 영상으로 디스플레이하고, 상기 디스플레이된 손톱 영상 및 상기 손톱 영상의 시간 정보를 저장하는 단계, 상기 입력된 동영상에 포함된 음성의 음소에 따라 화자의 음성을 인식하고, 상기 인식된 음성 데이터를 문자 데이터로 변환하고, 상기 변환된 문자데이터로부터 키워드를 추출하는 단계, 상기 입력된 동영상으로부터 자막을 검출하고, 상기 검출된 자막으로부터 자막 인식을 통하여 자막을 추출하는 단계, 상기 디스플레이된 동영상의 시작 프레임 및 화면 전환 프레임 중 시작 샷과 끝 샷을 지정하여 상기 지정된 시작 샷과 끝 샷 사이의 지정 동영상에 포함된 키워드 및 자막으로부터 메타데이터와 타이틀을 추출하고, 상기 추출된 메타데이터, 상기 시작 샷의 시간 정보, 상기 끝 샷의 시간 정보 및 상기 타이틀을 포함하는 동영상 데이터를 디스플레이하는 단계, 상기 디스플레이된 동영상 데이터를 XML 문서로 저장하는 단계, 및 특정 사용자가 탐색하고자 하는 검색어를 입력하면, 상기 XML 문서로 저장된 동영상의 저장 시간, 타이틀, 재생 길이 및 메타데이터를 출력하는 단계를 포함하는 메타데이터를 이용한 동영상 탐색 방법을 제공한다.

<26> 이하, 본 발명의 바람직한 실시예를 첨부도면에 의거하여 상세히 설명하기로 한다.

<27> 그러나, 다음에 예시하는 본 발명의 실시예는 여러 가지 다른 형태로 변형할 수 있으며, 본 발명의 범위가 다음에 상술하는 실시예에 한정되는 것은 아니다. 본 발명의 실시예는 당 업계에서 평균적인 지식을 가진 자에게 본 발명을 더욱 완전하게 설명하기 위하여 제공된다.

<28> 본 발명은 다양한 등장인물과 시청각 정보량이 많은 동영상을 대상으로 영상 및 음성 인식 기술을 활용하여 검색에 활용될 수 있는 메타데이터를 자동 추출한다.

<29> 이는 동영상을 재생하여 동영상에 포함된 음성 및 자막 인식을 통하여 각각의 키워드를 추출한 뒤, 추출된 핵심어의 융합을 통해 메타데이터를 추출한다.

<30> 이와 같이, 추출된 메타데이터를 동영상의 시간 정보와 함께 XML 형식의 파일로 저장하여 동영상의 기사 검색에 대한 정보로써 사용할 수 있다.

<31> 이에 있어서 본 발명에 핵심적으로 요구되는 기술 분야는 샷 검출 기술, 화자 독립 대어휘 연속어 음성 인식 기술과 영상에 대한 자막 검출 및 문자 인식 기술, 영상과 음성 인식을 통하여 얻은 결과로부터 핵심어 추출 및 융합하는 기술이다.

<32> 본 발명은 동영상의 재생으로부터 시작된다. 동영상을 재생하게 되면, 샷 검출 기술을 이용하여 동영상의 시작 장면 및 화면이 전환되는 주요 장면들을 샷의 시간 정보와 함께 화면에 손톱 영상으로 제시해 준다.

<33> 기본적으로 샷 검출 기술로는 30프레임마다 한 번씩 영상 프레임을 가져와서 이전에 획득한 영상 프레임과 히스토그램 분포 변동을 비교하여, 변동값이 일정값 이상 클 경우, 화면 전환 프레임으로 판단하고, 이렇게 검출된 화면 전환 프레임은 다양한 동영상을 포함하는 시청각 자료의 각 영역에 대한 잣대로 작용하게 하여 화면에 포함된 손톱영상을 클릭함으로써 영역의 시작 샷과 마지막 샷을 지정해주어 동영상의 구간을 지정할 수 있다.

<34> 예를 들어, 뉴스에서 기사 내용이 시작할 때, 첫 장면의 우측 상단에 기사에 대한 제목 자막을 표시하는 뉴스 특성을 이용하여 음성과 자막의 인식이 완료되고 난 후 뉴스 기사에 대한 시작 샷과 마지막 샷을 지정할 경우,

시작 샷에 해당하는 시간 정보를 가져와 시작 샷에 포함되어 있는 뉴스 제목 자막에 대한 자막 인식 결과를 찾아 기사 내용에 대한 제목으로 화면에 표현하여 주도록 하였다. 이를 통하여 주요 샷과 제목을 화면에 표현하여 기사 내용을 한눈에 파악할 수 있다.

<35> 핵심적으로 요구되는 음성 인식 기술 분야는 화자 독립 음성 인식 기술 및 연속 핵심어 인식 기술(Keyword spotting)이다. 방송에서 사용된 오디오 정보 중에 불특정 다수의 문장을 인식하기 위해서는 '화자 독립 음성 인식 기술'이 요구되고, 문장 중에서 메타데이터로 사용하기 위한 핵심 어휘를 추출하기 위하여 '연속 핵심어 인식 기술'이 필요하다. 본 발명에서는 보다 신뢰성 있는 핵심어 인식기 개발을 위해 먼저 대어휘 연속 음성 인식기(Large vocabulary continuous speech recognition)를 통하여 얻어진 전사(description) 데이터를 후처리 과정을 통하여 핵심어를 검출하는 방법을 사용하기 위하여 대어휘 연속 음성 인식기를 적용한다.

<36> 본 발명은 다양한 뉴스, 오락 프로그램 등의 동영상 데이터로부터 신뢰성 있는 음성 인식 성능을 얻기 위하여 불특정 다수의 음성을 인식할 수 있는 화자 독립 음성인식 기술을 사용한다. 이를 위하여 다양한 출생지의 1000명이 발음한 음성 데이터로부터 음향 모델을 생성하여, 어떠한 화자가 발성한 음성이라도 인식에 문제가 없도록 디자인한다. 또한, 음성 인식 성능 향상을 위하여 문맥에 따른 조음효과를 반영하기 위하여 음성 인식의 단위를 현재 음소의 앞, 뒤 음소를 함께 포함하는 문맥 종속형 모델인 프라이-폰(tri-phone)을 기본 단위로 하며, 대어휘 인식시스템에 적합하도록 압축에 다른 손실이 없는 높은 성능을 지닌 음성 인식 알고리즘인 연속 분포 은닉 마르코프 모델(continuous density Hidden Markov Model:continuous density HMM)을 사용한다.

<37> 또한, 동영상 데이터로부터 메타데이터를 추출하기 위하여 대어휘 연속어 음성인식 기술을 개발하여 문장 단위의 음성 데이터를 문자 데이터인 전사 데이터로 변환한다.

<38> 도 1은 본 발명의 일 실시예에 따른 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법의 흐름도이다.

<39> 도 1을 참조하면, 우선, 메타데이터를 포함하는 동영상을 입력하고, 상기 입력된 동영상의 시작 프레임 및 화면 전환 프레임을 추출한다(110 과정).

<40> 본 발명에 적용되는 동영상은 음성 정보와 문자 정보를 포함하는 동영상일 수 있다. 상기 동영상을 입력하면 우선 동영상의 시작 프레임을 추출한다. 그리고, 사용자에 의해 미리 설정된 영상 프레임 단위마다 영상 프레임을 저장한다.

<41> 상기 미리 설정된 영상 프레임 단위는 30프레임으로 설정할 수 있으며, 이는 동영상의 품질, 재생 환경, 사용자의 실시 형태에 따라서 설정 및 변경이 가능하다.

<42> 그러면, 미리 설정된 영상 프레임 단위별로 영상 프레임을 저장하고, 상기 저장된 영상 프레임과 직전에 저장된 이전 영상 프레임의 히스토그램 분포를 비교하여 히스토그램 분포의 변동 값이 사용자에 의해 미리 설정된 임계값을 초과하는 경우, 상기 저장된 영상 프레임을 화면 전환 프레임으로 추출한다.

<43> 예를 들면, 다양한 뉴스를 포함하는 동영상에서 앵커의 멘트와 앵커의 멘트에 따르는 영상 정보 간 영상의 히스토그램의 분포가 다르고, 동영상의 내용이 다를 경우 이에 따른 각각의 히스토그램 분포가 다르게 된다.

<44> 따라서, 영상 프레임 단위별로 영상 프레임을 로드한 후, 이전에 저장된 영상 프레임과의 히스토그램 분포를 비교하여 화면 전환 프레임을 추출하면 영상의 성격에 따라 개별적인 정보의 추출이 가능하게 된다.

<45> 한편, 미리 설정된 임계값은 다양한 뉴스, 오락프로그램, 드라마 등 과거의 동영상의 화면 전환 시간과 히스토그램 변화 값을 상호 비교하고, 상기 화면 전환 시간에 따른 히스토그램 변화 값의 평균값으로 임계값을 설정할 수 있다.

<46> 그 다음, 상기 추출된 시작 프레임 및 화면 전환 프레임을 손톱 영상으로 디스플레이하고, 상기 디스플레이된 손톱 영상 및 상기 손톱 영상의 시간 정보를 저장한다(120 과정).

<47> 즉, 동영상의 시작 프레임 및 상기 히스토그램 변화 값에 따라 화면 전환 프레임을 손톱 영상으로 디스플레이한다.

<48> 상기 디스플레이된 손톱 영상은 상기 손톱 영상의 시작 시간과 그 다음 손톱 영상의 시작 시간과 동일하도록 상기 손톱 영상의 끝 시간을 설정하여 저장한다.

<49> 이와 같이, 상기 손톱 영상에 시간 정보를 저장하게 함으로써 동영상을 각각 분리할 수 있게 된다.

<50> 그 다음, 상기 입력된 동영상에 포함된 음성의 음소에 따라 화자의 음성을 인식하고, 상기 인식된 음성 데이터를 문자 데이터로 변환하고, 상기 변환된 문자데이터로부터 키워드를 추출한다(130 과정).

<51> 실제로, 동영상에 등장하는 언어는 일반적인 대화체에서 시작하여 전문적인 용어까지 포함하는 대용량의 어휘를 필요로 한다. 그리고 음성의 인식은 어떤 단위로 인식하느냐에 따라서 탐색하여야 하는 공간의 크기와 인식에 소요되는 탐색 시간이 결정된다. 그리고, 방송 뉴스와 같은 대용량 어휘의 연속어 인식의 경우에는 어떤 인식 단위를 사용하느냐에 따라서 탐색 공간과 탐색 시간이 크게 좌우된다.

<52> 화자의 독립적인 음성 인식을 위하여 다양한 화자의 음성 특징에 대해 확률적 분포를 추정하고, 이를 인식과정에 이용하는 것이 필요하다. 일반적으로 확률 기반의 음향 모델링에 가장 많이 사용되는 알고리즘은 은닉 마르코프 모델(Hidden Markov Model:HMM) 알고리즘이다.

<53> HMM도 다양한 종류(DHMM, SCHMM, GMM 등)가 있지만 그 중 구현된 시스템의 환경이 임베디드 시스템과 같이 메모리나 CPU 성능에 크게 영향을 받지 않으면서 인식 성능을 최대로 하는 알고리즘은 확률의 분포를 연속분포로 사용하는 연속 밀도 은닉 마르코브 모델(Continuous Density Hidden Markov Model:CDHMM)을 사용할 수 있다. 이는 상태 내 확률 분포 표현에 연속된 가우시안 분포를 이용하므로, 어떠한 특징값에 대해서도 확률을 산출할 수 있다. 특히 특징이 입력 되었을 경우, 그에 해당하는 확률 분포 표현에 연속된 가우시안 분포(Gaussian mixture)를 이용할 수 있으므로 어떠한 특징 값에 대해서도 확률을 산출할 수 있다.

<54> 본 발명에서는 화자의 통계적 음향 모델을 양자화 에러 없이 보다 상세히 표현하기 위해 CHMM 알고리즘을 통해 음향모델을 생성한다. 또한 화자에 독립적인 통계적 특성을 모델링하기 위해 복수의 화자가 발성한 음성 데이터베이스로부터 문맥의 조음효과를 고려한 트라이-폰(tri-phone) 단위의 음소 단위로 음향 모델을 생성할 수 있다.

<55> 또한, 상기 화자의 음성은 문장과 문장 사이에 시간적인 갭(gap)이 존재하는 것을 이용하여 문장 단위로 인식될 수 있다.

<56> 예를 들어, 도 2를 참조하면, '취나물에서 이 농약이 기준치보다 육백 배가 넘게 나왔습니다'라는 단위 문장이 인식되었을 경우, 매 단어마다 서브워드(subword) 단위인 음소열로 단어를 구성하고, 각 서브워드인 음소는 이미 훈련된 음향모델의 CDHMM 파라미터로 대체하여 트라이-폰(tri-phone) 기반의 서브워드 노드로 구성한다. 또한 입력된 음성으로부터 각 단어 쌍의 확률인 바이-그램(Bi-gram)으로부터 단어간의 연결과 바이-그램 값을 적용하여 단어 간 네트워크를 구성한다.

<57> 이와 같이, 워드 네트워크의 각 단어마다 발음사전과 음향 모델을 참조하여 서브워드 모델인 트라이-폰 단위로 확장하여 전체 탐색 공간을 생성하게 된다.

<58> 한편, 음성 인식에 따라 키워드를 추출하기 위하여 기존의 단어에 대한 중요도를 판단하는 방법인 핵심어 후보 단어를 포함하는 데이터 베이스를 구성하여, 각 단어에 대하여 단어 빈도수(Term Frequency:TF) 및 역문헌 빈도수(Inverse Document Frequency:IDF)에 따라 키워드를 추출하는 방법을 사용할 수 있다.

<59> 이를 위하여, 예를 들어 과거의 음성으로부터 텍스트로 변환하는 음성 전사(Description) 자료을 이용하여 3만 단어 급의 핵심어 후보 단어의 데이터베이스를 구성하고, 상기 구성된 핵심어 후보 단어의 데이터베이스를 이용하여 사용자가 지정한 구간의 음성 인식 결과에서 계산된 단어 빈도 수를 곱한 값을 단어의 중요도에 대한 척도로 사용할 수 있다.

<60> 그 다음, 상기 입력된 동영상으로부터 자막을 검출하고, 상기 검출된 자막으로부터 자막 인식을 통하여 자막을 추출한다(140 과정).

<61> 실제로, 영상에서 사용된 자막의 경우 해당 프레임에 대한 요약 및 주제가 담겨있는 경우가 많으므로, 대부분의 자막을 메타데이터로 사용할 확률이 높다.

<62> 이를 제한된 영역에 대해 자막 인식을 통하여 자막 검출 및 자막 인식을 수행하여 자막 인식 결과를 얻게 된다.

<63> 그 다음, 사용자가 상기 디스플레이된 동영상의 시작 프레임 및 화면 전환 프레임 중 시작 샷과 끝 샷을 지정하면 상기 지정된 시작 샷, 끝 샷 및 상기 시작 샷과 끝 샷 사이의 손톱 영상에 포함된 키워드 및 자막으로부터 메타데이터와 타이틀을 추출한다(150 과정).

<64> 이는 사용자가 디스플레이된 동영상으로부터 추출된 시작 프레임 및 화면 전환 프레임 중 원하는 구간의 시작 샷과 끝 샷을 지정하면 상기 지정된 시작 샷, 끝 샷 및 상기 시작 샷과 끝 샷 사이의 손톱 영상으로부터 추출된

키워드 및 자막으로부터 메타데이터와 타이틀을 추출한다.

<65> 상세하게는, 상기 키워드와 자막이 일치하는 단어가 존재하는 경우 메타데이터로서의 중요도가 높을 수 있으므로, 상기 일치하는 단어를 메타데이터로 우선적으로 추출할 수 있다.

<66> 이는 상기 키워드와 자막이 일치하는 단어가 존재하는 경우, 상기 단어에 가중치를 부여하고, 상기 가중치가 부여된 단어를 메타데이터로 우선적으로 추출할 수 있고, 동일한 가중치가 부여된 단어가 복수 개 존재하는 경우, '가나다'순으로 메타데이터의 단어를 추출할 수 있다.

<67> 한편, 상기 타이틀은 상기 지정된 시작 샷에 포함된 자막으로 설정할 수 있다.

<68> 이는 시작 샷에 포함된 자막은 실제 동영상에서 핵심 단어를 포함하는 경우가 많으므로 이를 반영하여 상기 타이틀은 상기 지정된 시작 샷에 포함된 자막으로 설정하는 것이며, 사용자의 설정에 따라 상기 타이틀은 수정 변경이 가능할 수 있다.

<69> 마지막으로, 상기 추출된 메타데이터, 상기 시작 샷의 시간 정보, 상기 끝 샷의 시간 정보 및 상기 타이틀을 표시하여 사용자에게 상기 동영상의 특정 구간에 대한 동영상 정보를 제공한다(160 과정).

<70> 도 3은 도 1의 영상의 음성으로부터 키워드를 추출하는 단계(130 과정)의 상세 흐름도이다.

<71> 우선, 입력된 동영상으로부터 화자의 음성을 인식한다(331 과정).

<72> 본 발명에 적용되는 음성 인식에는 화자 독립 음성 인식 방법 및 연속 핵심어 인식 방법이 적용될 수 있다.

<73> 이는 동영상에 포함된 오디오 정보 중에 불특정 다수의 문장을 인식하기 위해서는 화자 독립 음성인식 방법이 요구되며, 문장 중에서 메타데이터로 사용하기 위한 키워드를 추출하기 위하여 연속 핵심어 인식 방법이 요구된다.

<74> 본 발명에서는 보다 신뢰성 있는 메타데이터를 추출하기 위하여 먼저 대어휘 연속어 음성 인식(Large vocabulary continuous speech recognition) 방법을 통하여 얻어진 전사(description) 데이터로부터 키워드를 검출하게 된다.

<75> 실제로 연속어 음성 인식 방법에 따른 연속 인식 탐색 공간 생성을 위하여 바이-그램(bi-gram) 기반의 탐색 공간을 생성하는데, 이는 기존의 동영상으로부터 텍스트화 된 어휘를 추출하여 바이-그램을 생성한다.

<76> 그리고, 대어휘 인식을 위하여 인식 시스템의 런-타임(run-time)시 생성 및 할당되는 메모리의 사용을 정적 할당으로 변환하여 불필요한 하드웨어 접근 시간을 극소화 하고, 확률 계산시 필수적인 지수, 로그 함수를 테이블 룩 어헤드(table-look ahead) 방식을 통하여 복잡한 수식 연산을 미리 연산하여 근사화 함으로써 연산속도를 줄일 수 있다.

<77> 그 다음, 상기 추출된 화면 전환 프레임에 따라 음성 인식 결과 구간을 설정한다(332 과정).

<78> 상술한 바와 같이 화면 전환 프레임이 추출되면, 추출된 화면 전환 프레임은 다음 화면 전환 프레임이 추출되기 전까지의 영상 정보를 저장하게 되므로, 상기 추출된 화면 전환 프레임 사이의 구간을 음성 인식 결과 구간으로 설정할 수 있다.

<79> 이와 같이, 추출된 화면 전환 프레임을 기반으로 음성 인식 결과 구간의 음성 정보로부터 키워드를 추출하게 된다.

<80> 마지막으로, 설정된 음성 인식 결과 구간의 음성 정보로부터 키워드를 추출한다(333 과정).

<81> 본 발명에 있어서, 음성 정보로부터 키워드를 추출하는 방법은 통계적인 각 단어의 빈도수에 기반한다. 이는 어떤 단어가 설정된 음성 인식 결과 구간에 자주 나타나는 단어라면 일반적으로 중요한 단어라고 판단할 수 있다. 그러나, '오늘은'이나, '~기자입니다.'와 같은 의례적으로 모든 음성 인식 결과 구간에 나타나기 쉬운 단어에 있어서 빈도수 만을 이용하여 키워드를 추출한다면, 실제 키워드에 더해서 필요없는 단어까지 같이 검출이 되는 문제가 있다.

<82> 따라서, 이를 동시에 고려하는 것이 TF-IDF(Term Frequency and Inverted Document Frequency)로서 하기의 수학식 1로 표현될 수 있다.

<83> 하기 수학식 1에서 N은 구성된 데이터베이스에 존재하는 모든 단어 수를 의미하며, n은 구성된 데이터 베이스에 존재하는 특정 단어의 빈도수를 의미하고, p는 키워드를 추출하고자 하는 현재 동영상 구간에 존재하는 전체 단

어의 수를 의미하고, t는 특정 단어의 현재 동영상 구간에서의 빈도 수를 의미한다.

수학식 1

$$TF-IDF=\log\left(\frac{N}{n+1}\right)\times\frac{t}{p}$$

<84>

<85> 즉, 상기 수학식 1에 따라 역문헌 빈도 수인 해당 단어의 구성된 데이터베이스에서의 빈도의 역수를 취하고, 해당 동영상 구간에서는 빈도 수를 곱함으로써 다른 동영상에서는 빈도 수가 낮으면서, 해당 동영상 구간에 빈도 수가 높은 단어가 TF-IDF에 따른 키워드로 추출되게 된다.

<86> 도 4는 도 1의 자막 인식을 통하여 자막을 추출하는 단계(140 과정)의 상세흐름도이다.

<87> 도 4를 참조하면 우선, 입력 영상의 전처리로 가우시안 필터링 및 컬러도메인에서의 필터링을 수행하여 자막 후보 영역을 검출한다(441 과정).

<88> 즉, 우선적으로 자막이 포함된 자막 프레임의 검출을 위하여 자막이 가지는 고주파 성분과 배경과의 강한 대조비를 이루는 특성을 고려하여 샘플된 프레임에서 자막 후보 영역을 검출한다.

<89> 여기서, 강한 고주파 성분을 추출하기 위하여 각 프레임의 미분 연산자(Gradient operator)를 사용하며, 컬러 정보의 손실을 방지하기 위하여 각 프레임에서 컬러 정보를 가지고 있는 빨강, 녹색, 파랑(Red, Green, Blue:RGB) 성분에 대해서 연산자를 사용한다.

<90> 컬러미분연산자(Color gradient operator)를 사용함으로써 자막이 가지고 있는 고주파 성분 추출 및 강한 대조비를 그대로 유지시킬 수 있다.

<91> 이는 흑백(Gray-level) 영상 변환 후 1차원 성분에서의 미분연산자를 사용하는 경우보다 3차원의 도메인을 가지는 컬러 영상에 대한 정보를 보존하면서도 상기 컬러 영상을 이용할 수 있기 때문에 자막이 가지는 고주파 성분 추출에 강인함을 보일 수 있다.

<92> 각 3차원(R, G, B) 도메인에서 얻은 미분 연산값으로부터 벡터연산을 통하여 강한 대조비를 가지는 성분을 추출하고, 미분연산과 벡터 연산을 통하여 배경성분을 최대한 배제하면서 자막 성분을 추출할 수 있는 특징 영상(Edge image)을 구할 수 있다.

<93> 상술한, 자막의 특징이 있는 배경이 존재할 수 있으므로 특징 영상에서 다시 미분연산을 한 후, 특징 영상의 증감변화에 대한 쌍대성(Edge pair) 및 자막이 가지는 거리성분을 고려하여 수평방향 및 수직방향의 좌표값을 얻을 수 있다.

<94> 그 다음, 자막 후보 영역에 대한 검증을 수행한다(442 과정).

<95> 이는 자막 후보 영역에는 배경에 존재하는 글자 혹은 자막이 가지는 특징과 유사한 성질을 가지는 배경이 존재할 수 있기 때문에 자막 후보 영역에 대한 검증 과정이 필요하기 때문이다. 자막 후보 영역의 검증은 정상 영역(실제 인위적인 자막 영역)만을 분류하고 이를 실제 자막으로 추출할 수 있는 영역과 배경에 존재하는 문자 혹은 기타 영역을 분리하여 자막 후보 영역만을 정확하게 추출하기 위함이다. 이로 인하여 자막 부분에 대한 검출율을 향상시킬 수 있다.

<96> 자막 후보 영역에 대한 검증을 위하여 투 클래스(Two-Class)인 자막 영역과 비 자막 영역으로 분류하는 것이 아니라, 실제 자막 영역만을 가지고 데이터 도메인에 대한 클래스를 구성하는 원 클래스(One-Class) 문제로 분석한다.

<97> 즉, 훈련(Training)을 통해 정상 영역에 대한 자막 후보 영역을 구성하고, 상기 자막 후보 영역이 구성된 영역 밖에 존재하는 영상 프레임 내에서의 영역은 아웃라이어(Outlier)로 간주하여 자막 후보 영역을 검증하게 된다.

<98> 이와 같은 자막 후보 영역의 검증은 훈련 데이터의 수를 상당히 줄일 수 있는 장점과 인위적인 자막(Superimposed caption)과 상당히 유사한 배경에 존재하는 글자를 포함하는 글자영역을 구별할 수 있다.

<99> 마지막으로, 검출된 자막 영역의 이진화(Binarization), 자막 영역에서의 글자단위 분리, 한글 자모음 유형 분류 및 자소 분리, 자소단위 인식 단계를 거쳐 자막을 인식한다(443 과정).

<100> 이는, 이진화의 성능에 따라서 문자 인식의 성능이 좌우되므로, 자막 영역의 휘도 성분을 추출하여 히스토그램을 분석하는 방법을 사용하여 정교한 이진화를 실시할 수 있다. 휘도 성분의 히스토그램 분석을 위하여 동적 분산 값과 동적 필터 크기를 적용한 평활화를 적용하여 배경과 자막 성분과 관련된 두 개의 첨점(Peak)과 계곡값(Valley)을 이용하여 1차 동적 문턱값(threshold)을 구한다.

<101> 그리고, 자막의 경계 부분의 정확한 이진화를 위하여 자막의 높이를 고려하여 동적인 마스크를 생성하여 마스크 내의 자막과 배경성분의 값을 분석하여 2차 동적 문턱값을 생성한다.

<102> 그런 다음, 이진화된 자막 후보 영역 내의 글자 단위의 비에 따른 분리를 수행하는데, 이는 글자의 높이 대 폭을 분석하고, 글자의 경계 부근의 수직방향의 투영값을 분석하여 자막성분의 글자의 분리를 수행한다.

<103> 하나의 글자에서 자음과 모음이 분리되어 수직방향 투영시에는 2개의 글자로 분리되는 경우를 대비하여 높이 대 폭을 분석한 뒤, 수직 성분 유무를 판단하여 정확한 영역 내의 글자단위 분리를 실시한다.

<104> 글자 단위 분리 후 수직, 수평성분의 투영값의 최대, 최소, 평균 등을 특징벡터로 추출하여 한글 유형 분류를 하고, 자모음 유형분류 후 얻어진 모음 성분과 종성 성분 유무의 정보를 가지고 자소 분리 과정을 수행한다. 각 유형별 모음성분을 고려하여 투영값의 최대값, 최소값 및 변화율 등을 분석하여 자소 분리를 수행한다.

<105> 상기 분리된 자음과 모음으로부터 중심점을 계산하여 정규화된 크기로 투영하여 중심점으로부터의 일정거리 성분 및 방향성분의 분포를 특징벡터로 사용하여 각 분리된 자소를 신경망을 통해 인식하고, 상기 인식된 자소인식결과를 조합하여 글자 단위로 자막을 인식하고, 상기 인식된 자막을 추출한다.

<106> 도 5는 도 1의 입력 동영상의 시작 프레임 및 화면 전환 프레임 추출 단계(110 과정)의 상세 흐름도이다.

<107> 도 5를 참조하면 우선, 입력된 동영상을 재생한다(511 과정).

<108> 상기 입력 동영상은 실시간 동영상 및 녹화된 동영상일 수 있으며, 바람직하게는 상기 입력 동영상은 MPEG-2 표준화 규격에 따른 동영상을 포함할 수 있다.

<109> MPEG-2 표준화 규격에 따른 동영상은 텔레비전 방송, 통신, 오디오·비디오 기기 등 광범위한 적용 분야를 대상으로 하는 고품질의 규격을 가지는 동영상 표준으로 고선명 텔레비전(HDTV)의 화질을 감안하여 4~100Mbps고화질의 규격을 가지기 때문에 영상의 전송, 디지털 비디오 디스크(DVD) 등의 디지털 저장 매체의 개발과 소프트웨어 제작에 활용도가 높으므로 본 발명에 쉽게 접목할 수 있다.

<110> 그 다음, 상기 입력된 동영상의 시작 프레임의 화면을 손톱 영상으로 디스플레이하고(512 과정), 사용자에 의해 설정된 프레임 단위마다 영상 프레임을 저장한다(513 과정).

<111> 한편, 사용자에 의해 설정된 프레임 단위는 영상에 따라서 영상의 화면 변화의 동적 변화 정도 및 사용 상태에 따라 임의로 설정할 수 있으나, 바람직하게는, 영상의 움직임을 더욱 세밀하게 묘사할 수 있는 30 프레임 단위로 설정할 수 있다.

<112> 그 다음, 저장된 영상 프레임의 히스토그램 분포 정보를 획득한다(514 과정). 히스토그램은 이미지에서 밝기 수치의 구분을 그래픽으로 설명한 것으로 가로축은, 어두운 화소는 왼쪽으로, 밝은 화소는 오른쪽으로 향하는 밝기의 수치(0에서 255)를 나타내게 된다. 세로축은 각 단계의 밝기에서 화소들의 비율을 나타낸다. 따라서, 영상 프레임의 히스토그램을 확인함으로써 사용자는 이미지의 전반적인 노출을 측정할 수 있으므로, 히스토그램이 왼쪽으로 치우쳐서 높게 나타나면 이미지가 대부분 어두운 화소로 구성되며 이미지가 어둡게 나타나는 특성을 가지는 바, 영상의 히스토그램 분포를 살펴보면 영상의 특성을 알 수 있다.

<113> 그 다음, 저장된 영상 프레임의 히스토그램 분포와 직전에 저장된 영상 프레임의 히스토그램 분포를 비교하여 히스토그램의 변화 값이 사용자에 의해 미리 설정된 임계값보다 큰 값을 가지는 지의 여부를 판단하여(515 과정), 상기 히스토그램 분포 변화가 임계값보다 큰 경우, 손톱 영상으로 상기 저장된 영상 프레임을 디스플레이하고(516 과정), 상기 히스토그램 분포 변화가 임계값보다 작은 경우 다시 사용자에 의해 설정된 프레임 단위별로 영상 프레임을 로드하여 저장하게 된다(513 과정).

<114> 실제로, 상술한 바와 같이 다양한 뉴스를 포함하는 동영상에서 앵커의 멘트와 앵커의 멘트에 따르는 영상 정보간 영상의 히스토그램의 분포가 다르고, 동영상의 내용이 다를 경우 이에 따른 각각의 히스토그램 분포가 다르게 된다.

<115> 따라서, 영상 프레임 단위별로 영상 프레임을 로드한 후, 이전에 저장된 영상 프레임과의 히스토그램 분포를 비

교하여 화면 전환 프레임을 추출하면 영상의 성격에 따라 개별적인 정보의 추출이 가능하게 된다.

<116> 한편, 미리 설정된 임계값은 다양한 뉴스, 오락프로그램, 드라마 등 과거의 동영상의 화면 전환 시간과 히스토그램 변화 값을 상호 비교하고, 상기 화면 전환 시간에 따른 히스토그램 변화 값의 평균값으로 임계값을 설정할 수 있다.

<117> 도 6은 도 1의 메타데이터와 타이틀을 추출하는 단계(150 과정)의 상세 흐름도이다.

<118> 도 6을 참조하면, 우선 사용자에 의해 손톱 영상으로 디스플레이된 동영상의 시작 프레임 및 화면 전환 프레임 중 시작 샷과 끝 샷을 지정한다(651 과정).

<119> 즉, 사용자는 동영상의 재생시에 히스토그램의 변화값으로 추출된 화면 전환 프레임과 디스플레이된 동영상으로부터 추출된 시작 프레임 중 메타데이터를 추출하고자 하는 시작 샷과 끝 샷을 지정한다.

<120> 그 다음, 화면 전환 프레임 중 원하는 구간의 시작 샷과 끝 샷을 지정하면 상기 지정된 시작 샷 및 끝 샷의 시간 정보를 획득한다(652 과정).

<121> 상기 시간 정보를 획득함으로써, 지정된 시작 샷 및 끝 샷 사이에 포함되는 화면 전환 프레임의 데이터를 함께 로드한다.

<122> 그 다음, 상기 지정된 시작 샷의 자막 인식 결과를 획득한다(653 과정). 상기 자막 인식 결과는 상술한 자막 추출 및 인식 과정에 의할 수 있다.

<123> 마지막으로, 상기 획득한 시작 샷의 자막 인식 결과를 타이틀로 설정하고, 지정된 시작 샷 및 끝 샷의 메타데이터를 추출한다(654 과정).

<124> 이는 전술한 바와 같이, 시작 샷에 포함된 자막은 실제 동영상에서 핵심 단어를 포함하는 경우가 많으므로 이를 반영하여 상기 타이틀은 상기 지정된 시작 샷에 포함된 자막으로 설정하는 것이며, 사용자의 설정에 따라 상기 타이틀은 수정 변경이 가능할 수 있다.

<125> 한편, 상기 메타데이터의 추출은 자막과 키워드를 융합하여 생성하되, 상기 자막이 문장형으로 추출될 경우 상기 자막 중 명사인 단어를 자막으로 추출할 수 있다.

<126> 일반적으로 문장 내에 포함되어 있는 명사가 물체의 행동 혹은 상태를 서술함으로써 의미를 전달한다. 즉, 문장에서 전달하고자 하는 주요 의미는 명사에 대한 표현에 집중되어 있다. 따라서, 문장에서 표현하고자 하는 명사가 그 문장의 핵심어를 대부분 포함하고 있는 것을 알 수 있다. 이러한 자막을 추출하기 위하여 문장형으로 표현되는 자막에 대하여 형태소 분석을 통하여 문장 중에 명사로 판단되는 단어를 자막으로 추출한다.

<127> 아울러, 명사를 추출함과 동시에 자막으로 추출된 성분에 대하여 불필요한 단어를 제거할 수 있다.

<128> 이는 본 발명에 있어서 자막은 한글 자막에 대한 것으로, 숫자나 특수 기호 뒤에 올 수 있는 단어에 대해 핵심어로 출력되지 않도록 할 수 있다.

<129> 즉, 숫자 다음에 올 수 있는 명사의 품사를 가지지만 의미상 핵심어로 판단되지 않는 '단위 표현'인 경우, 자막 성분에서 제외할 수 있다.

<130> 도 7은 본 발명의 다른 실시예에 따른 메타데이터를 이용한 동영상 탐색 방법의 흐름도이다.

<131> 우선, 동영상을 입력하고, 상기 입력된 동영상의 시작 프레임 및 화면 전환 프레임을 추출하여 상기 시작 프레임 및 화면 전환 프레임을 손톱 영상으로 디스플레이하고, 상기 디스플레이된 손톱 영상 및 상기 손톱 영상의 시간 정보를 저장한다(710 과정).

<132> 그 다음, 상기 입력된 동영상에 포함된 음성의 음소에 따라 화자의 음성을 인식하고, 상기 인식된 음성 데이터를 문자 데이터로 변환하고, 상기 변환된 문자데이터로부터 키워드를 추출한다(720 과정).

<133> 그 다음, 상기 입력된 동영상으로부터 자막을 검출하고, 상기 검출된 자막으로부터 자막 인식을 통하여 자막을 추출한다(730 과정).

<134> 그 다음, 상기 디스플레이된 동영상의 시작 프레임 및 화면 전환 프레임 중 시작 샷과 끝 샷을 지정하여 상기 지정된 시작 샷과 끝 샷 사이의 지정 동영상에 포함된 키워드 및 자막으로부터 메타데이터와 타이틀을 추출하고, 상기 추출된 메타데이터, 상기 시작 샷의 시간 정보, 상기 끝 샷의 시간 정보 및 상기 타이틀을 포함하는 동영상 데이터를 디스플레이한다(740 과정).

<135> 상기 710과정 내지 740과정은 전술한 바와 동일하므로 중복된 상술은 생략하기로 한다.

<136> 그 다음, 상기 디스플레이된 동영상 데이터를 XML 문서로 저장한다(750 과정).

<137> 상기 추출된 메타데이터, 상기 시작 샷의 시간 정보, 상기 끝 샷의 시간 정보 및 상기 타이틀을 포함하는 동영상 데이터를 디스플레이된 정보를 그래픽 유저 인터페이스(Graphic User Interface:GUI) 상에서 사용자에 의해 XML 문서로 저장할 수 있다.

<138> 또한, 상기 XML 문서는 MPEG-7의 표준에 맞춘 XML 문서를 포함할 수 있다.

<139> 도 8을 함께 참조하면, 상기 추출된 메타데이터, 상기 시작 샷의 시간 정보, 상기 끝 샷의 시간 정보 및 상기 타이틀을 포함하는 동영상 데이터를 그래픽 유저 인터페이스(Graphic User Interface:GUI) 상에서 확인할 수 있으며 사용자에 의해 XML 문서로 쉽게 임포팅(importing)이 가능하다.

<140> 마지막으로, 특정 사용자가 탐색하고자 하는 검색어를 입력하면, 상기 XML 문서로 저장된 동영상의 저장 시간, 타이틀, 재생 길이 및 메타데이터를 출력한다(760 과정).

<141> 도 9를 함께 참조하면, 도 9에서 볼 수 있는 바와 같이, 특정 사용자가 탐색하고자 하는 검색어를 입력하면 웹 브라우저 상에서 상기 검색어와 동일한 단어를 메타 데이터로 하는 동영상의 정보를 출력하고, 이를 통하여 특정 사용자는 상기 검색어에 해당하는 동영상의 정보를 미리 알 수 있으며, 실제로 상기 출력된 동영상 정보에 해당하는 동영상을 재생할 수 있다.

<142> 본 발명에 따른 상기의 각 단계는 일반적인 프로그래밍 기법을 이용하여 소프트웨어적으로 또는 하드웨어적으로 다양하게 구현할 수 있다는 것은 이 분야에 통상의 기술을 가진 자라면 용이하게 알 수 있는 것이다.

<143> 그리고 본 발명의 일부 단계들은, 또한, 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치를 포함한다. 컴퓨터가 읽을 수 있는 기록매체의 예로는 ROM, RAM, CD-ROM, CD-RW, 자기 테이프, 플로피디스크, HDD, 광 디스크, 광자기 저장장치 등이 있으며, 또한 캐리어 웨이브(예를 들어 인터넷을 통한 전송)의 형태로 구현되는 것도 포함한다. 또한 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산방식으로 컴퓨터가 읽을 수 있는 코드로 저장되고 실행될 수 있다.

<144> 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 본 개시된 실시예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 상기의 설명에 포함된 예들은 본 발명에 대한 이해를 위해 도입된 것이며, 이 예들은 본 발명의 사상과 범위를 한정하지 않는다. 상기의 예들 외에도 본 발명에 따른 다양한 실시 태양이 가능하다는 것은, 본 발명이 속한 기술 분야에 통상의 지식을 가진 사람에게는 자명할 것이다. 본 발명의 범위는 전술한 설명이 아니라 청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

발명의 효과

<145> 상술한 바와 같이 본 발명에 의하면, 음성 인식 결과와 동영상 자체의 개방형 자막 인식 결과인 방송 콘텐츠에 포함되어 있는 메타 데이터 정보를 자동으로 추출하고, 새로운 방송 자료들의 수작업에 의한 아카이빙(archiving) 작업시간을 단축할 수 있으며, 과거의 방대한 양의 방송 데이터들에 대한 콘텐츠 관리 및 색인 작업을 자동으로 실행하여 방송 제작자들에게 양질의 콘텐츠를 제작할 수 있게 함으로써, 과거의 자료를 이용한 콘텐츠 제작에 소요되는 시간과 비용을 절약할 수 있도록 할 수 있을 뿐만 아니라 현재 광범위하게 적용되고 있는 인터넷 검색 사이트와 연동하여 실시간으로 제작되는 콘텐츠에 대한 멀티미디어 자료들의 검색을 가능하게 하고, 과거의 자료들에 대한 열람을 용이하게 함으로써, 사용자의 콘텐츠 사용의 편의성을 제공하는 효과가 있다.

도면의 간단한 설명

<1> 도 1은 본 발명의 일 실시예에 따른 동영상의 음성 인식과 자막 인식을 통한 메타데이터 추출 방법의 흐름도이다.

<2> 도 2는 본 발명에 적용되는 연속어 음성인식 탐색공간을 도시한 것이다.

<3> 도 3은 도 1의 영상의 음성으로부터 키워드를 추출하는 단계(130 과정)의 상세 흐름도이다.

<4> 도 4는 도 1의 자막 인식을 통하여 자막을 추출하는 단계(140 과정)의 상세흐름도이다.

<5> 도 5는 도 1의 동영상의 시작 프레임 및 화면 전환 프레임 추출 단계(110 과정)의 상세 흐름도이다.

<6> 도 6은 도 1의 메타데이터와 타이틀을 추출하는 단계(150 과정)의 상세 흐름도이다.

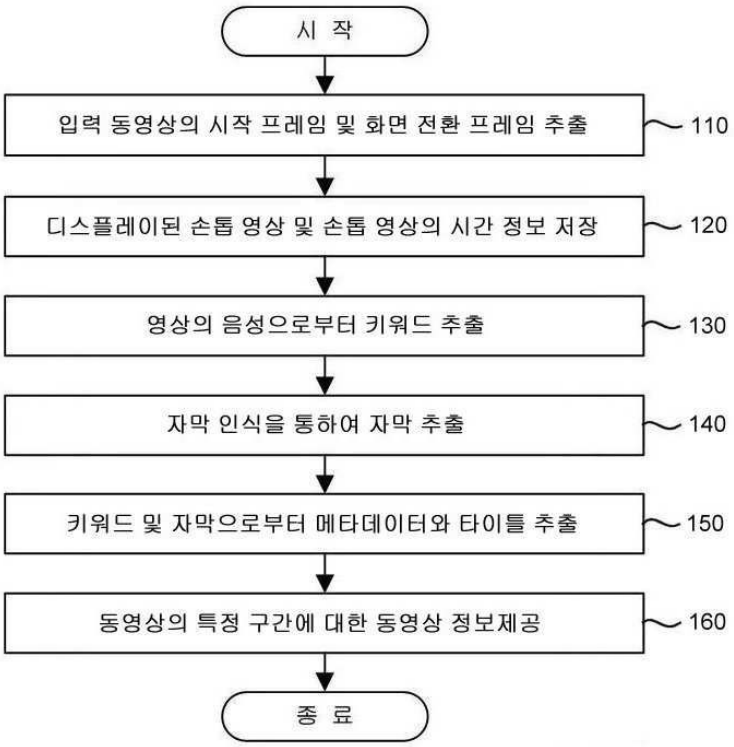
<7> 도 7은 본 발명의 다른 실시예에 따른 메타데이터를 이용한 동영상 탐색 방법의 흐름도이다.

<8> 도 8은 XML 문서로 저장되는 동영상 데이터의 그래픽 유저 인터페이스 환경을 도시한 것이다.

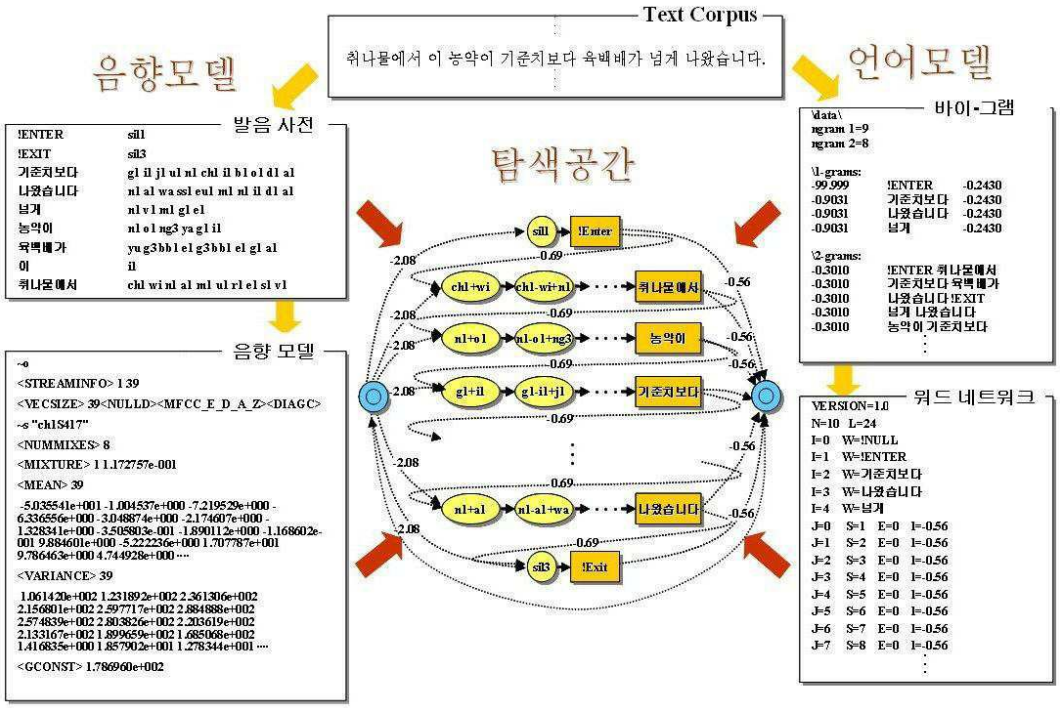
<9> 도 9는 XML 문서로 저장되는 동영상 데이터의 검색 윈도우를 도시한 것이다.

도면

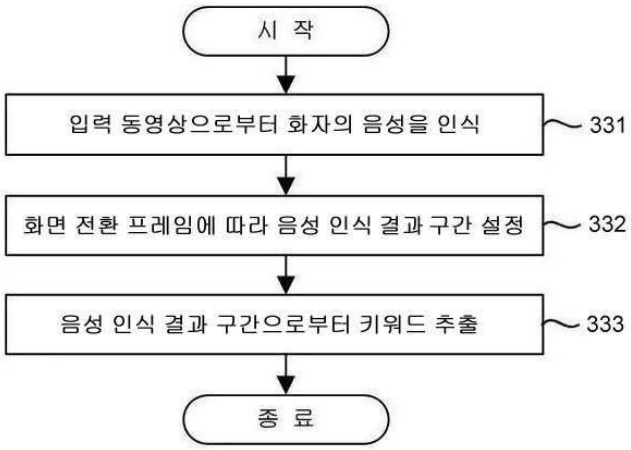
도면1



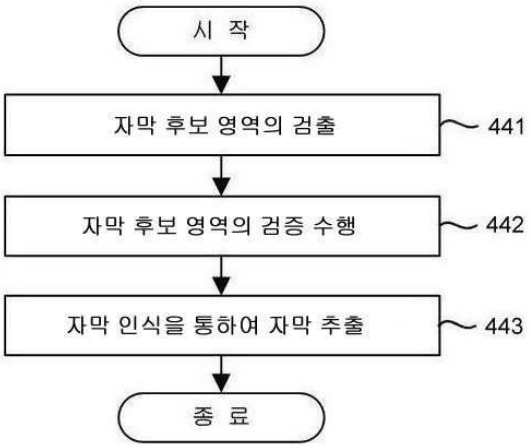
도면2


Text Corpus
취나물에서 이 농약이 기준치보다 육백배가 넘게 나왔습니다.
음향모델
언어모델
발음 사전
!ENTER sil
!EXIT sil3
기준치보다 gl il jl ul nl chl il b l o l dl al
나왔습니다 nl al wa ssl eul ml nl il dl al
넘게 nl v l ml gl el
농약이 nl o l ng3 ya g l il
육백배가 yu g3bb l el g3bb l el gl al
이 il
취나물에서 chl wi nl al ml ul r l el sl v l
음향 모델
탐색공간
바이-그램
\data\
ngram 1=9
ngram 2=8
\1-grams:
-99.999 !ENTER -0.2430
-0.9031 기준치보다 -0.2430
-0.9031 나왔습니다 -0.2430
-0.9031 넘게 -0.2430
\2-grams:
-0.3010 !ENTER 취나물에서
-0.3010 기준치보다 육백배가
-0.3010 나왔습니다 !EXIT
-0.3010 넘게 나왔습니다
-0.3010 농약이 기준치보다
워드 네트워크
VERSION=1.0
N=10 L=24
I=0 W=!NULL
I=1 W=!ENTER
I=2 W=기준치보다
I=3 W=나왔습니다
I=4 W=넘게
J=0 S=1 E=0 l=-0.56
J=1 S=2 E=0 l=-0.56
J=2 S=3 E=0 l=-0.56
J=3 S=4 E=0 l=-0.56
J=4 S=5 E=0 l=-0.56
J=5 S=6 E=0 l=-0.56
J=6 S=7 E=0 l=-0.56
J=7 S=8 E=0 l=-0.56
sil
!Enter
chl+wi
chl-wi+nl
취나물에서
nl+ol
nl-ol+ng3
농약이
gl+il
gl-il+jl
기준치보다
nl+al
nl-al+wa
나왔습니다
sil3
!Exit
-2.08
-0.69
-0.56

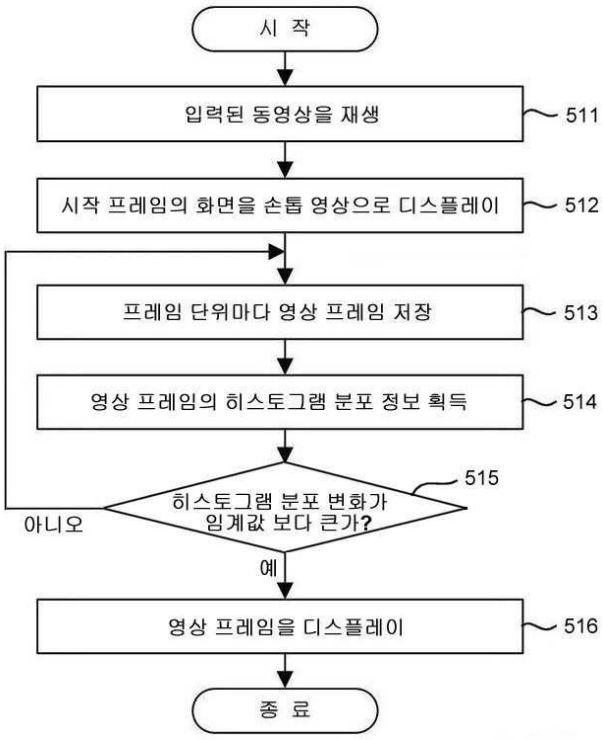
도면3


시 작
입력 동영상으로부터 화자의 음성을 인식
331
화면 전환 프레임에 따라 음성 인식 결과 구간 설정
332
음성 인식 결과 구간으로부터 키워드 추출
333
종 료

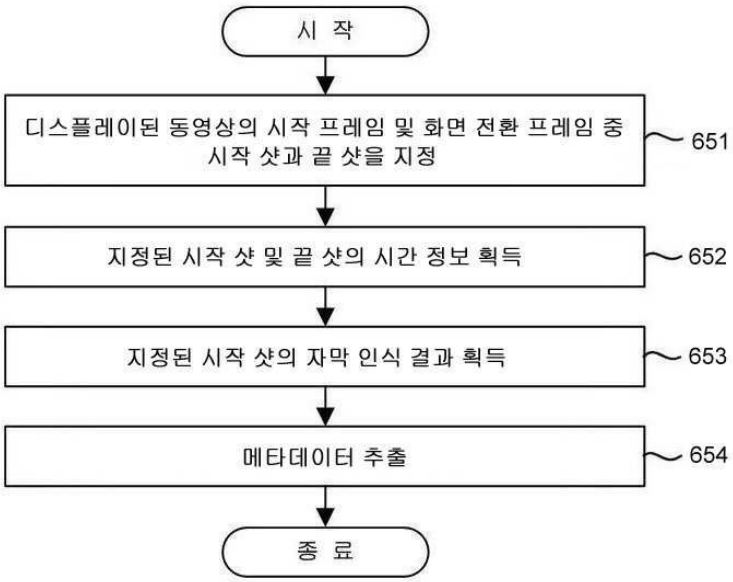
도면4


시 작
자막 후보 영역의 검출
441
자막 후보 영역의 검증 수행
442
자막 인식을 통하여 자막 추출
443
종 료

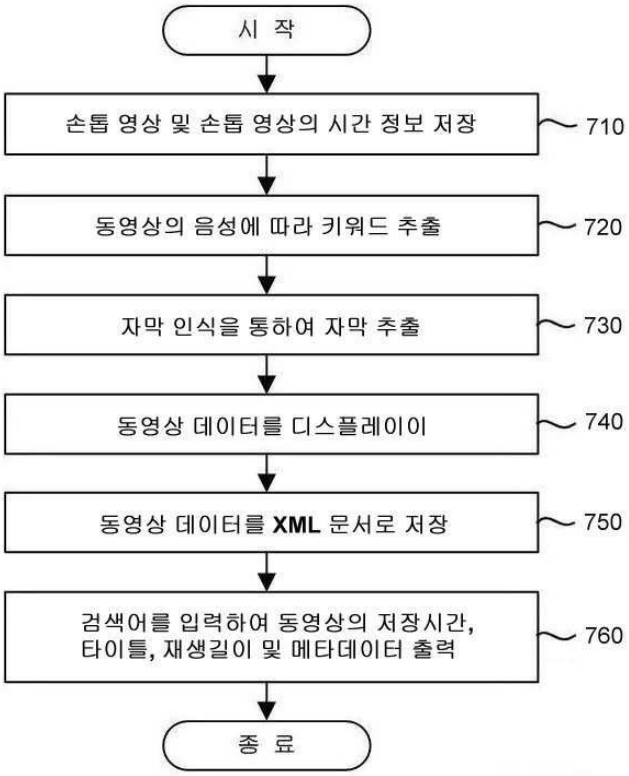
도면5


시 작
입력된 동영상을 재생
511
시작 프레임의 화면을 손톱 영상으로 디스플레이
512
프레임 단위마다 영상 프레임 저장
513
영상 프레임의 히스토그램 분포 정보 획득
514
히스토그램 분포 변화가 임계값 보다 큰가?
515
아니오
예
영상 프레임을 디스플레이
516
종 료

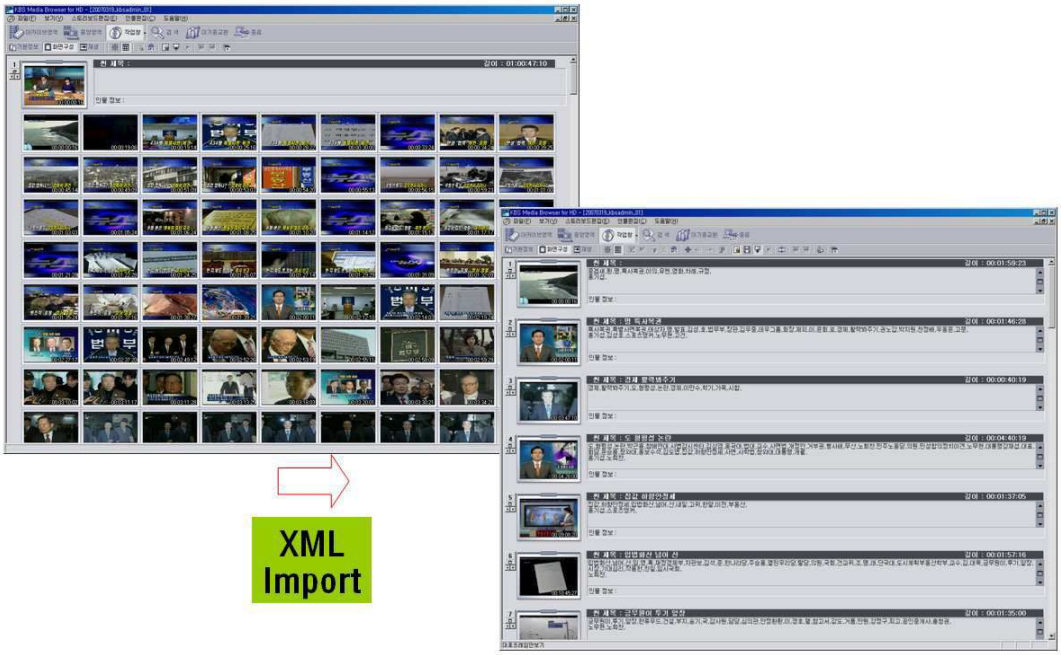
도면6


시 작
디스플레이된 동영상의 시작 프레임 및 화면 전환 프레임 중 시작 샷과 끝 샷을 지정
651
지정된 시작 샷 및 끝 샷의 시간 정보 획득
652
지정된 시작 샷의 자막 인식 결과 획득
653
메타데이터 추출
654
종 료

도면7


시 작
손톱 영상 및 손톱 영상의 시간 정보 저장
710
동영상의 음성에 따라 키워드 추출
720
자막 인식을 통하여 자막 추출
730
동영상 데이터를 디스플레이이
740
동영상 데이터를 XML 문서로 저장
750
검색어를 입력하여 동영상의 저장시간, 타이틀, 재생길이 및 메타데이터 출력
760
종 료

도면8


XML
Import

도면9


일반
고급
분류
검색어 투기
촬영일
2007-01-26 부터 2007-04-26 까지
자료종류
소재
프로그램
결과보기방법
정렬 촬영일
순서 내림차순
결과내 재검색
초기화
오늘영상
검색
2 긍무원이 투기 앞장
클립 ID
20070300135
저 장 소
중앙
길 이
00:01:35:00
내 용
긍무원이,투기,앞장,한류우드,건설,부지,송기,국,감사원,담당,심의관,안정환란,이,경호,랠,
참고서,값도,거품,만원,강정구,죄고,공인중개사,충청권,
노무현,노회찬,