딥러닝 기반 얼굴 비식별화 및 얼굴인식 기술

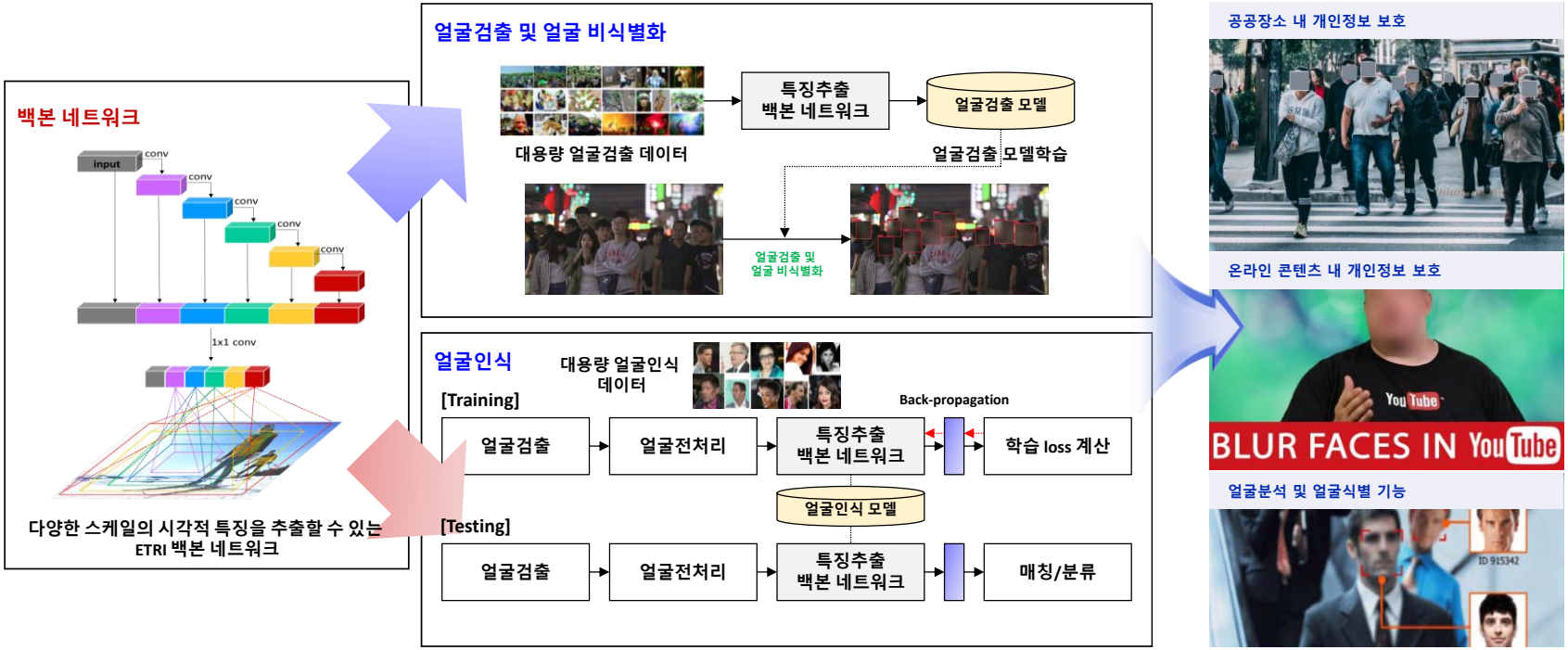
목 차

1. 기술의 개요
2. 기술이전 내용 및 범위
3. 경쟁기술과 비교
4. 기술의 사업성
 - 활용분야 및 기대효과
5. 국내외 시장 동향

1. 기술의 개요

▣ 영상 내 존재하는 얼굴 영역을 검출하고 얼굴 영상을 분석하기 위한 딥러닝 기반 기술로 얼굴검출, 얼굴 비식별화, 얼굴인식 기술로 구성됨

본 기술이전은 딥러닝 기술을 바탕으로 영상 내 존재하는 얼굴 영역을 검출하고 개인정보보호를 위해 해당 얼굴 영역의 비식별화 및 선별적 비식별화를 위한 얼굴인식 기술이 포함되며 얼굴 기반 콘텐츠 분석 및 콘텐츠 관리 응용, 출입통제 활용 등의 응용 서비스에 활용 가능함



2. 기술이전 내용 및 범위

▣ 기술이전 내용

- ❖ 영상 내 얼굴 영역을 찾기 위한 딥러닝 기반 얼굴검출 기술
- ❖ 검출된 얼굴 영역에 대한 얼굴 비식별화 기술
- ❖ 검출된 얼굴 영역에 대한 얼굴 특징추출 및 인식 기술

▣ 기술이전 범위

- ❖ 특허 (실시권)
- ❖ 관련 소스 코드 및 샘플 프로그램
- ❖ 학습된 모델
- ❖ 기술문서
- ❖ 시험 절차서 및 결과서

2. 기술이전 내용 및 범위

▣ 기술 개발 현황

❖ 기술성숙도(TRL : Technology Readiness Level) 단계 : (6)단계

기초 연구단계	1단계	기초 이론/실험	▪ 기초이론 정립 단계
	2단계	실용 목적의 아이디어, 특허 등 개념정립	▪ 기술개발 개념 정립 및 아이디어에 대한 특허 출원 단계
실험단계	3단계	실험실 규모의 기본성능 검증	▪ 실험실 환경에서 실험 또는 전산 시뮬레이션을 통해 기본성능이 검증될 수 있는 단계 ▪ 개발하려는 부품/시스템의 기본 설계도면을 확보하는 단계
	4단계	실험실 규모의 소재/부품/시스템 핵심성능 평가	▪ 시험샘플을 제작하여 핵심성능에 대한 평가가 완료된 단계 ▪ 3단계에서 도출된 다양한 결과 중에서 최적의 결과를 선택하려는 단계 ▪ 컴퓨터 모사가 가능한 경우 최적화를 완료하는 단계
시작품 단계	5단계	확정된 소재/부품/시스템시작품 제작 및 성능 평가	▪ 확정된 소재/부품/시스템의 실험실 시작품 제작 및 성능 평가가 완료된 단계 ▪ 개발 대상의 생산을 고려하여 설계하나 실제 제작한 시작품 샘플은 1~수개 미만인 단계 ▪ 경제성을 고려하지 않고 기술의 핵심성능으로만 볼 때, 실제로 판매가 될 수 있는 정도로 목표 성능을 달성한 단계
	6단계	파일롯 규모 시작품 제작 및 성능 평가	▪ 파일롯 규모(복수 개~양산규모의 1/10정도)의 시작품 제작 및 평가가 완료된 단계 ▪ 파일롯 규모 생산품에 대해 생산량, 생산용량, 불량률 등 제시 ▪ 파일롯 생산을 위한 대규모 투자가 동반되는 단계 ▪ 생산기업이 수요기업 적용환경에 유사하게 자체 현장테스트를 실시하여 목표 성능을 만족시킨 단계 ▪ 성능 평가 결과에 대해 가능하면 공인인증 기관의 성적서 확보
실용화 단계	7단계	신뢰성평가 및 수요기업 평가	▪ 실제 환경에서 성능 검증이 이루어지는 단계 ▪ 부품 및 소재개발의 경우 수요업체에서 직접 파일롯 시작품을 현장 평가(성능 및 신뢰성 평가) ▪ 가능하면 인증기관의 신뢰성 평가 결과 제출
	8단계	시제품 인증 및 표준화	▪ 표준화 및 인허가 취득 단계
사업화	9단계	사업화	▪ 본격적인 양산 및 사업화 단계 ▪ 6-시그마 등 품질관리가 중요한 단계

3. 경쟁기술과 비교

▣ 기술의 주요 특징

❖ **딥러닝 기반 얼굴검출 및 얼굴인식 기술**:
- 이미지 내에 등장하는 얼굴 영역을 검출하고 특징추출을 통한 인식 기술 포함
- 자원 효율적인 백본 네트워크가 적용되어 실시간 처리가 가능
- 공간적인 주목 (attention) 기법을 이용한 검출 및 인식 성능 개선 기술 적용

❖ **검출된 얼굴 영역에 대한 얼굴 비식별화 기술**:
- 검출된 얼굴 영역에 대해 개인 정보인 얼굴을 비식별화하는 기술
- 얼굴인식 기술을 활용한 선택적 비식별화 기능 구현 가능

4. 기술의 사업성

▣ 활용 분야

예상 제품 / 서비스	예상 수요자
얼굴 기반 콘텐츠 분석/처리	– 방송 콘텐츠 제작 및 사업자 – 데이터 수집 및 공개 업체
출입통제 시스템	– CCTV 분석 사업자 – 공공기관/지자체, 기업체 – 쇼핑 매장 및 문화센터 보유 사업자

▣ 기대 효과

❖ 본 기술은 딥러닝 기반의 얼굴 비식별화 및 얼굴인식 기술은 다양한 응용 분야에서 활용이 가능한 기술로서 사업화 시나리오에 따라 추가 모델 학습, 전달되는 모듈 기술을 바탕으로 새로운 파이프라인을 구성하는 등의 추가 기술 개발이 필요함

❖기대 활용처

1. 얼굴 기반 콘텐츠 분석 및 관리: 영상 정보 공유, 공공 데이터 제작, 또는 비디오 콘텐츠 제작 시 개인정보를 비식별화 처리할 수 있는 응용 활용, 얼굴인식 기반 장면 분할, 인물 별 콘텐츠 관리 활용 가능
2. 출입통제 활용: 얼굴검출 기능을 통해 특정 장소에 입장하는 사람 counting 기능, 특정 장소에 출입하는 인원 중 사전에 등록된 인물 여부를 확인하는 시나리오 활용 가능

5. 국내외 시장 동향

▣ 시장전망

❖ 얼굴인식 관련 세계 시장은 2015년 1,522백만달러에서 2020년 2,836백만달러로 연평균 13.3% 성장세를 전망, 국내 시장은 2015년 86,873백만원에서 2020년 151,417백만원으로 연평균 11.8%의 성장세 전망

(단위 : 백만달러, 백만원)

관련 제품/서비스	시장	1차년도 (2021)	2차년도 (2022)	3차년도 (2023)	4차년도 (2024)	5차년도 (2025)	합계
얼굴인식 시스템	해외	3,213	3,640	4,124	4,673	5,295	20,945
	국내	168,345	187,166	208,092	231,356	258,656	1,053,615

(출처: KISTI Market Report 2017-06 참조)

* 세계 시장은 CAGR 13.3%, 국내 시장은 CAGR 11.8%로 산정, (환율 $1 = 1,200원)

감사합니다.

www.etri.re.kr

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0119425
(43) 공개일자 2020년10월20일

(51) 국제특허분류(Int. Cl.)
G06K 9/00 (2006.01) *G06N 20/00* (2019.01)
(52) CPC특허분류
G06K 9/00228 (2013.01)
G06K 9/00268 (2013.01)
(21) 출원번호 10-2019-0038049
(22) 출원일자 2019년04월01일
심사청구일자 2019년12월05일

(71) 출원인
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
김형일
대전광역시 유성구 문지로299번길 108, 302호(문지동)
권용진
대전광역시 유성구 가정로 65, 108동 801호(신성동, 대림두레아파트)
(뒷면에 계속)
(74) 대리인
특허법인지명

전체 청구항 수 : 총 17 항

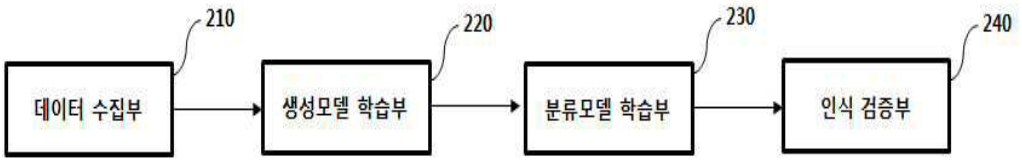
(54) 발명의 명칭 **도메인 적응 기반 객체 인식 장치 및 그 방법**

(57) 요 약

본 발명은 도메인 적응 기반의 객체 인식 장치 및 그 방법에 관한 것이다

본 발명에 따른 도메인 적응 기반 객체 인식 장치는 도메인 적응 기반 객체 인식 프로그램이 저장된 메모리 및 프로그램을 실행시키는 프로세서를 포함하되, 프로세서는 입력 프로브 영상에 대해 도메인 적응 기반으로 갤러리 영상과 유사한 영상 또는 특징으로 생성시키기 위한 생성모델을 학습하고, 갤러리 영상과 프로브 영상의 학습 데이터베이스를 이용하여 객체인식 분류 모델을 학습하여, 입력 프로브 영상을 이용한 객체 인식을 수행하는 것을 특징으로 한다.

대 표 도 - 도2



(52) CPC특허분류
G06K 9/00288 (2013.01)
G06N 20/00 (2019.01)
(72) 발명자
문진영
대전광역시 유성구 지족로 343, 206동 604호 (지족동, 반석마을아파트2단지)
박종열
대전광역시 중구 서문로 96, 203동 1503호 (문화동, 센트럴파크2단지아파트)
오성찬
서울특별시 송파구 백제고분로18길 30, 107동 303호(잠실동, 우성아파트)
윤기민
대전광역시 유성구 전민로26번길 14, 102호(전민동, 그레이스빌)
이전우
충청남도 계룡시 두마면 사계로 51, 103동 502호 (계룡대림e편한세상아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호	2014-3-00123
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원(IITP)
연구사업명	ICT융합산업원천기술개발사업
연구과제명	(딥뷰-1세부) 실시간 대규모 영상 데이터 이해 · 예측을 위한 고성능 비주얼 디스커버리 플랫폼 개발
기 여 율	1/1
과제수행기관명	한국전자통신연구원
연구기간	2018.01.01 ~ 2018.12.31

명 세 서

청구범위

청구항 1

도메인 적응 기반 객체 인식 프로그램이 저장된 메모리; 및

상기 프로그램을 실행시키는 프로세서를 포함하되,

상기 프로세서는 입력 프로브 영상에 대해 도메인 적응 기반으로 갤러리 영상과 유사한 영상 또는 특징으로 생성시키기 위한 생성모델을 학습하고, 갤러리 영상과 프로브 영상의 학습 데이터베이스를 이용하여 객체인식 분류 모델을 학습하여, 상기 입력 프로브 영상을 이용한 객체 인식을 수행하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 2

제1항에 있어서,

상기 프로세서는 객체의 특징 정보를 이용한 전처리를 수행하여 상기 학습 데이터베이스를 구축하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 3

제1항에 있어서,

상기 프로세서는 상기 학습 데이터베이스와 갤러리에 등록되지 않은 외부 영상 데이터베이스를 활용하여 전처리를 수행하여 상기 생성모델을 학습하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 4

제3항에 있어서,

상기 프로세서는 영상 소스를 분류하고, 도메인 적응 기반의 새로운 영상을 생성하고, 객체 ID를 판별하여 갤러리 영상의 스타일을 학습하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 5

제1항에 있어서,

상기 프로세서는 상기 학습 데이터베이스에 대해 전처리 수행, 특징 추출 수행에 따라 객체 ID 분류기를 학습시켜, 상기 객체인식 분류 모델을 학습하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 6

제1항에 있어서,

상기 프로세서는 수신된 입력 영상으로부터 객체 영역을 검출하고, 상기 생성모델을 이용하여 입력 영상을 갤러리 영상과 유사한 새로운 영상 또는 특징으로 생성하고, 생성된 새로운 영상에 대해 상기 객체인식 분류 모델을 이용한 특징 추출을 수행하여, 객체의 ID 정보를 획득하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 7

얼굴 영상을 수집하는 데이터 수집부;

갤러리 얼굴 영상의 스타일을 학습하는 생성모델 학습부;

얼굴 인식 및 매칭을 수행하기 위해 사전에 등록이 필요한 인물 정보를 이용하여 분류모델을 학습하는 분류모델 학습부; 및

생성모델 및 분류모델을 이용하여 실제 입력 얼굴 영상에 대한 인식을 수행하는 인식 검증부

를 포함하는 도메인 적응 기반 객체 인식 장치.

청구항 8

제7항에 있어서,

상기 데이터 수집부는 상기 얼굴 영상에 대해 특징점 정보를 이용하여 전처리를 수행하고, 갤러리 얼굴 영상 데이터베이스를 구축하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 9

제8항에 있어서,

상기 생성모델 학습부는 상기 갤러리 얼굴 영상 데이터베이스와 외부 얼굴 영상 데이터베이스를 이용하여 얼굴 영상 생성모델을 학습하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 10

제9항에 있어서,

상기 생성모델 학습부는 입력된 영상이 상기 갤러리 얼굴 영상 데이터베이스에 포함되는지 여부를 판별하고, 학습된 갤러리 얼굴 영상의 스타일과 유사하게 새로운 얼굴 영상을 생성하고, 입력된 영상의 ID를 판별하여 상기 얼굴 영상 생성모델을 학습하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 11

제8항에 있어서,

상기 분류모델 학습부는 상기 갤러리 얼굴 영상 데이터베이스를 이용한 전처리 및 특징 추출에 따른 얼굴 ID 분류 결과에 따라 오류 계산을 수행하여, 얼굴 ID 분류기를 학습시키는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 12

제7항에 있어서,

상기 인식 검증부는 비디오 입력으로부터 얻은 각 프레임으로부터 얼굴 영역을 검출하고, 상기 생성모델을 이용하여 입력된 얼굴 영상이 갤러리 얼굴 영상과 유사하도록 새로운 얼굴 영상을 생성하고, 상기 분류모델을 이용하여 특징 추출 및 매칭을 수행하여, ID 정보를 획득하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 13

(a) 객체 영상을 수집하는 단계;

(b) 갤러리 영상의 스타일을 학습하여 생성모델을 학습하는 단계;

(c) 객체 인식을 위해 사전에 등록이 필요한 정보를 이용하여 분류모델을 학습하는 단계; 및

(d) 상기 생성모델 및 분류모델을 이용하여 영상 내 객체를 인식하는 단계

를 포함하는 도메인 적응 기반 객체 인식 방법.

청구항 14

제13항에 있어서,

상기 (a) 단계는 특징점 정보를 이용하여 상기 객체 영상에 대한 전처리를 수행하고, 갤러리 영상 데이터베이스를 구축하는 것

인 도메인 적응 기반 객체 인식 방법.

청구항 15

제13항에 있어서,

상기 (b) 단계는 갤러리 영상 데이터베이스와 외부 영상 데이터베이스를 이용하여, 입력 영상을 상기 갤러리 영상의 스타일에 부합되는 새로운 영상 또는 특징으로 생성하기 위한 상기 생성모델을 학습하는 것

인 도메인 적응 기반 객체 인식 방법.

청구항 16

제13항에 있어서,

상기 (c) 단계는 갤러리 영상 데이터베이스를 이용하여 전처리 및 특징 추출을 수행하고, ID 분류 결과에 따른 오류 계산을 수행하여 ID 분류기를 학습시키는 것

인 도메인 적응 기반 객체 인식 방법.

청구항 17

제13항에 있어서,

상기 (d) 단계는 비디오 입력으로부터 얻은 각 프레임으로부터 객체 영역을 검출하고, 상기 생성모델을 이용하여 객체가 상기 갤러리 영상과 유사하도록 새로운 영상 또는 특징을 생성하고, 상기 분류모델을 이용하여 특징을 추출하고 매칭을 수행하여, 객체의 ID 정보를 획득하는 것

인 도메인 적응 기반 객체 인식 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 도메인 적응 기반의 객체 인식 장치 및 그 방법에 관한 것이다.

배 경 기 술

[0003] 종래의 객체 인식 기술은 객체 검출, 전처리, 특징 추출, 인식/매칭의 과정을 통해 수행된다.

[0004] 객체 인식 기술은 사전에 등록된 정보를 기반으로 현재 입력되는 정보를 인식하게 되는데, 다양한 환경 변화를 보상시키기 위한 전처리 또는 환경 변화에 강인한 특징 추출 기법이 제안되었으나, 실제 발생되는 모든 변화를 다룰 수 없는 한계점이 있고, 강인한 특징 추출 학습을 위해 대량의 데이터가 요구되는 문제점이 있다.

발명의 내용

해결하려는 과제

[0006] 본 발명은 전술한 문제점을 해결하기 위하여 제안된 것으로, 제한된 집합의 갤러리 영상과 프로브 영상을 이용하여 갤러리 영상 또는 특징의 스타일을 학습하고, 프로프 영상을 도메인 적응을 통해 갤러리 영상의 스타일과 유사한 새로운 영상 또는 특징을 생성함으로써, 외부 환경 변화로부터 강인한 객체 인식이 가능한 장치 및 방법을 제공하는데 그 목적이 있다.

과제의 해결 수단

[0008] 본 발명에 따른 도메인 적응 기반 객체 인식 장치는 도메인 적응 기반 객체 인식 프로그램이 저장된 메모리 및 프로그램을 실행시키는 프로세서를 포함하되, 프로세서는 입력 프로브 영상에 대해 도메인 적응 기반으로 갤러리 영상과 유사한 영상 또는 특징으로 생성시키기 위한 생성모델을 학습하고, 갤러리 영상과 프로브 영상의 학습 데이터베이스를 이용하여 객체인식 분류 모델을 학습하여, 입력 프로브 영상을 이용한 객체 인식을 수행하는 것을 특징으로 한다.

[0009] 본 발명에 따른 도메인 적응 기반 객체 인식 장치는 얼굴 영상을 수집하는 데이터 수집부와, 갤러리 얼굴 영상의 스타일을 학습하는 생성모델 학습부와, 얼굴 인식 및 매칭을 수행하기 위해 사전에 등록이 필요한 인물 정보를 이용하여 분류모델을 학습하는 분류모델 학습부 및 생성모델 및 분류모델을 이용하여 실제 입력 얼굴 영상에 대한 인식을 수행하는 인식 검증부를 포함하는 것을 특징으로 한다.

[0010] 본 발명에 따른 도메인 적응 기반 객체 인식 방법은 객체 영상을 수집하는 단계와, 갤러리 영상의 스타일을 학습하여 생성모델을 학습하는 단계와, 객체 인식을 위해 사전에 등록이 필요한 정보를 이용하여 분류모델을 학습하는 단계 및 생성모델과 분류모델을 이용하여 영상 내 객체를 인식하는 단계를 포함하는 것을 특징으로 한다.

발명의 효과

[0012] 본 발명의 실시예에 따르면, 갤러리 얼굴 영상과 프로브 얼굴 영상 사이의 차이가 큰 신분증(주민등록증, 여권 등) 인식, 출입통제 시스템 등에 적용되어, 제약된 환경에서 촬영된 갤러리 얼굴 영상과 다양한 변화를 갖는 얼굴 영상을 이용하여 갤러리 얼굴 영상의 스타일을 학습하고, 프로브 얼굴 영상 입력을 학습 모델에 의해 새로운 영상(갤러리 영상의 스타일과 유사한 영상) 또는 특징으로 생성함으로써, 갤러리 및 프로브 얼굴 영상 사이의

불일치를 줄이고 외부 환경 변화로부터 강인한 얼굴인식 수행이 가능한 효과가 있다.

[0013] 본 발명에 따르면 얼굴영상 생성모델 학습과 얼굴인식 분류모델 학습을 동시에 사용함으로써, 외부 환경 변화로부터 강인한 얼굴 인식 수행의 신뢰성을 높이는 것이 가능한 효과가 있다.

[0014] 본 발명의 효과는 이상에서 언급한 것들에 한정되지 않으며, 언급되지 아니한 다른 효과들은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0016] 도 1 및 도 2는 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 장치를 나타내는 블록도이다.

도 3은 본 발명의 실시예에 따른 데이터 수집부를 나타내는 블록도이다.

도 4는 본 발명의 실시예에 따른 생성모델 학습부를 나타내는 블록도이다.

도 5는 본 발명의 실시예에 따른 분류모델 학습부를 나타내는 블록도이다.

도 6은 본 발명의 실시예에 따른 인식 검증부를 나타내는 블록도이다.

도 7은 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 방법을 나타내는 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0017] 본 발명의 전술한 목적 및 그 이외의 목적과 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다.

[0018] 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 이하의 실시예들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 목적, 구성 및 효과를 용이하게 알려주기 위해 제공되는 것일 뿐으로서, 본 발명의 권리범위는 청구항의 기재에 의해 정의된다.

[0019] 한편, 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 및/또는 "포함하는(comprising)"은 언급된 구성소자, 단계, 동작 및/또는 소자가 하나 이상의 다른 구성소자, 단계, 동작 및/또는 소자의 존재 또는 추가됨을 배제하지 않는다.

[0021] 이하에서는, 당업자의 이해를 돕기 위하여 본 발명이 제안된 배경에 대하여 먼저 서술하고, 본 발명의 실시예에 대하여 서술하기로 한다.

[0022] 종래 기술에 따른 얼굴인식 기술은 얼굴검출, 전처리(preprocessing), 특징추출, 인식 또는 매칭의 과정을 통해 수행된다.

[0023] 이러한 얼굴인식 기술은 사전에 등록된 갤러리 얼굴 영상과 실제 입력으로 들어오는 프로브 얼굴 영상을 비교하여 인물 정보를 인식하는 기술과, 입력으로 두 장의 영상이 들어왔을 때 동일 인물인지 여부를 판단하는 얼굴검증 기술로 분류된다.

[0024] 이러한 환경에서 사전에 등록된 갤러리 얼굴 영상 정보들은 상대적으로 제약된 환경(고정된 조도, 촬영위치 등)에서 촬영된 반면, 입력으로 들어오는 프로브 얼굴 영상의 경우에는 조도변화, 포즈변화, 저해상도 등 다양한 환경에서 취득되어 열화된(degraded) 영상이 입력된다.

[0025] 종래 기술에 따르면, 이러한 환경에서 효과적인 얼굴인식을 수행하기 위해 다양한 환경 변화를 보상 시키기 위한 전처리 기술(조명보정 및 필터링, 포즈보정, 초 해상화 등), 환경 변화에 강인한 특징 추출 기법 등이 주로 개발되어 왔다.

[0026] 하지만, 전처리 기술을 통해서는 실제 발생하는 모든 변화를 다룰 수 없고, 전처리 알고리즘은 실험적으로(heuristically) 고안된 것으로 모든 문제를 자동적으로 탐지하여 보정하는데 한계가 있다.

[0027] 또한, 환경 변화에 강인한 특징을 추출하기 위해 심층 학습(deep learning) 기반 방법이 개발되고 있으나, 학습

에 사용되는 얼굴 영상들을 갤러리 영상과 비교할 때, 스타일의 차이가 존재하며, 강인한 특징 추출기를 학습시키기 위해서는 다양한 변화를 포함하는 대량의 데이터가 요구되는 문제점이 있다.

[0029] 본 발명은 전술한 문제점을 해결하기 위하여 제안된 것으로, 스마트 관제 또는 출입통제 시스템에서 얼굴인식 수행 시에 사전에 등록된 갤러리(gallery) 얼굴 영상들을 이용한 학습을 통해, 실세계에서 취득된 다양한 변화를 갖게 되는 프로브(probe) 얼굴 영상의 도메인 적응(domain adaptation)을 통해 갤러리 얼굴 영상 스타일과 유사한 새로운 영상 또는 특징을 생성시킴으로써, 갤러리 영상과 프로브 영상 사이의 불일치(mismatch) 문제를 줄이고, 효과적인 인식/매칭(matching)을 수행하는 것이 가능한 도메인 적응 기반 객체 인식 장치 및 그 방법을 제안한다.

[0030] 본 발명의 실시예에 따르면, 제한된 집합의 갤러리 얼굴 영상과 프로브 얼굴 영상을 이용하여 갤러리 얼굴 영상의 스타일을 학습하고, 학습된 모델을 이용하여 프로브 얼굴 영상의 도메인 적응을 통해, 갤러리 얼굴 영상들의 스타일과 유사한 새로운 영상 또는 특징을 생성한다.

[0031] 도메인 적응은 복수의 도메인이 존재할 때 서로 다른 도메인과 유사한 데이터를 생성하거나, 특정 도메인에서 학습된 모델이 다른 도메인에서 사용될 때 효과적으로 작동하게 하는 기술이다.

[0032] 본 발명의 실시예에 따르면, 생성된 프로브 얼굴 영상과 갤러리 얼굴 영상의 특징 추출에 따라 얼굴인식을 수행하게 되며, 갤러리 얼굴 영상 및 프로브 얼굴 영상 사이의 불일치를 줄임으로써 효과적인 얼굴인식이 가능하다.

[0034] 도 1은 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 장치를 나타내는 블록도이다.

[0035] 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 장치는 도메인 적응 기반 객체 인식 프로그램이 저장된 메모리(100) 및 프로그램을 실행시키는 프로세서(200)를 포함하되, 프로세서(200)는 입력 프로브 영상에 대해 도메인 적응 기반으로 갤러리 영상과 유사한 영상 또는 특징으로 생성시키기 위한 생성모델을 학습하고, 갤러리 영상과 프로브 영상의 학습 데이터베이스를 이용하여 객체인식 분류 모델을 학습하여, 입력 프로브 영상을 이용한 객체 인식을 수행하는 것을 특징으로 한다.

[0036] 프로세서(200)는 객체의 특징 정보를 이용한 전처리를 수행하여 학습 데이터베이스를 구축하고, 갤러리 영상 데이터베이스와 갤러리에 등록되지 않은 외부 영상 데이터베이스를 활용하여 전처리를 수행한 결과에 따라 생성모델을 학습한다.

[0037] 프로세서(200)는 입력되는 영상이 학습 데이터베이스에 포함되었는지 여부를 판별하여 영상 소스를 분류하고, 도메인 적응 기반의 새로운 영상 또는 특징을 생성하며, 객체 ID를 판별하여 갤러리 영상의 스타일을 학습한다.

[0038] 프로세서(200)는 학습 데이터베이스에 대해 전처리 수행, 특징 추출 수행에 따라 객체 ID 분류기를 학습시켜, 객체인식 분류 모델을 학습한다.

[0039] 이 때, 객체 ID 분류기를 통해 출력된 결과에 대해 오류 계산을 수행하여, 객체 ID 분류기를 학습시키게 된다.

[0040] 프로세서(200)는 수신된 입력 영상으로부터 객체 영역을 검출하고, 생성모델을 이용하여 입력 영상을 갤러리 영상과 유사한 새로운 영상 또는 특징으로 생성하고, 객체인식 분류 모델을 이용한 특징 추출을 수행하여, 객체의 ID 정보를 획득한다.

[0042] 도 2는 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 장치를 나타내는 블록도이다.

[0043] 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 장치는 모델 학습과 분류 및 매칭에 필요한 얼굴 영상들을 수집하는 데이터 수집부(210)와, 갤러리 얼굴 영상의 스타일을 학습하여 입력 프로브 영상에 대해 도메인 적응을 통해 갤러리 얼굴 영상과 유사한 새로운 얼굴 영상을 생성시키기 위한 생성모델 학습부(220)와, 얼굴 인식 및 매칭을 수행하기 위해 사전에 등록이 필요한 인물 정보를 이용하여 분류모델을 학습하는 분류모델 학습부(230) 및 생성모델과 분류모델을 이용하여 실제 입력 얼굴 영상에 대한 인식을 수행하는 인식 검증부(240)를 포함한다.

[0045] 도 3은 본 발명의 실시예에 따른 데이터 수집부를 나타내는 블록도이다.

[0046] 본 발명의 실시예에 따른 데이터 수집부는 얼굴 영상에 대해 특징점 정보를 이용하여 전처리를 수행하고, 갤러리 얼굴 영상 데이터베이스를 구축한다.

[0047] 도 3을 참조하면, 데이터 수집부는 얼굴 검출기(211), 전처리기(212)를 포함한다.

[0048] 얼굴 검출기(211)는 입력 영상(I)에 대해 얼굴이 존재하는 영역을 검출하고, 전처리기(212)는 검출된 얼굴 영상에 대해 얼굴의 특징점(feature point) 정보를 이용한 얼굴 정렬 또는 밝기 값 정규화와 같은 전처리를 수행하여 갤러리 얼굴 영상 데이터베이스(213)를 구축한다.

[0049] 데이터 수집부는 얼굴인식을 위해 사전에 등록시킬 인물에 대해 오프라인으로 촬영을 통해 갤러리 얼굴 영상 데이터베이스를 구축하거나, 추가적으로 훈련에 필요한 영상을 웹으로부터 확보하는 것이 가능하다.

[0051] 도 4는 본 발명의 실시예에 따른 생성모델 학습부를 나타내는 블록도이다.

[0052] 생성모델 학습부는 갤러리 얼굴 영상 데이터베이스(213)와 외부 얼굴 영상 데이터베이스(214)를 이용하여 얼굴 영상 생성모델을 학습하며, 전처리기(221), 영상소스 분류기(222), 얼굴영상 생성기(223), 얼굴 ID 분류기(224), 오류 계산 및 학습기(225)를 포함한다.

[0053] 생성모델 학습부는 입력된 영상이 갤러리 얼굴 영상 데이터베이스에 포함되는지 여부를 판별하고, 학습된 갤러리 얼굴 영상의 스타일과 유사하게 새로운 얼굴 영상을 생성하고, 입력된 영상의 ID를 판별하여 얼굴 영상 생성모델을 학습한다.

[0054] 전처리기(221)는 사전에 구축된 갤러리 얼굴 영상 데이터베이스(213)와 갤러리에 등록된 얼굴이 아닌 외부 얼굴 영상 데이터베이스(214)를 활용하여 전처리(픽셀 값 정규화, 영상 크기 정규화 등)를 수행한다.

[0055] 본 발명의 실시예에 따르면, 얼굴영상 생성모델(226)은 generative adversarial network 학습 방식으로 학습되는데, 이 모델을 학습하기 위해 영상소스 분류기(222), 얼굴영상 생성기(223), 얼굴 ID 분류기(224)의 3가지 모델이 동시에 학습된다.

[0056] 영상소스 분류기(222)는 입력으로 들어오는 영상이 갤러리 영상 데이터베이스에 포함되는지 여부를 판별하고, 얼굴영상 생성기(223)는 새로운 영상 또는 특징을 생성하는 모델이 되며, 얼굴 ID 분류기(224)는 입력 얼굴 영상의 ID를 판별한다.

[0057] 얼굴 ID 분류기(224)는 입력 얼굴 영상의 ID를 판별함으로써, 얼굴영상 생성 시 ID를 유지하면서도 스타일이 비슷한 영상으로 생성하도록 만드는 것이다.

[0058] 오류 계산 및 학습기(225)는 전술한 3가지의 모델을 통해 출력되는 결과로부터 오류를 계산하고, 반복적으로 학습을 수행하여, 영상소스 분류기(222) 학습을 통해 갤러리 얼굴 영상의 스타일을 학습하는 것과 동시에, 자신의 ID 정보는 잃지 않는 얼굴영상 생성모델(226)을 학습한다.

[0060] 도 5는 본 발명의 실시예에 따른 분류모델 학습부를 나타내는 블록도이다.

[0061] 본 발명의 실시예에 따른 분류모델 학습부는 갤러리 얼굴 영상 데이터베이스(213)를 이용한 전처리 및 특징 추출에 따른 얼굴 ID 분류 결과에 따라 오류 계산을 수행하여, 얼굴 ID 분류기(233)를 학습시키고, 사전에 등록이 필요한 인물 정보를 이용하여 얼굴인식 분류모델(235)을 학습한다.

[0062] 전처리기(231)는 사전에 수집된 갤러리 얼굴 영상 데이터베이스(213)에 대해 전처리를 수행하고, 특징 추출기(232)의 특징추출 후에 얼굴 ID 분류기(233)를 통해 나온 출력을 이용하여, 오류 계산을 통해 얼굴 ID 분류기(233)를 학습시키게 된다.

[0063] 이 때, 딥 네트워크(deep network)의 경우에는 특징 추출기(232) 및 얼굴 ID 분류기(233) 모두 신경망으로 구성되며, 초기 값은 대용량의 얼굴 데이터로 학습된 백본 네트워크(예: VGG Face)를 이용하여 세팅된다.

[0065] 도 6은 본 발명의 실시예에 따른 인식 검증부를 나타내는 블록도이다.

[0066] 인식 검증부의 얼굴 검출기는 비디오 입력으로부터 얻은 각 프레임으로부터 얼굴 영역을 검출하고, 얼굴영상 생성기(243)는 얼굴영상 생성모델(226)을 이용하여 입력된 얼굴 영상이 갤러리 얼굴 영상과 유사하도록 새로운 얼굴 영상을 생성하고, 특징추출 및 매칭기(244)는 얼굴인식 분류모델(235)을 이용하여 특징 추출 및 매칭을 수행하여, ID 정보(245)를 획득한다.

[0068] 도 7은 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 방법을 나타내는 순서도이다.

[0069] 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 방법은 객체 영상을 수집하는 단계(S710)와, 갤러리 영상의 스타일을 학습하여 생성모델을 학습하는 단계(S720)와, 객체 인식을 위해 사전에 등록이 필요한 정보를 이용하여 분류모델을 학습하는 단계(S730) 및 생성모델과 분류모델을 이용하여 영상 내 객체를 인식하는 단계(S740)를 포함한다

[0070] S710 단계는 특징점 정보를 이용하여 객체 영상에 대한 전처리를 수행하고, 갤러리 영상 데이터베이스를 구축한다

[0071] S720 단계는 갤러리 영상 데이터베이스와 외부 영상 데이터베이스를 이용하여, 입력 영상을 갤러리 영상의 스타일에 부합되는 새로운 영상 또는 특징으로 생성하기 위한 생성모델을 학습한다

[0072] S730 단계는 갤러리 영상 데이터베이스를 이용하여 전처리 및 특징 추출을 수행하고, ID 분류 결과에 따른 오류 계산을 수행하여 ID 분류기를 학습시킨다.

[0073] S740단계는 비디오 입력으로부터 얻은 각 프레임으로부터 객체 영역을 검출하고, 생성모델을 이용하여 객체가 갤러리 영상과 유사하도록 새로운 영상 또는 특징을 생성하고, 분류모델을 이용하여 특징을 추출하고 매칭을 수행하여, 객체의 ID 정보를 획득한다.

[0075] 한편, 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 방법은 컴퓨터 시스템에서 구현되거나, 또는 기록매체에 기록될 수 있다. 컴퓨터 시스템은 적어도 하나 이상의 프로세서와, 메모리와, 사용자 입력 장치와, 데이터 통신 버스와, 사용자 출력 장치와, 저장소를 포함할 수 있다. 전술한 각각의 구성 요소는 데이터 통신 버스를 통해 데이터 통신을 한다.

[0076] 컴퓨터 시스템은 네트워크에 커플링된 네트워크 인터페이스를 더 포함할 수 있다. 프로세서는 중앙처리 장치(central processing unit (CPU))이거나, 혹은 메모리 및/또는 저장소에 저장된 명령어를 처리하는 반도체 장치일 수 있다.

[0077] 메모리 및 저장소는 다양한 형태의 휘발성 혹은 비휘발성 저장매체를 포함할 수 있다. 예컨대, 메모리는 ROM 및 RAM을 포함할 수 있다.

[0078] 따라서, 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 방법은 컴퓨터에서 실행 가능한 방법으로 구현될 수 있다. 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 방법이 컴퓨터 장치에서 수행될 때, 컴퓨터로 판독 가능한 명령어들이 본 발명에 따른 객체 인식 방법을 수행할 수 있다.

[0079] 한편, 상술한 본 발명에 따른 도메인 적응 기반 객체 인식 방법은 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현되는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록 매체로는 컴퓨터 시스템에 의하여 해독될 수 있는 데이터가 저장된 모든 종류의 기록 매체를 포함한다. 예를 들어, ROM(Read Only Memory), RAM(Random Access Memory), 자기 테이프, 자기 디스크, 플래시 메모리, 광 데이터 저장장치 등이 있을 수 있다. 또한, 컴퓨터로 판독 가능한 기록매체는 컴퓨터 통신망으로 연결된 컴퓨터 시스템에 분산되어, 분산방식으로 읽을 수 있는 코드로서 저장되고 실행될 수 있다.

[0081] 이제까지 본 발명의 실시예들을 중심으로 살펴보았다. 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 개시된 실시예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

부호의 설명

[0083] 100: 메모리 200: 프로세서

210: 데이터 수집부 211: 얼굴 검출기

212: 전처리기 213: 갤러리 얼굴 영상 DB

214: 외부 얼굴 영상 DB 220: 생성모델 학습부

221: 전처리기 222: 영상소스 분류기

223: 얼굴영상 생성기 224: 얼굴 ID 분류기

225: 오류 계산 및 학습기 226: 얼굴영상 생성모델

230: 분류모델 학습부 231: 전처리기

232: 특징 추출기 233: 얼굴 ID 분류기

234: 오류 계산 및 학습기 235: 얼굴인식 분류모델

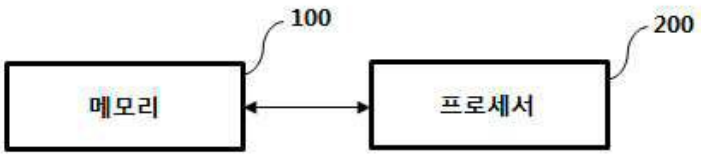
240: 인식 검증부 241: 얼굴 검출기

242: 전처리기 243: 얼굴영상 생성기

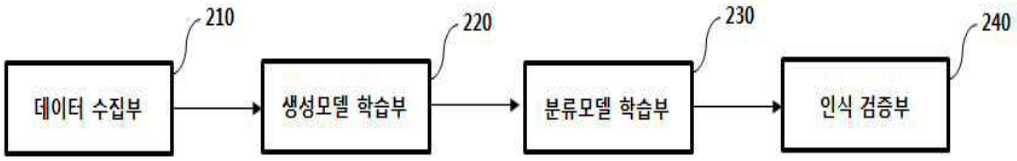
244: 특징추출 및 매칭기 245: ID 정보

도면

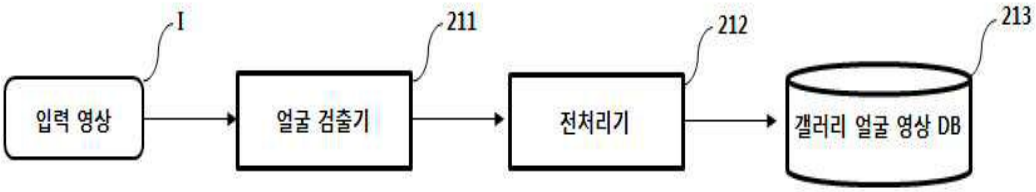
도면1



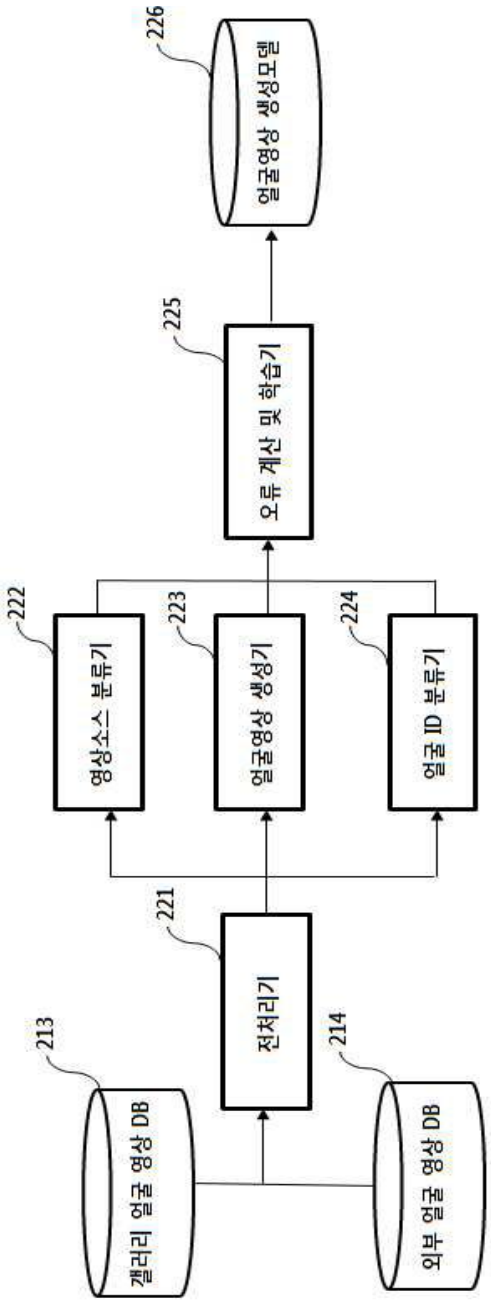
도면2



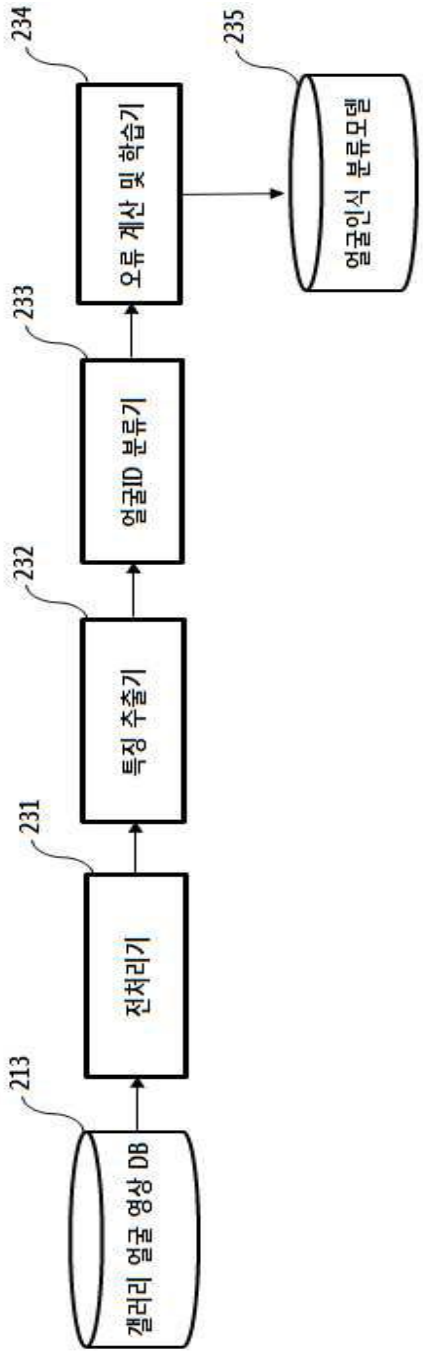
도면3


I
211
212
213
입력 영상
얼굴 검출기
전처리기
갤러리 얼굴 영상 DB

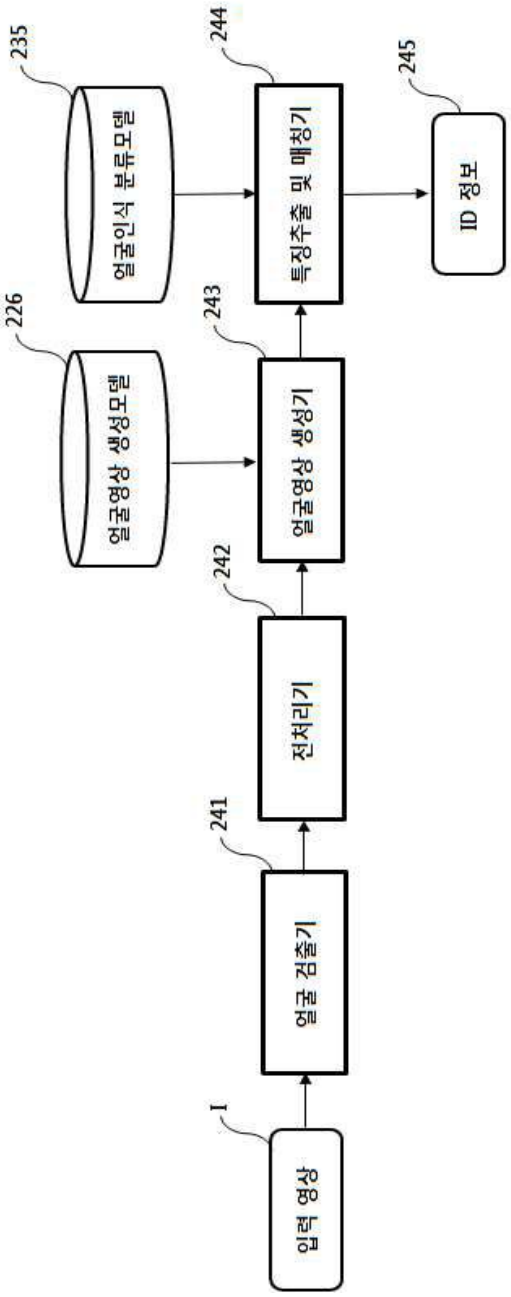
도면4


226
얼굴영상 생성모델
225
오류 계산 및 학습기
222
영상소스 분류기
223
얼굴영상 생성기
224
얼굴 ID 분류기
221
전처리기
213
갤러리 얼굴 영상 DB
214
외부 얼굴 영상 DB

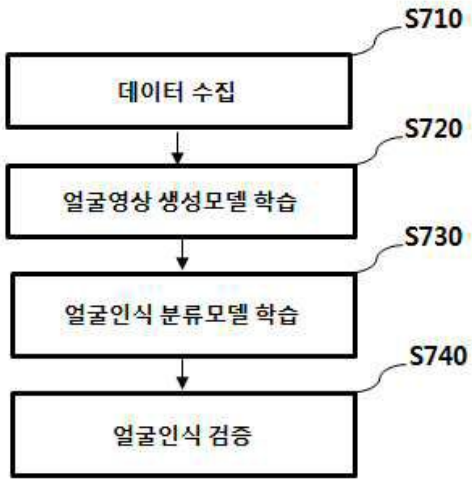
도면5


213
갤러리 얼굴 영상 DB
231
전처리기
232
특징 추출기
233
얼굴ID 분류기
234
오류 계산 및 학습기
235
얼굴인식 분류모델

도면6


235
244
245
얼굴인식 분류모델
특징추출 및 매칭기
ID 정보
226
243
얼굴영상 생성모델
얼굴영상 생성기
242
전처리기
241
얼굴 검출기
I
입력 영상

도면7



딥러닝 기반의 사람 상태 이해 기술

목 차

1. 기술의 개요
2. 기술이전 내용 및 범위
3. 경쟁기술과 비교
4. 기술의 사업성
 - 활용분야 및 기대효과
5. 국내외 시장 동향

1. 기술의 개요

▣ 딥러닝 기반의 사람 상태 이해 기술의 핵심인 객체탐지 및 분할, 사람 관절 추정 기술, 쓰러진 사람 탐지 기술로 구성됨

본 기술이전은 객체 탐지 및 객체 분할 기술, 사람 관절 추정 기술, 쓰러진 사람 탐지 기술로 구성되며, 기술을 활용하는 방법에 따라 지능형 감시 시스템 또는 객체 인식 기반 응용 서비스 등으로 활용 가능함

< 딥러닝 기반의 사람 상태 이해 기술 >

< 핵심기술을 활용한 응용 서비스 >

2. 기술이전 내용 및 범위

▣ 기술이전 내용

- ❖ 객체 탐지 및 분할 기술
- ❖ 사람 관절 추정 기술
- ❖ 쓰러진 사람 탐지 기술

▣ 기술이전 범위

- ❖ 특허 (실시권)
- ❖ 관련 소스 코드 및 샘플 프로그램
- ❖ 학습된 모델
- ❖ 기술문서
- ❖ 시험 절차서 및 결과서

2. 기술이전 내용 및 범위

▣ 기술 개발 현황

❖ 기술성숙도(TRL : Technology Readiness Level) 단계 : (6)단계

기초 연구단계	1단계	기초 이론/실험	▪ 기초이론 정립 단계
	2단계	실용 목적의 아이디어, 특허 등 개념정립	▪ 기술개발 개념 정립 및 아이디어에 대한 특허 출원 단계
실험단계	3단계	실험실 규모의 기본성능 검증	▪ 실험실 환경에서 실험 또는 전산 시뮬레이션을 통해 기본성능이 검증될 수 있는 단계 ▪ 개발하려는 부품/시스템의 기본 설계도면을 확보하는 단계
	4단계	실험실 규모의 소재/부품/시스템 핵심성능 평가	▪ 시험샘플을 제작하여 핵심성능에 대한 평가가 완료된 단계 ▪ 3단계에서 도출된 다양한 결과 중에서 최적의 결과를 선택하려는 단계 ▪ 컴퓨터 모사가 가능한 경우 최적화를 완료하는 단계
시작품 단계	5단계	확정된 소재/부품/시스템시작품 제작 및 성능 평가	▪ 확정된 소재/부품/시스템의 실험실 시작품 제작 및 성능 평가가 완료된 단계 ▪ 개발 대상의 생산을 고려하여 설계하나 실제 제작한 시작품 샘플은 1~수개 미만인 단계 ▪ 경제성을 고려하지 않고 기술의 핵심성능으로만 볼 때, 실제로 판매가 될 수 있는 정도로 목표 성능을 달성한 단계
	6단계	파일롯 규모 시작품 제작 및 성능 평가	▪ 파일롯 규모(복수 개~양산규모의 1/10정도)의 시작품 제작 및 평가가 완료된 단계 ▪ 파일롯 규모 생산품에 대해 생산량, 생산용량, 불량률 등 제시 ▪ 파일롯 생산을 위한 대규모 투자가 동반되는 단계 ▪ 생산기업이 수요기업 적용환경에 유사하게 자체 현장테스트를 실시하여 목표 성능을 만족시킨 단계 ▪ 성능 평가 결과에 대해 가능하면 공인인증 기관의 성적서 확보
실용화 단계	7단계	신뢰성평가 및 수요기업 평가	▪ 실제 환경에서 성능 검증이 이루어지는 단계 ▪ 부품 및 소재개발의 경우 수요업체에서 직접 파일롯 시작품을 현장 평가(성능 및 신뢰성 평가) ▪ 가능하면 인증기관의 신뢰성 평가 결과 제출
	8단계	시제품 인증 및 표준화	▪ 표준화 및 인허가 취득 단계
사업화	9단계	사업화	▪ 본격적인 양산 및 사업화 단계 ▪ 6-시그마 등 품질관리가 중요한 단계

3. 경쟁기술과 비교

▣ 기술의 주요 특징

❖ **객체 탐지 및 분할 기술**:
- 이미지/동영상에서 등장하는 사물의 위치와 종류(80종)를 빠르게 인식하는 기술
- 객체 탐지는 효율적인 백본 네트워크가 적용되어 실시간 처리가 가능
- 공간적인 주목 (attention) 기법을 이용한 분할 성능 개선 기술 적용

❖ **사람 관절 추정 기술**:
- 객체 탐지 기술을 확장하여 사람의 자세를 인식하기 위해 활용되는 기술
- 사람 위치 탐지 후, 이를 기준으로 각 관절 위치와 신뢰도를 추정하는 기술

❖ **쓰러진 사람 탐지 기술:**
- 객체 탐지된 사람과 분할 결과, 사람 관절 추정 결과를 종합적으로 활용하여 사람의 상태를 이해하는 기술
- 6가지의 사람 상태 (Lying, Crouch, Sitting, Walking, Standing, Running) 검출 가능
- 사람의 상태 이해 결과 이외에도 사람 관절 정보와 픽셀 단위로 분할된 세그멘테이션 정보도 같이 제공하여 사람과 관련된 여러 종합적인 정보를 동시에 파악할 수 있음

4. 기술의 사업성

▣ 활용 분야

예상 제품 / 서비스	예상 수요자
지능형 영상 관제 시스템	- CCTV 업체 - 지자체 관제 센터 - 소규모 관제
시각 데이터 분석 엔진	- 영상 분석 사업자

▣ 기대 효과

❖ 본 기술은 이미지/동영상에서 딥러닝 기반의 사람 상태 이해 기술은 다양한 응용 분야에서 활용이 가능한 기술로서 다음과 같은 시나리오를 고려할 수 있으며, 본 기술은 이미지와 동영상에서 사람을 검출하고 사람의 상태를 이해하는 기술에 최적화되어 있어 사업화에 필요한 추가적인 기술 개발이 필요함

❖기대 활용처

1. 고령자 상태 및 행동 분석: 사람의 관절 추정 기술을 활용하여 쓰러짐 등과 같은 위험 상태를 탐지하고 분석하는 기술 분야
2. 쓰러진 사람 탐지: 실세계 횡단 보도 또는 보행로에서 사람이 쓰러질 경우 이를 탐지하여 최대한 빨리 관제사에게 전달하는 지능형 감시 시스템으로 활용

5. 국내외 시장 동향

▣ 시장전망

❖ 지능형 영상 분석 관련 세계 시장은 2018년 65.5억달러에서 2025년 144.4억 달러로 연평균 21.4% 성장세를 전망, 국내 시장은 2018년 1,210억원에서 2025년 3,212억원으로 연평균 20.9%의 성장세 전망

(단위 : 억달러, 억원)

관련 제품/서비스	시장	1차년도 (2021)	2차년도 (2022)	3차년도 (2023)	4차년도 (2024)	5차년도 (2025)	합계
지능형 영상 분석	해외	65.5	79.5	96.8	118.2	144.4	504.4
	국내	1,210	1,536	1,960	2,508	3,212	10,426

(출처: Allied Market Research, Global Video Analytics Market (2018-2025))

* 국내 시장은 아시아 지역 시장의 합에서 20% 정도의 시장 규모로 산정, (환률 1$ = 1,100원)

감사합니다.

www.etri.re.kr

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0123682
(43) 공개일자 2021년10월14일

(51) 국제특허분류(Int. Cl.)
G06T 7/246 (2017.01) *G06Q 50/26* (2012.01)
G06T 7/194 (2017.01) *H04N 7/18* (2006.01)
(52) CPC특허분류
G06T 7/251 (2017.01)
G06Q 50/26 (2013.01)
(21) 출원번호 10-2020-0041134
(22) 출원일자 2020년04월03일
심사청구일자 2020년10월29일

(71) 출원인
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
배강민
대전광역시 유성구 대덕대로578번길 26-26
윤기민
대전광역시 유성구 봉명로 93 도안6단지센트럴시티 602동 2103호
(뒷면에 계속)
(74) 대리인
특허법인지명

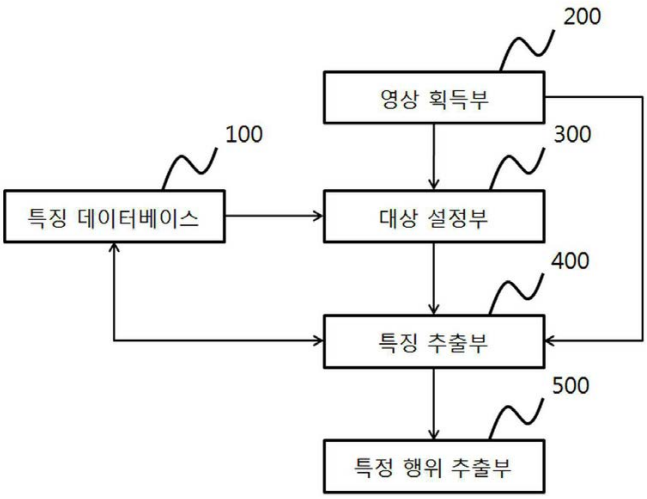
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템 및 방법**

(57) 요 약

본 발명은 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템 및 방법에 관한 것으로, 대상의 특징을 추출하기 위한 특징 데이터가 저장된 특징 데이터베이스; 기 설치된 카메라를 통해 촬영되는 영상을 획득하는 영상 획득부; 상기 영상 획득부를 통해 획득한 영상에서 기 학습된 딥러닝 모델을 이용하여 상기 물체를 들고 있는 사람을 특징 추출 대상으로 설정하는 대상 설정부; 설정된 특징 추출 대상에서 특정 행위를 하는지를 판단하기 위한 특징 데이터를 추출하는 특징 추출부; 및 상기 특징 추출부를 통해 추출된 특징 정보를 이용하여 쓰레기 투기 행위를 탐지하는 특정 행위 탐지부를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06T 7/194 (2017.01)
H04N 7/18 (2013.01)
G06T 2207/20084 (2013.01)

(72) 발명자
권용진
대전광역시 유성구 가정로 65, 108동 801호(신성동, 대림두레아파트)
김형일
대전광역시 유성구 문지로299번길 108, 302호(문지동)
문진영
대전광역시 유성구 지족로 343, 206동 604호 (지족동, 반석마을아파트2단지)
박종열
대전광역시 중구 서문로 96, 203동 1503호 (문화동, 센트럴파크2단지아파트)
배유석
대전광역시 유성구 관평1로 12, 704동 1401호 (관평동, 대덕테크노밸리7단지아파트)
이영완
대전광역시 유성구 송림로54번길 55 302호

이 발명을 지원한 국가연구개발사업

과제고유번호	1711103312
과제번호	2014-3-00123
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	2019년 RnD 재발견프로젝트
연구과제명	(딥뷰-1세부) 실시간 대규모 영상 데이터 이해·예측을 위한 고성능 비주얼 디스커버리 플랫폼 개발
기 여 율	1/1
과제수행기관명	한국전자통신연구원
연구기간	2019.01.01 ~ 2019.12.31

명 세 서

청구범위

청구항 1

대상의 특징을 추출하기 위한 특징 데이터가 저장된 특징 데이터베이스;

기 설치된 카메라를 통해 촬영되는 영상을 획득하는 영상 획득부;

상기 영상 획득부를 통해 획득한 영상에서 기 학습된 딥러닝 모델을 이용하여 상기 물체를 들고 있는 사람을 특징 추출 대상으로 설정하는 대상 설정부;

설정된 특징 추출 대상에서 특정 행위를 하는지를 판단하기 위한 특징 데이터를 추출하는 특징 추출부; 및

상기 특징 추출부를 통해 추출된 특징 정보를 이용하여 쓰레기 투기 행위를 탐지하는 특정 행위 탐지부를 포함하는 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템.

청구항 2

제 1항에 있어서,

상기 영상 획득부는,

웹을 통해 접속하는 IP 카메라인 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템.

청구항 3

제 1항에 있어서,

상기 특징 데이터베이스는,

대상의 특징을 추출하기 위해,

영상 획득부가 촬영하는 영상에서 추출되는 대상 주변의 이미지 데이터;

추출된 대상의 관절 영역 데이터를 추출하기 위한 관절 데이터;

추출된 대상의 특정 행동을 추출하기 위한 움직임 데이터; 및

상기 움직임 데이터를 통해 선정된 대상이 들고 있는 물체를 식별하기 위해 상기 영상 획득부가 촬영하는 영상에서의 주변 배경 데이터가 각각 저장된 것인 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템.

청구항 4

제 1항에 있어서,

상기 대상이 들고 있는 물체는,

사람이 들 수 있는 물체만을 학습하는 것인 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템.

청구항 5

제 1항에 있어서,

상기 특징 추출부는,

획득한 영상에서 대상 주변의 이미지 데이터를 추출하는 대상 주변 이미지 추출부;

추출된 대상의 관절 영역 데이터를 추출하는 대상 관절 영역 추정부;

추출된 대상의 쓰레기 투기 움직임 데이터를 추출하는 대상 주변 움직임 추출부; 및

상기 대상 주변 움직임 추출부를 통해 추출된 대상의 움직임이 행해진 대상의 주변 배경 데이터를 추출하는 대상 주변 배경 추출부를 포함하는 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템.

청구항 6

제 1항에 있어서,

상기 특징 추출부는,

DNN(Deep Neural network)을 이용하여 추출된 복수의 다중 특징 정보를 각각 학습하는 다중 특징 학습부를 더 포함하는 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템.

청구항 7

제 6항에 있어서,

상기 다중 특징 학습부는,

쓰레기 투기 영상과 일반적인 행동 인식 영상을 각각 학습하는 것인 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템.

청구항 8

제 1항에 있어서,

상기 특정 행위 탐지부는,

CCTV 데이터 분류기를 이용하는 것인 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템.

청구항 9

제 8항에 있어서,

사람의 세부 상태 분류기, 사람 자세 예측기 및 사람의 상태 분류기 중 적어도 하나 이상을 이용하여 상기 특정 행위 탐지부를 통해 탐지한 투기 행위에 대한 검증을 수행하는 투기 행동 검증부를 포함하는 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템.

청구항 10

제 9항에 있어서,

상기 투기 행동 검증부는,

상기 특정 행위 탐지부를 통해 탐지한 투기 행위의 검증여부에 따라, 관제사에게 이벤트 송부를 결정하는 것인

감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템.

청구항 11

영상 획득부에 의해, 기 설치된 카메라를 통해 촬영되는 영상을 획득하는 단계;

대상 설정부에 의해, 상기 획득한 영상에서 기 학습된 딥러닝 모델을 이용하여 물체를 들고 있는 사람을 특징 추출 대상으로 설정하는 단계;

특징 추출부에 의해, 상기 설정된 특징 추출 대상에서 특정 행위를 하는지를 판단하기 위한 특징 데이터를 추출하는 단계; 및

특정 행위 탐지부에 의해, 상기 추출된 특징 정보를 이용하여 쓰레기 투기 행위를 탐지하는 단계를 포함하는 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 방법.

청구항 12

제 11항에 있어서,

상기 영상 획득하는 단계는,

웹을 통해 접속하는 IP 카메라인 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 방법.

청구항 13

제 12항에 있어서,

상기 영상 획득하는 단계는,

IP, 아이디, 패스워드를 통해서 인증 받은 후 웹으로 인터넷을 접속할 수 있는 주소를 이용하여 접속하는 것인 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 방법.

청구항 14

제 11항에 있어서,

상기 대상이 손으로 잡고 있는 물체는,

사람이 들 수 있는 물체만을 학습하는 것인 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 방법.

청구항 15

제 11항에 있어서,

상기 특징 추출하는 단계는,

대상 주변 이미지 추출부에 의해, 상기 획득한 영상에서 대상 주변의 이미지 데이터를 추출하는 단계;

대상 관절 영역 추정부에 의해, 상기 추출된 대상의 관절 영역 데이터를 추출하는 단계;

대상 주변 움직임 추출부에 의해, 상기 추출된 대상의 쓰레기 투기 움직임 데이터를 추출하는 단계; 및

대상 주변 배경 추출부에 의해, 상기 추출된 대상의 움직임이 행해진 대상의 주변 배경 데이터를 추출하는 단계를 포함하는 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 방법.

청구항 16

제 11에 있어서,

CCTV 데이터 분류기를 이용하여 투기 행위를 검출하는 단계 및

상기 투기 행동 검증부에 의해, 상기 특정 행위 탐지부를 통해 탐지한 투기 행위의 검증여부에 따라, 관제사에게 이벤트 송부를 결정하는 단계;를 포함하는 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 방법.

청구항 17

제 16항에 있어서,

상기 탐지한 투기 행위에 대한 검증을 수행하는 단계는,

투기 행동 검증부에 의해, 사람의 세부 상태 분류기, 사람 자세 예측기 및 사람의 상태 분류기 중 적어도 하나 이상을 이용하여 상기 특정 행위 탐지부를 통해 탐지한 투기 행위에 대한 검증을 수행하는 것인 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 방법.

청구항 18

영상 획득부에 의해, 획득한 영상에서 대상 주변의 이미지 데이터를 저장하는 단계;

대상의 관절 영역 데이터를 추출하기 위한 관절 데이터를 저장하는 단계;

특정 움직임을 판단하기 위해, 추출된 대상의 움직임 데이터를 저장하는 단계; 및

추출된 대상이 들고 있는 물체를 판단하기 위해, 상기 영상 획득부에 의해, 획득한 영상에서 대상의 주변 배경 데이터를 저장하는 단계를 포함하는 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 학습방법.

청구항 19

제 18항에 있어서,

상기 각 데이터를 저장하는 단계는,

DNN(Deep Neural network)을 이용하여 추출된 복수의 다중 특징 정보를 각각 학습하는 단계를 더 포함하는 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 학습방법.

청구항 20

제 19항에 있어서,

상기 각 데이터를 저장하는 단계는,

쓰레기 투기 영상과 일반적인 행동 인식 영상을 각각 학습하는 것인 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 학습방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템 및 방법에 관한 것으

로, 더욱 상세하게는 감시카메라를 통해 제공되는 촬영 영상으로부터 특정 행위인 쓰레기 투기 행위자를 탐지하는 시스템 및 방법에 관한 것이다.

배 경 기 술

[0002] 환경과 공공을 위한 인공지능의 개발이 화두다. 플라스틱이나 쓰레기 등을 무분별한 사용과 버려짐으로 인해 많은 환경이 파괴되고 있다. 또한, 공공의 문제로서 무단으로 버려진 쓰레기 더미를 치우는데 많은 사회적 비용이 발생한다. 이를 해결하기 위해, 관제사가 CCTV를 돌려보거나 간단한 움직임 센서 등을 이용한 방법을 사용하고 있지만, 비효율성으로 인해서 사용 효과가 미비하다.

[0003] 기존의 유사한 연구로서는 놓고 간 물체 탐지 방법이 있다. 한번 발생한 움직임이 오랫동안 지속되면, 특정 물체가 사람에 의해서 장면에 등장한 이후에 두고 갔다라는 가정을 이용한다.

[0004] 그러나, 이 방식은 주차된 차량의 경우에도 놓고 간 물체라고 판단되는 문제가 있으며 또한 버려진 물체가 가려짐 없이 확연히 보여야만 동작하는 한계가 있다.

[0005] 종래의 쓰레기 투기 탐지 방법의 기술들은 미리 정해진 가정을 많이 활용한다.

[0006] 예를 들어, 쓰레기가 버려져 있는 영역에 사람이 나타나서 오래 머물러있는다든지, 투기 감시 카메라 아래에서 움직이는 물체가 1개에서 2개로 나타난다든지, 혹은 손 주변 움직임이 발생한다든지 하는 제약사항이 많다.

발명의 내용

해결하려는 과제

[0007] 본 발명은 종래 문제점을 해결하기 위해 안출된 것으로, 실제 투기 행위에서 발생하는 사람의 자세 및 배경 정보의 형상을 학습 기반의 방식을 통해 쓰레기 투기 행위를 감지하고, 사용자에게 이벤트를 알려주는 장치 및 방법을 제시하고자 한다.

[0008] 또한, 본 발명은 추가적인 일반 행동 분류기를 통해서, 쓰레기 투기 행위뿐만 아니라 일반적인 행동 분류를 지원하고, 이를 이용하여 오탐지가 적은 쓰레기 투기 감지 시스템 및 방법을 제공하고자 한다.

[0009] 본 발명의 목적은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 또 다른 목적들은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0010] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템은 대상의 특징을 추출하기 위한 특징 데이터가 저장된 특징 데이터베이스; 기 설치된 카메라를 통해 촬영되는 영상을 획득하는 영상 획득부; 상기 영상 획득부를 통해 획득한 영상에서 기 학습된 딥러닝 모델을 이용하여 상기 물체를 들고 있는 사람을 특징 추출 대상으로 설정하는 대상 설정부; 설정된 특징 추출 대상에서 특정 행위를 하는지를 판단하기 위한 특징 데이터를 추출하는 특징 추출부; 및 상기 특징 추출부를 통해 추출된 특징 정보를 이용하여 쓰레기 투기 행위를 탐지하는 특정 행위 탐지부를 포함한다.

[0011] 상기 영상 획득부는 웹을 통해 접속하는 IP 카메라이다.

[0012] 상기 특징 데이터베이스는 대상의 특징을 추출하기 위해, 영상 획득부가 촬영하는 영상에서 추출되는 대상 주변의 이미지 데이터; 추출된 대상의 관절 영역 데이터를 추출하기 위한 관절 데이터; 추출된 대상의 특정 행동을 추출하기 위한 움직임 데이터; 및 상기 움직임 데이터를 통해 선정된 대상이 들고 있는 물체를 식별하기 위해 상기 영상 획득부가 촬영하는 영상에서의 주변 배경 데이터가 각각 저장되고, 학습된다.

[0013] 상기 대상이 들고 있는 물체는, 사람이 들 수 있는 물체만을 학습한다.

[0014] 상기 특징 추출부는, 획득한 영상에서 대상 주변의 이미지 데이터를 추출하는 대상 주변 이미지 추출부; 추출된 대상의 관절 영역 데이터를 추출하는 대상 관절 영역 추정부; 추출된 대상의 쓰레기 투기 움직임 데이터를 추출하는 대상 주변 움직임 추출부; 및 상기 대상 주변 움직임 추출부를 통해 추출된 대상의 움직임이 행해진 대상의 주변 배경 데이터를 추출하는 대상 주변 배경 추출부를 포함한다.

[0015] 상기 특징 추출부는, DNN(Deep Neural network)을 이용하여 추출된 복수의 다중 특징 정보를 각각 학습하는 다중 특징 학습부를 더 포함하고, 상기 다중 특징 학습부는 쓰레기 투기 영상과 일반적인 행동 인식 영상을 각각

학습한다.

[0016] 상기 특정 행위 탐지부는 CCTV 데이터 분류기를 이용하고, 사람의 세부 상태 분류기, 사람 자세 예측기 및 사람의 상태 분류기 중 적어도 하나 이상을 이용하여 상기 특정 행위 탐지부를 통해 탐지한 투기 행위에 대한 검증을 수행하는 투기 행동 검증부를 포함한다.

[0017] 그리고 상기 투기 행동 검증부는, 상기 특정 행위 탐지부를 통해 탐지한 투기 행위의 검증여부에 따라, 관제사에게 이벤트 송부를 결정한다.

[0019] 그리고 본 발명의 일 실시예에 따른 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 방법은 영상 획득부에 의해, 기 설치된 카메라를 통해 촬영되는 영상을 획득하는 단계; 대상 설정부에 의해, 상기 획득한 영상에서 기 학습된 딥러닝 모델을 이용하여 물체를 들고 있는 사람을 특징 추출 대상으로 설정하는 단계; 특징 추출부에 의해, 상기 설정된 특징 추출 대상에서 특정 행위를 하는지를 판단하기 위한 특징 데이터를 추출하는 단계; 및 특정 행위 탐지부에 의해, 상기 추출된 특징 정보를 이용하여 쓰레기 투기 행위를 탐지하는 단계를 포함한다.

[0020] 상기 영상 획득하는 단계는, 웹을 통해 접속하는 IP 카메라이고, IP, 아이디, 패스워드를 통해서 인증 받은 후 웹으로 인터넷을 접속할 수 있는 주소를 이용하여 접속하는 것이 바람직하다.

[0021] 상기 대상이 손으로 잡고 있는 물체는, 사람이 들 수 있는 물체만을 학습할 수 있다.

[0022] 상기 특징 추출하는 단계는, 대상 주변 이미지 추출부에 의해, 상기 획득한 영상에서 대상 주변의 이미지 데이터를 추출하는 단계; 대상 관절 영역 추정부에 의해, 상기 추출된 대상의 관절 영역 데이터를 추출하는 단계; 대상 주변 움직임 추출부에 의해, 상기 추출된 대상의 쓰레기 투기 움직임 데이터를 추출하는 단계; 및 대상 주변 배경 추출부에 의해, 상기 추출된 대상의 움직임이 행해진 대상의 주변 배경 데이터를 추출하는 단계를 포함한다.

[0023] 또한, CCTV 데이터 분류기를 이용하여 투기 행위를 검출하는 단계 및 상기 투기 행동 검증부에 의해, 상기 특정 행위 탐지부를 통해 탐지한 투기 행위의 검증여부에 따라, 관제사에게 이벤트 송부를 결정하는 단계;를 포함한다.

[0025] 상기 탐지한 투기 행위에 대한 검증을 수행하는 단계는, 투기 행동 검증부에 의해, 사람의 세부 상태 분류기, 사람 자세 예측기 및 사람의 상태 분류기 중 적어도 하나 이상을 이용하여 상기 특정 행위 탐지부를 통해 탐지한 투기 행위에 대한 검증을 수행하는 것이 바람직하다.

[0027] 본 발명의 다른 실시에에 따른 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 학습방법은 영상 획득부에 의해, 획득한 영상에서 대상 주변의 이미지 데이터를 저장하는 단계; 대상의 관절 영역 데이터를 추출하기 위한 관절 데이터를 저장하는 단계; 특정 움직임을 판단하기 위해, 추출된 대상의 움직임 데이터를 저장하는 단계; 및 추출된 대상이 들고 있는 물체를 판단하기 위해, 상기 영상 획득부에 의해, 획득한 영상에서 대상의 주변 배경 데이터를 저장하는 단계를 포함한다.

[0028] 상기 각 데이터를 저장하는 단계는, DNN(Deep Neural network)을 이용하여 추출된 복수의 다중 특징 정보를 각각 학습하는 단계를 더 포함한다.

[0029] 상기 각 데이터를 저장하는 단계는, 쓰레기 투기 영상과 일반적인 행동 인식 영상을 각각 학습하는 것이 바람직하다.

발명의 효과

[0030] 본 발명의 일 실시예에 따르면, 실제 학습 데이터로 다양하게 수집한 영상을 통해서, 투기 행동과 그렇지 않은 행동을 잘 구별할 수 있는 분류기를 학습하여 기존의 규칙 기반의 방법보다 오탐지가 적고 개선된 결과를 얻을 수 있는 효과가 있다.

도면의 간단한 설명

[0031] 도 1은 본 발명의 일 실시예에 따른 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템을 설명하기 위한 기능블럭도.

도 2는본 발명의 일 실시예에서 영상 획득부를 통해 촬영된 영상의 일 예를 설명한 도면.

도 3은 도 1의 특징 추출부를 상세히 설명하기 위한 기능블록도.

도 4 내지 도 7은 본 발명의 일 실시예에 따른 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템에서 특정 행위를 감지하기 위한 과정을 설명하기 위한 참고도.

도 8은 본 발명의 일 실시예에서 획득한 영상에 복수개의 대상이 설정된 상태를 설명하기 위한 참고도.

도 9는 본 발명의 일 실시예에서 획득한 영상에 복수개의 대상이 들고 있는 물체를 설명하기 위한 참고도.

도 10은 본 발명의 일 실시예에서 특정 행위를 검증하기 위한 과정을 설명하기 위한 기능블록도.

도 11은 본 발명의 일 실시예에 따른 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 방법을 설명하기 위한 순서도.

도 12는 도 11의 특징 데이터 추출 단계를 설명하기 위한 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0032] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 것이며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하며, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 한편, 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 및/또는 "포함하는(comprising)"은 언급된 구성소자, 단계, 동작 및/또는 소자는 하나 이상의 다른 구성소자, 단계, 동작 및/또는 소자의 존재 또는 추가를 배제하지 않는다.

[0033] 도 1은 본 발명의 일 실시예에 따른 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템을 설명하기 위한 기능블럭도이다.

[0034] 도 1에 도시된 바와 같이, 본 발명의 일 실시예에 따른 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 시스템은 특징 데이터베이스(100), 영상 획득부(200), 대상 설정부(300), 특징 추출부(400) 및 특정 행위 탐지부(500)를 포함한다.

[0035] 특징 데이터베이스(100)는 대상의 특징을 추출하기 위한 특징 데이터가 저장된다.

[0036] 영상 획득부(200)는 기 설치된 카메라를 통해 촬영되는 영상을 획득한다. 도 2는 영상 획득부(200)를 통해 촬영된 영상의 예이다. 이러한 영상 획득부(200)는 CCTV 카메라일 수 있고, 웹을 통해 접속하는 IP 카메라이며, IP, 아이디, 패스워드를 통해서 인증 받은 후 웹으로 인터넷을 접속할 수 있는 주소를 이용하여 접속한다.

[0037] 대상 설정부(300)는 도 2에 도시된 바와 같이, 영상 획득부(200)를 통해 획득한 영상에서 기 학습된 딥러닝 모델을 이용하여 물체를 들고 있는 사람을 특징 추출 대상으로 설정한다. 이러한 대상 설정부(300)는 영상 획득부(200)를 통해 촬영된 영상에서 사람을 탐지하고, 사람이 소유하고 있는 물체를 탐지할 수 있다.

[0038] 특징 추출부(400)는 설정된 특징 추출 대상에서 특정 행위를 하는지를 판단하기 위한 특징 데이터를 추출한다. 여기서, 본 발명의 일 실시예에서의 특정 행위는 쓰레기를 투기하는 행위이다.

[0039] 도 3은 도 1의 특징 추출부(400)를 설명하기 위한 구성블록도이다.

[0040] 도 3에 도시된 바와 같이, 하기에서는 본 발명의 일 실시예에 따른 특징 추출부(400)는 대상 주변 이미지 추출부(410), 대상 관절 영역 추정부(420), 대상 주변 움직임 추출부(430) 및 대상 주변 배경 추출부(440)를 포함한다.

[0041] 대상 주변 이미지 추출부(410)는 특징 데이터베이스(100)에 저장된 특정 정보를 이용하여 도 4에 도시된 바와 같이, 상기 영상에서 대상 주변의 이미지 데이터를 추출한다.

[0042] 대상 관절 영역 추정부(420)는 특징 데이터베이스(100)에 저장된 특정 정보를 이용하여 획득한 영상에서 도 5에 도시된 바와 같이, 추출된 대상의 관절 영역 데이터를 추출한다.

[0043] 대상 주변 움직임 추출부(430)는 특징 데이터베이스(100)에 저장된 특정 정보를 이용하여 획득한 영상에서 도 7에 도시된 바와 같이, 추출된 대상의 쓰레기 투기 움직임 데이터를 추출한다.

[0044] 대상 주변 배경 추출부(440)는 특징 데이터베이스(100)에 저장된 특정 정보를 이용하여 도 8에 도시된 바와 같이, 대상 주변 움직임 추출부(430)를 통해 추출된 대상의 움직임이 행해진 대상의 주변 배경 데이터를 추출한다.

[0045] 이후, 특정 행위 탐지부(500)는 특징 추출부(400)를 통해 추출된 특징 정보를 이용하여 특정 행위인 쓰레기 투기 행위를 탐지한다.

[0046] 일 예로, 도 2에 도시된 바와 같이, 영상 획득부(200)로부터 촬영된 영상 정보가 수집되면, 대상 설정부(300)는 영상 획득부(200)를 통해 획득한 영상에서 기 학습된 딥러닝 모델을 이용하여 물체를 들고 있는 사람을 특징 추출 대상과 대상이 들고 있는 물체(T1)을 설정한다. 한편, 특징 추출 대상은 도 8에 도시된 바와 같이, 복수의 사람(01 내지 03)이 특징 추출 대상으로 설정되거나 해당 대상이 들고 있는 물체(T1)이 될 수 있다.

[0047] 이를 위해, 특징 추출부(400)는 설정된 특징 추출 대상에서 특정 행위를 하는지를 판단하기 위한 특징 데이터를 추출한다.

[0048] 먼저, 대상 주변 이미지 추출부(410)는 특징 데이터베이스(100)에 저장된 특정 정보를 이용하여 상기 영상에서 대상 주변의 이미지 데이터를 추출한다.

[0049] 대상 관절 영역 추정부(420)는 특징 데이터베이스(100)에 저장된 특정 정보를 이용하여 획득한 영상에서 추출된 대상의 관절 영역 데이터를 추출한다.

[0050] 대상 주변 움직임 추출부(430)는 특징 데이터베이스(100)에 저장된 특정 정보를 이용하여 획득한 영상에서 추출된 대상의 쓰레기 투기 움직임 데이터를 추출한다.

[0051] 대상 주변 배경 추출부(440)는 특징 데이터베이스(100)에 저장된 특정 정보를 이용하여 상기 대상 주변 움직임 추출부(430)를 통해 추출된 대상의 움직임이 행해진 대상의 주변 배경 데이터를 추출한다.

[0052] 이후, 특정 행위 탐지부(500)는 특징 추출부(400)를 통해 추출된 특징 정보를 이용하여 특정 행위인 쓰레기 투기 행위를 탐지한다.

[0053] 특정 행위 탐지부(500)는 특징 추출부(400)에 의해, 추출된 특징 데이터를 이용하여 추출된 특징 추출 대상의 특징 행위를 탐지할 수 있다.

[0054] 즉, 특정 행위 탐지부(500)는 추출된 특징 추출 대상에 대하여 추출한 대상 주변의 이미지 데이터를 비교하여 사람이 들고 있는 물체 즉, 쓰레기를 검출한다. 이와 같이, 사림이 들고 있는 물체가 검출되면 해당 물체를 들고 있는 대상만을 선택하여 정밀 분석할 수 있다.

[0055] 그리고 특징 행위 탐지부(500)는 추출된 특징 추출 대상에 대하여 관절 영역 데이터를 비교하여 쓰레기의 투기 행동을 탐지한다. 즉, 쓰레기 투기시의 관절 영역 데이터를 비교함으로써, 쓰레기 투기 행동을 수행하고자 하는 대상인지를 탐지할 수 있다.

[0056] 또한, 특징 행위 탐지부(500)는 추출된 특징 추출 대상에 대하여 대상 쓰레기 투기 움직임 데이터와 비교하여 검출된 대상이 쓰레기를 투기 하는 움직임을 수행함을 탐지할 수 있다.

[0057] 그리고 특징 행위 탐지부(500)는 추출된 특징 추출 대상에 대하여 대상의 주변 배경 데이터와 비교하여 대상이 특징 물체를 들고 있는지를 판단하거나 대상이 들고 있는 물체를 투기하는지를 탐지하여 쓰레기 투기 대상 탐지를 보다 정확하게 할 수 있다.

[0058] 한편 특징 추출부(400)는 DNN(Deep Neural network)을 이용하여 추출된 복수의 다중 특징 정보를 각각 학습하는 다중 특징 학습부를 더 포함할 수 있다. 여기서, 본 실시예에서의 대상이 손으로 잡고 있는 물체는 사람이 들 수 있는 물체만을 학습하는 것이 바람직하다. 이와 같이, 학습 대상을 한정함으로써 보다 정확하게 물체를 학습할 수 있다.

[0059] 한편, 다중 특징 학습부는 쓰레기 투기 영상과 일반적인 행동 인식 영상을 각각 학습할 수 있다. 이와 같이, 쓰

레기 투기 영상과 일반적인 행동 인식 영상을 각각 학습함으로써, 그 학습 대상 영상 데이터를 학습하여 그 정확도를 높일 수 있다.

[0060] 도 10은 본 발명의 일 실시예에서 특정 행위를 검증하기 위한 과정을 설명하기 위한 기능블록도이다.

[0061] 한편, 본 발명의 일 실시예에서는, 특정 행위 추출부(500)를 통해 추출된 특정 행위를 검증하여 검출의 정확성을 높일 수 있다.

[0062] 이를 위해, 본 발명의 일 실시예에서는, CCTV 데이터 분류기를 이용하여 쓰레기를 투기하는 대상을 탐지하고, 사람의 세부 상태 분류기, 사람 자세 예측기 및 사람의 상태 분류기 중 적어도 하나 이상을 이용하여 특정 행위 탐지부(500)를 통해 탐지한 투기 행위에 대한 검증을 수행하는 투기 행동 검증부(600)를 포함할 수 있다. 여기서, 사람의 세부 상태 분류기는 사람의 포괄적 행동 특징(예시; 서있다, 누워있다, 앉아있다 등)을 분류할 수 있는 MPHB 분류기이고, 사람의 상태 분류기는 사람의 세부적 행동 특징(예시; 자전거를 타다, 카트를 밀다 등)을 분류할 수 있는 Standford40 분류기이다.

[0063] 그리고, 투기 행동 검증부(600)는 상기 특정 행위 탐지부(500)를 통해 탐지한 투기 행위의 검증여부에 따라, 관제사에게 이벤트 송부를 결정할 수 있다.

[0065] 이하, 하기에서는 본 발명의 일 실시예에 따른 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 탐지 방법에 대하여 도 11을 참조하여 설명하기로 한다.

[0066] 먼저, 영상 획득부(200)에 의해, 기 설치된 카메라를 통해 촬영되는 영상을 획득한다(S100). 상기 영상 획득하는 단계(S100)는 웹을 통해 접속하는 IP 카메라인 것이 바람직하고, IP, 아이디, 패스워드를 통해서 인증 받은 후 웹으로 인터넷을 접속할 수 있는 주소를 이용하여 접속할 수 있다.

[0067] 이후, 대상 설정부(300)에 의해, 상기 획득한 영상에서 기 학습된 딥러닝 모델을 이용하여 물체를 들고 있는 사람을 특징 추출 대상으로 설정한다(S200).

[0068] 이어서, 특징 추출부(400)에 의해, 상기 설정된 특징 추출 대상에서 쓰레기를 투기하는 행위를 하는지를 판단하기 위한 특징 데이터를 추출한다(S300).

[0069] 이하, 하기에서는 본 발명의 일 실시예에 따른 특징 추출하는 단계(S300)에 대하여 도 12를 참조하여 설명하기로 한다.

[0070] 먼저, 대상 주변 이미지 추출부(410)에 의해, 상기 획득한 영상에서 대상 주변의 이미지 데이터를 추출한다(S310).

[0071] 그리고 대상 관절 영역 추정부(420)에 의해, 상기 추출된 대상의 관절 영역 데이터를 추출한다(S320).

[0072] 또한, 대상 주변 움직임 추출부(430)에 의해, 상기 추출된 대상의 쓰레기 투기 움직임 데이터를 추출한다(S330).

[0073] 대상 주변 배경 추출부(440)에 의해, 상기 추출된 대상의 움직임이 행해진 대상의 주변 배경 데이터를 추출한다(S340).

[0074] 이후, 특정 행위 탐지부(500)에 의해, 상기 추출된 특징 정보를 이용하여 쓰레기 투기 행위를 탐지한다(S400).

[0075] 이후, 탐지된 쓰레기 투기 행위를 검증한다(S500). 즉, 특정 행위 추출부(500)를 통해 추출된 특정 행위를 검증하여 검출의 정확성을 높일 수 있다.

[0076] 상기 검증 단계는, 투기 행동 검증부(600)에 의해, 상기 CCTV 데이터 분류기를 이용하여 쓰레기를 투기하는 대상을 탐지할 수 있다. 그리고, 상기 검증 단계는, 사람의 세부 상태 분류기, 사람 자세 예측기 및 사람의 상태 분류기 중 적어도 하나 이상을 이용하여 특정 행위 탐지부(500)를 통해 탐지한 투기 행위에 대한 검증을 수행할 수 있다. 여기서, 사람의 세부 상태 분류기는 사람의 포괄적 행동 특징(예시; 서있다, 누워있다, 앉아있다 등)을 분류할 수 있는 MPHB 분류기이고, 사람의 상태 분류기는 사람의 세부적 행동 특징(예시; 자전거를 타다, 카트를 밀다 등)을 분류할 수 있는 Standford40 분류기이다.

[0077] 그리고, 투기 행동 검증부(600)에 의해, 탐지한 투기 행위의 검증여부에 따라, 관제사에게 이벤트 송부를 결정할 수 있다.

[0078] 한편, 상기 대상이 손으로 잡고 있는 물체는 사람이 들 수 있는 물체만을 학습할 수 있다.

[0080] 본 발명의 일 실시예에 따른 감시카메라 환경에서 다중 특징 정보를 이용한 쓰레기 투기 행위자 학습 방법은 획득한 영상에서 대상 주변의 이미지 데이터를 특징 데이터베이스(100)에 저장한다. 이러한 이미지 데이터는 검출된 대상이 손으로 들고 있는 물건 또는 쓰레기 더미에 쌓여 있는 물건을 검출하는데 이용된다.

[0081] 이어서, 대상의 관절 영역 데이터를 추출하기 위한 관절 데이터를 특징 데이터베이스(100)에 저장한다. 여기서, 관절 데이터는 대상이 특정 행위를 수행하는지를 검출하는데 이용된다.

[0082] 그리고 특정 움직임을 판단하기 위해, 추출된 대상의 움직임 데이터를 특징 데이터베이스(100)에 저장한다. 여기서, 움직임 데이터는 대상이 특정 행위를 수행하는지를 검출하는데 이용된다. 이를 위해, 관절 데이터와 움직임 데이터의 경우 특정 행위와 일반적인 행동 인식 영상을 데이터로 이용하여 학습함으로써, 실제 CCTV를 통해 촬영된 영상뿐만 아니라, 특정 동작을 검출하기 위해 다른 영상 매체를 통해 촬영된 영상 또한 이용될 수 있다.

[0083] 또한 추출된 대상이 들고 있는 물체를 판단하기 위해, 상기 획득한 영상에서 대상의 주변 배경 데이터를 특징 데이터베이스(100)에 저장한다.

[0084] 여기서, 주변 배경 데이터는 CCTV를 통해 촬영된 영상내에 복수의 대상이 검출되는 경우, 실제 쓰레기를 투기하는 대상 즉, 특정 행동을 수행하는 대상을 검출하기 위한 데이터로 이용된다. 따라서, 실제 CCTV를 통해 촬영된 영상으로부터 획득한 데이터가 저장되고 학습되는 것이 바람직하다.

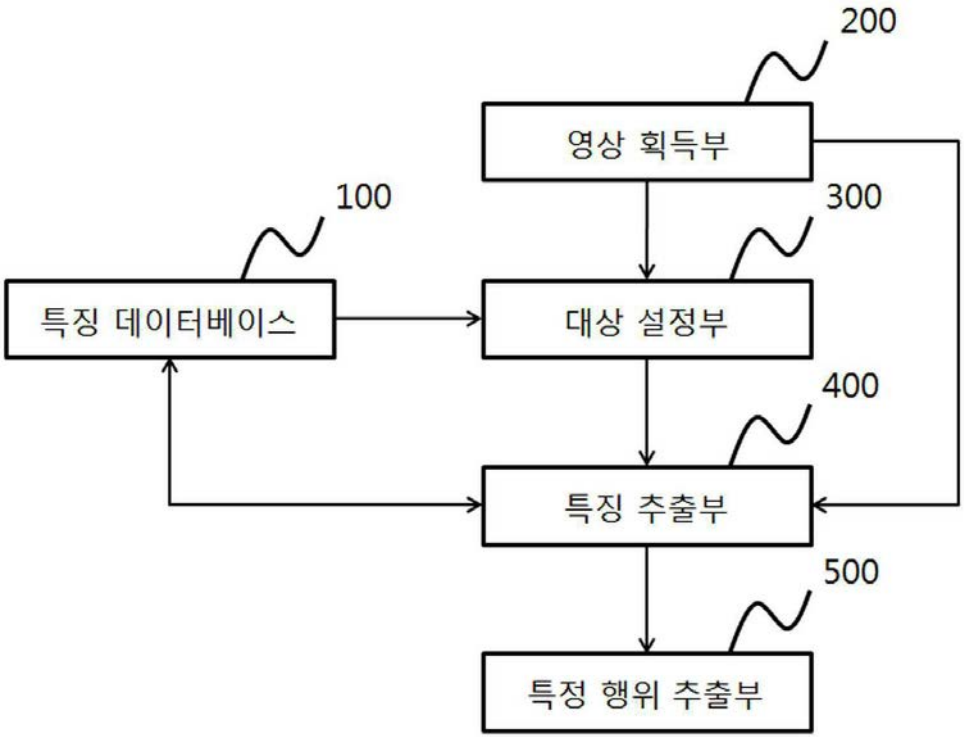
[0085] 여기서, 상기 각 데이터를 저장하는 단계는 DNN(Deep Neural network)을 이용하여 추출된 복수의 다중 특징 정보를 각각 학습하는 것이 바람직하다.

[0086] 또한 상기 각 데이터를 저장하는 단계는 쓰레기 투기 영상과 일반적인 행동 인식 영상을 각각 학습할 수 있다.

[0087] 이상, 본 발명의 구성에 대하여 첨부 도면을 참조하여 상세히 설명하였으나, 이는 예시에 불과한 것으로서, 본 발명이 속하는 기술분야에 통상의 지식을 가진자라면 본 발명의 기술적 사상의 범위 내에서 다양한 변형과 변경이 가능함은 물론이다. 따라서 본 발명의 보호 범위는 전술한 실시예에 국한되어서는 아니되며 이하의 특허청구범위의 기재에 의하여 정해져야 할 것이다.

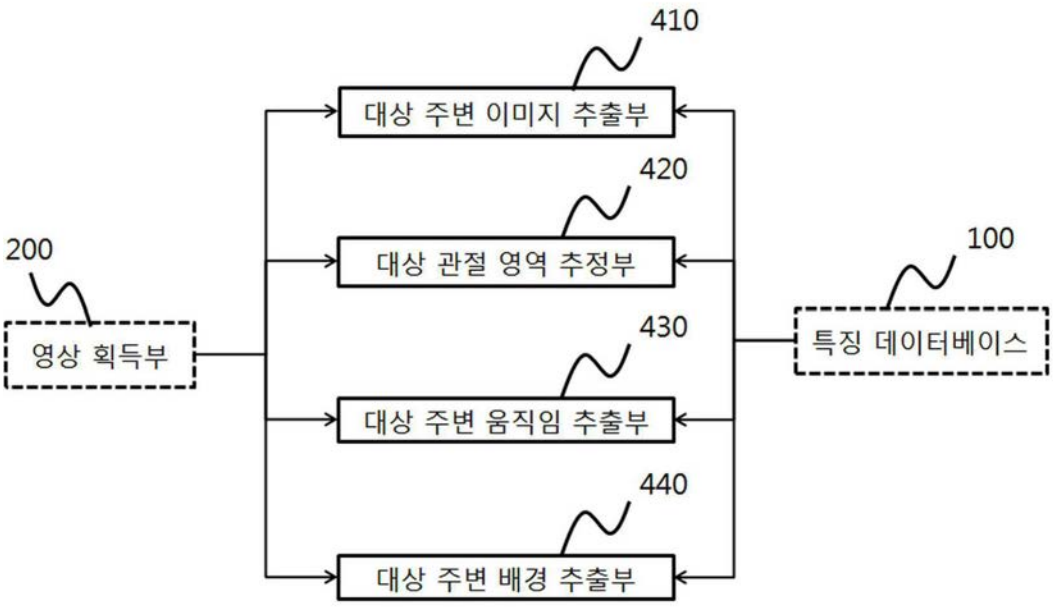
도면

도면1



도면2

도면3


410
대상 주변 이미지 추출부
420
대상 관절 영역 추정부
200
영상 획득부
100
특징 데이터베이스
430
대상 주변 움직임 추출부
440
대상 주변 배경 추출부

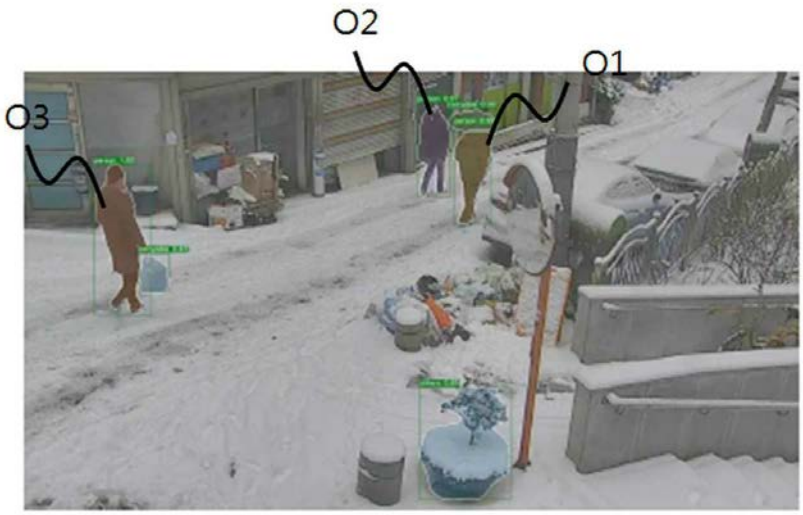
도면4

도면5

도면6

도면7

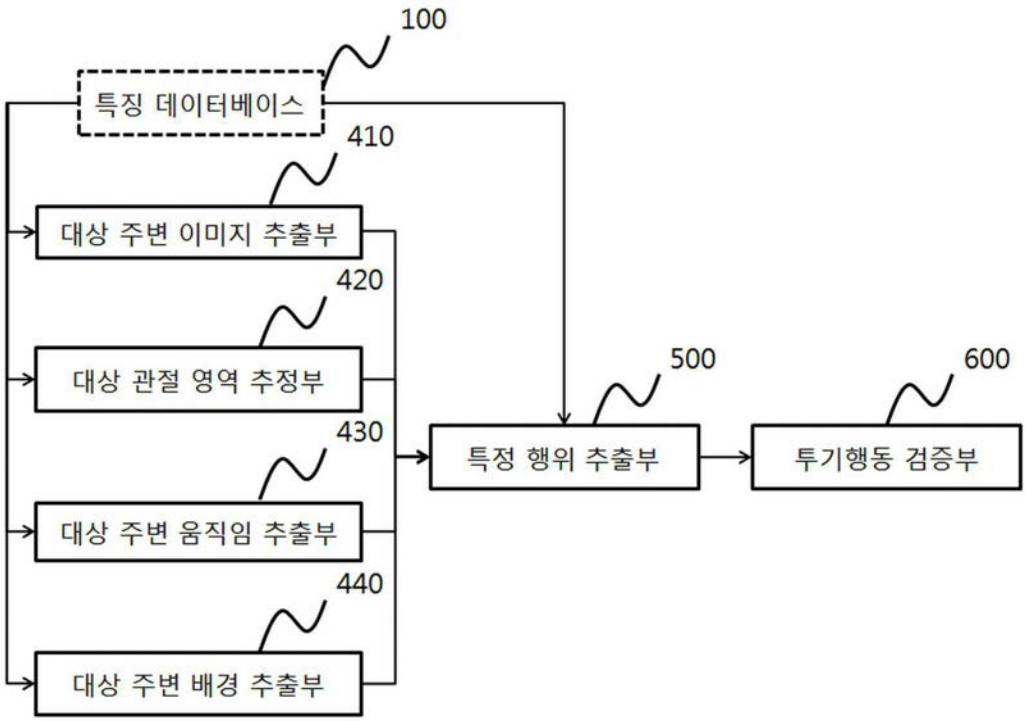
도면8



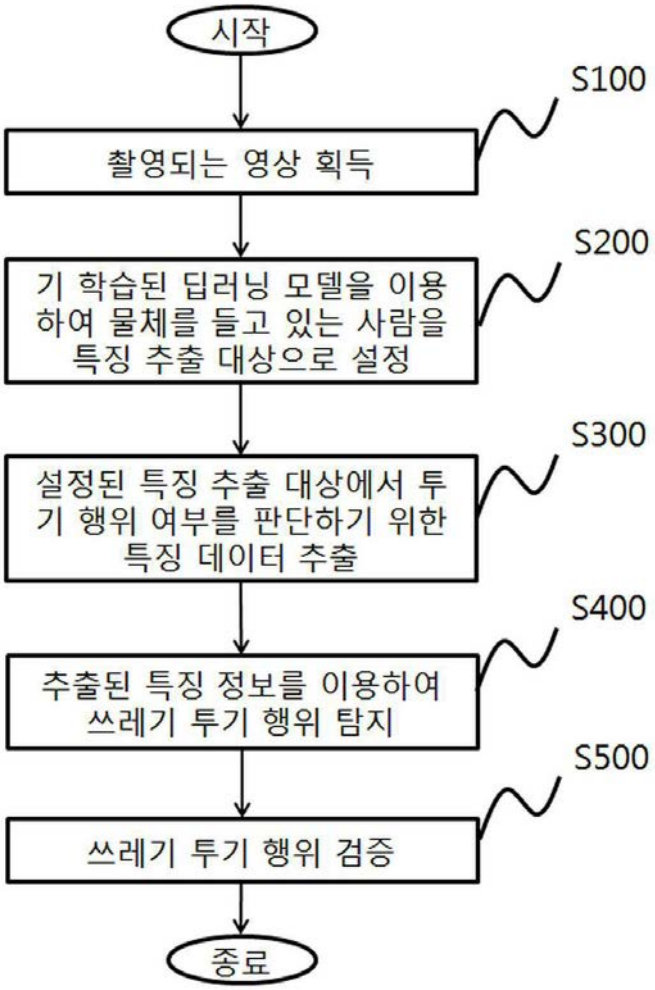
도면9


O4
T1

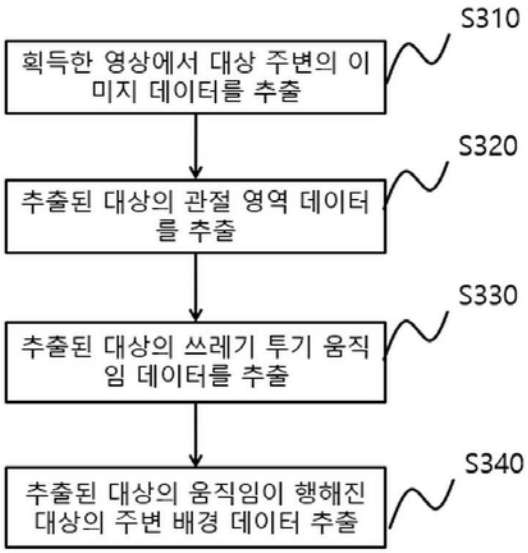
도면10


100
특징 데이터베이스
410
대상 주변 이미지 추출부
420
대상 관절 영역 추정부
430
대상 주변 움직임 추출부
440
대상 주변 배경 추출부
500
특정 행위 추출부
600
투기행동 검증부

도면11


시작
S100
촬영되는 영상 획득
S200
기 학습된 딥러닝 모델을 이용하여 물체를 들고 있는 사람을 특징 추출 대상으로 설정
S300
설정된 특징 추출 대상에서 투기 행위 여부를 판단하기 위한 특징 데이터 추출
S400
추출된 특징 정보를 이용하여 쓰레기 투기 행위 탐지
S500
쓰레기 투기 행위 검증
종료

도면12


S310
획득한 영상에서 대상 주변의 이미지 데이터를 추출
S320
추출된 대상의 관절 영역 데이터를 추출
S330
추출된 대상의 쓰레기 투기 움직임 데이터를 추출
S340
추출된 대상의 움직임이 행해진 대상의 주변 배경 데이터 추출

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0119425
(43) 공개일자 2020년10월20일

(51) 국제특허분류(Int. Cl.)
G06K 9/00 (2006.01) *G06N 20/00* (2019.01)
(52) CPC특허분류
G06K 9/00228 (2013.01)
G06K 9/00268 (2013.01)
(21) 출원번호 10-2019-0038049
(22) 출원일자 2019년04월01일
심사청구일자 2019년12월05일

(71) 출원인
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
김형일
대전광역시 유성구 문지로299번길 108, 302호(문지동)
권용진
대전광역시 유성구 가정로 65, 108동 801호(신성동, 대림두레아파트)
(뒷면에 계속)
(74) 대리인
특허법인지명

전체 청구항 수 : 총 17 항

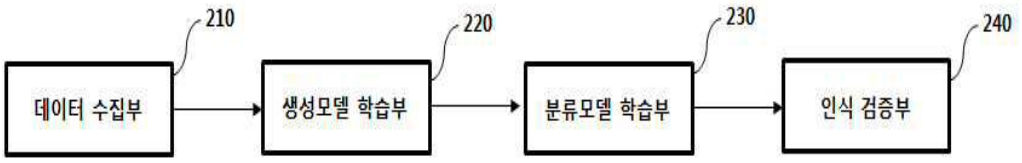
(54) 발명의 명칭 **도메인 적응 기반 객체 인식 장치 및 그 방법**

(57) 요 약

본 발명은 도메인 적응 기반의 객체 인식 장치 및 그 방법에 관한 것이다

본 발명에 따른 도메인 적응 기반 객체 인식 장치는 도메인 적응 기반 객체 인식 프로그램이 저장된 메모리 및 프로그램을 실행시키는 프로세서를 포함하되, 프로세서는 입력 프로브 영상에 대해 도메인 적응 기반으로 갤러리 영상과 유사한 영상 또는 특징으로 생성시키기 위한 생성모델을 학습하고, 갤러리 영상과 프로브 영상의 학습 데이터베이스를 이용하여 객체인식 분류 모델을 학습하여, 입력 프로브 영상을 이용한 객체 인식을 수행하는 것을 특징으로 한다.

대 표 도 - 도2



(52) CPC특허분류
G06K 9/00288 (2013.01)
G06N 20/00 (2019.01)
(72) 발명자
문진영
대전광역시 유성구 지족로 343, 206동 604호 (지족동, 반석마을아파트2단지)
박종열
대전광역시 중구 서문로 96, 203동 1503호 (문화동, 센트럴파크2단지아파트)
오성찬
서울특별시 송파구 백제고분로18길 30, 107동 303호(잠실동, 우성아파트)
윤기민
대전광역시 유성구 전민로26번길 14, 102호(전민동, 그레이스빌)
이전우
충청남도 계룡시 두마면 사계로 51, 103동 502호 (계룡대림e편한세상아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호	2014-3-00123
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원(IITP)
연구사업명	ICT융합산업원천기술개발사업
연구과제명	(딥뷰-1세부) 실시간 대규모 영상 데이터 이해 · 예측을 위한 고성능 비주얼 디스커버리 플랫폼 개발
기 여 율	1/1
과제수행기관명	한국전자통신연구원
연구기간	2018.01.01 ~ 2018.12.31

명 세 서

청구범위

청구항 1

도메인 적응 기반 객체 인식 프로그램이 저장된 메모리; 및

상기 프로그램을 실행시키는 프로세서를 포함하되,

상기 프로세서는 입력 프로브 영상에 대해 도메인 적응 기반으로 갤러리 영상과 유사한 영상 또는 특징으로 생성시키기 위한 생성모델을 학습하고, 갤러리 영상과 프로브 영상의 학습 데이터베이스를 이용하여 객체인식 분류 모델을 학습하여, 상기 입력 프로브 영상을 이용한 객체 인식을 수행하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 2

제1항에 있어서,

상기 프로세서는 객체의 특징 정보를 이용한 전처리를 수행하여 상기 학습 데이터베이스를 구축하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 3

제1항에 있어서,

상기 프로세서는 상기 학습 데이터베이스와 갤러리에 등록되지 않은 외부 영상 데이터베이스를 활용하여 전처리를 수행하여 상기 생성모델을 학습하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 4

제3항에 있어서,

상기 프로세서는 영상 소스를 분류하고, 도메인 적응 기반의 새로운 영상을 생성하고, 객체 ID를 판별하여 갤러리 영상의 스타일을 학습하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 5

제1항에 있어서,

상기 프로세서는 상기 학습 데이터베이스에 대해 전처리 수행, 특징 추출 수행에 따라 객체 ID 분류기를 학습시켜, 상기 객체인식 분류 모델을 학습하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 6

제1항에 있어서,

상기 프로세서는 수신된 입력 영상으로부터 객체 영역을 검출하고, 상기 생성모델을 이용하여 입력 영상을 갤러리 영상과 유사한 새로운 영상 또는 특징으로 생성하고, 생성된 새로운 영상에 대해 상기 객체인식 분류 모델을 이용한 특징 추출을 수행하여, 객체의 ID 정보를 획득하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 7

얼굴 영상을 수집하는 데이터 수집부;

갤러리 얼굴 영상의 스타일을 학습하는 생성모델 학습부;

얼굴 인식 및 매칭을 수행하기 위해 사전에 등록이 필요한 인물 정보를 이용하여 분류모델을 학습하는 분류모델 학습부; 및

생성모델 및 분류모델을 이용하여 실제 입력 얼굴 영상에 대한 인식을 수행하는 인식 검증부

를 포함하는 도메인 적응 기반 객체 인식 장치.

청구항 8

제7항에 있어서,

상기 데이터 수집부는 상기 얼굴 영상에 대해 특징점 정보를 이용하여 전처리를 수행하고, 갤러리 얼굴 영상 데이터베이스를 구축하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 9

제8항에 있어서,

상기 생성모델 학습부는 상기 갤러리 얼굴 영상 데이터베이스와 외부 얼굴 영상 데이터베이스를 이용하여 얼굴 영상 생성모델을 학습하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 10

제9항에 있어서,

상기 생성모델 학습부는 입력된 영상이 상기 갤러리 얼굴 영상 데이터베이스에 포함되는지 여부를 판별하고, 학습된 갤러리 얼굴 영상의 스타일과 유사하게 새로운 얼굴 영상을 생성하고, 입력된 영상의 ID를 판별하여 상기 얼굴 영상 생성모델을 학습하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 11

제8항에 있어서,

상기 분류모델 학습부는 상기 갤러리 얼굴 영상 데이터베이스를 이용한 전처리 및 특징 추출에 따른 얼굴 ID 분류 결과에 따라 오류 계산을 수행하여, 얼굴 ID 분류기를 학습시키는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 12

제7항에 있어서,

상기 인식 검증부는 비디오 입력으로부터 얻은 각 프레임으로부터 얼굴 영역을 검출하고, 상기 생성모델을 이용하여 입력된 얼굴 영상이 갤러리 얼굴 영상과 유사하도록 새로운 얼굴 영상을 생성하고, 상기 분류모델을 이용하여 특징 추출 및 매칭을 수행하여, ID 정보를 획득하는 것

인 도메인 적응 기반 객체 인식 장치.

청구항 13

(a) 객체 영상을 수집하는 단계;

(b) 갤러리 영상의 스타일을 학습하여 생성모델을 학습하는 단계;

(c) 객체 인식을 위해 사전에 등록이 필요한 정보를 이용하여 분류모델을 학습하는 단계; 및

(d) 상기 생성모델 및 분류모델을 이용하여 영상 내 객체를 인식하는 단계

를 포함하는 도메인 적응 기반 객체 인식 방법.

청구항 14

제13항에 있어서,

상기 (a) 단계는 특징점 정보를 이용하여 상기 객체 영상에 대한 전처리를 수행하고, 갤러리 영상 데이터베이스를 구축하는 것

인 도메인 적응 기반 객체 인식 방법.

청구항 15

제13항에 있어서,

상기 (b) 단계는 갤러리 영상 데이터베이스와 외부 영상 데이터베이스를 이용하여, 입력 영상을 상기 갤러리 영상의 스타일에 부합되는 새로운 영상 또는 특징으로 생성하기 위한 상기 생성모델을 학습하는 것

인 도메인 적응 기반 객체 인식 방법.

청구항 16

제13항에 있어서,

상기 (c) 단계는 갤러리 영상 데이터베이스를 이용하여 전처리 및 특징 추출을 수행하고, ID 분류 결과에 따른 오류 계산을 수행하여 ID 분류기를 학습시키는 것

인 도메인 적응 기반 객체 인식 방법.

청구항 17

제13항에 있어서,

상기 (d) 단계는 비디오 입력으로부터 얻은 각 프레임으로부터 객체 영역을 검출하고, 상기 생성모델을 이용하여 객체가 상기 갤러리 영상과 유사하도록 새로운 영상 또는 특징을 생성하고, 상기 분류모델을 이용하여 특징을 추출하고 매칭을 수행하여, 객체의 ID 정보를 획득하는 것

인 도메인 적응 기반 객체 인식 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 도메인 적응 기반의 객체 인식 장치 및 그 방법에 관한 것이다.

배 경 기 술

[0003] 종래의 객체 인식 기술은 객체 검출, 전처리, 특징 추출, 인식/매칭의 과정을 통해 수행된다.

[0004] 객체 인식 기술은 사전에 등록된 정보를 기반으로 현재 입력되는 정보를 인식하게 되는데, 다양한 환경 변화를 보상시키기 위한 전처리 또는 환경 변화에 강인한 특징 추출 기법이 제안되었으나, 실제 발생되는 모든 변화를 다룰 수 없는 한계점이 있고, 강인한 특징 추출 학습을 위해 대량의 데이터가 요구되는 문제점이 있다.

발명의 내용

해결하려는 과제

[0006] 본 발명은 전술한 문제점을 해결하기 위하여 제안된 것으로, 제한된 집합의 갤러리 영상과 프로브 영상을 이용하여 갤러리 영상 또는 특징의 스타일을 학습하고, 프로프 영상을 도메인 적응을 통해 갤러리 영상의 스타일과 유사한 새로운 영상 또는 특징을 생성함으로써, 외부 환경 변화로부터 강인한 객체 인식이 가능한 장치 및 방법을 제공하는데 그 목적이 있다.

과제의 해결 수단

[0008] 본 발명에 따른 도메인 적응 기반 객체 인식 장치는 도메인 적응 기반 객체 인식 프로그램이 저장된 메모리 및 프로그램을 실행시키는 프로세서를 포함하되, 프로세서는 입력 프로브 영상에 대해 도메인 적응 기반으로 갤러리 영상과 유사한 영상 또는 특징으로 생성시키기 위한 생성모델을 학습하고, 갤러리 영상과 프로브 영상의 학습 데이터베이스를 이용하여 객체인식 분류 모델을 학습하여, 입력 프로브 영상을 이용한 객체 인식을 수행하는 것을 특징으로 한다.

[0009] 본 발명에 따른 도메인 적응 기반 객체 인식 장치는 얼굴 영상을 수집하는 데이터 수집부와, 갤러리 얼굴 영상의 스타일을 학습하는 생성모델 학습부와, 얼굴 인식 및 매칭을 수행하기 위해 사전에 등록이 필요한 인물 정보를 이용하여 분류모델을 학습하는 분류모델 학습부 및 생성모델 및 분류모델을 이용하여 실제 입력 얼굴 영상에 대한 인식을 수행하는 인식 검증부를 포함하는 것을 특징으로 한다.

[0010] 본 발명에 따른 도메인 적응 기반 객체 인식 방법은 객체 영상을 수집하는 단계와, 갤러리 영상의 스타일을 학습하여 생성모델을 학습하는 단계와, 객체 인식을 위해 사전에 등록이 필요한 정보를 이용하여 분류모델을 학습하는 단계 및 생성모델과 분류모델을 이용하여 영상 내 객체를 인식하는 단계를 포함하는 것을 특징으로 한다.

발명의 효과

[0012] 본 발명의 실시예에 따르면, 갤러리 얼굴 영상과 프로브 얼굴 영상 사이의 차이가 큰 신분증(주민등록증, 여권 등) 인식, 출입통제 시스템 등에 적용되어, 제약된 환경에서 촬영된 갤러리 얼굴 영상과 다양한 변화를 갖는 얼굴 영상을 이용하여 갤러리 얼굴 영상의 스타일을 학습하고, 프로브 얼굴 영상 입력을 학습 모델에 의해 새로운 영상(갤러리 영상의 스타일과 유사한 영상) 또는 특징으로 생성함으로써, 갤러리 및 프로브 얼굴 영상 사이의

불일치를 줄이고 외부 환경 변화로부터 강인한 얼굴인식 수행이 가능한 효과가 있다.

[0013] 본 발명에 따르면 얼굴영상 생성모델 학습과 얼굴인식 분류모델 학습을 동시에 사용함으로써, 외부 환경 변화로부터 강인한 얼굴 인식 수행의 신뢰성을 높이는 것이 가능한 효과가 있다.

[0014] 본 발명의 효과는 이상에서 언급한 것들에 한정되지 않으며, 언급되지 아니한 다른 효과들은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0016] 도 1 및 도 2는 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 장치를 나타내는 블록도이다.

도 3은 본 발명의 실시예에 따른 데이터 수집부를 나타내는 블록도이다.

도 4는 본 발명의 실시예에 따른 생성모델 학습부를 나타내는 블록도이다.

도 5는 본 발명의 실시예에 따른 분류모델 학습부를 나타내는 블록도이다.

도 6은 본 발명의 실시예에 따른 인식 검증부를 나타내는 블록도이다.

도 7은 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 방법을 나타내는 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0017] 본 발명의 전술한 목적 및 그 이외의 목적과 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다.

[0018] 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 이하의 실시예들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 목적, 구성 및 효과를 용이하게 알려주기 위해 제공되는 것일 뿐으로서, 본 발명의 권리범위는 청구항의 기재에 의해 정의된다.

[0019] 한편, 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 및/또는 "포함하는(comprising)"은 언급된 구성소자, 단계, 동작 및/또는 소자가 하나 이상의 다른 구성소자, 단계, 동작 및/또는 소자의 존재 또는 추가됨을 배제하지 않는다.

[0021] 이하에서는, 당업자의 이해를 돕기 위하여 본 발명이 제안된 배경에 대하여 먼저 서술하고, 본 발명의 실시예에 대하여 서술하기로 한다.

[0022] 종래 기술에 따른 얼굴인식 기술은 얼굴검출, 전처리(preprocessing), 특징추출, 인식 또는 매칭의 과정을 통해 수행된다.

[0023] 이러한 얼굴인식 기술은 사전에 등록된 갤러리 얼굴 영상과 실제 입력으로 들어오는 프로브 얼굴 영상을 비교하여 인물 정보를 인식하는 기술과, 입력으로 두 장의 영상이 들어왔을 때 동일 인물인지 여부를 판단하는 얼굴검증 기술로 분류된다.

[0024] 이러한 환경에서 사전에 등록된 갤러리 얼굴 영상 정보들은 상대적으로 제약된 환경(고정된 조도, 촬영위치 등)에서 촬영된 반면, 입력으로 들어오는 프로브 얼굴 영상의 경우에는 조도변화, 포즈변화, 저해상도 등 다양한 환경에서 취득되어 열화된(degraded) 영상이 입력된다.

[0025] 종래 기술에 따르면, 이러한 환경에서 효과적인 얼굴인식을 수행하기 위해 다양한 환경 변화를 보상 시키기 위한 전처리 기술(조명보정 및 필터링, 포즈보정, 초 해상화 등), 환경 변화에 강인한 특징 추출 기법 등이 주로 개발되어 왔다.

[0026] 하지만, 전처리 기술을 통해서는 실제 발생하는 모든 변화를 다룰 수 없고, 전처리 알고리즘은 실험적으로(heuristically) 고안된 것으로 모든 문제를 자동적으로 탐지하여 보정하는데 한계가 있다.

[0027] 또한, 환경 변화에 강인한 특징을 추출하기 위해 심층 학습(deep learning) 기반 방법이 개발되고 있으나, 학습

에 사용되는 얼굴 영상들을 갤러리 영상과 비교할 때, 스타일의 차이가 존재하며, 강인한 특징 추출기를 학습시키기 위해서는 다양한 변화를 포함하는 대량의 데이터가 요구되는 문제점이 있다.

[0029] 본 발명은 전술한 문제점을 해결하기 위하여 제안된 것으로, 스마트 관제 또는 출입통제 시스템에서 얼굴인식 수행 시에 사전에 등록된 갤러리(gallery) 얼굴 영상들을 이용한 학습을 통해, 실세계에서 취득된 다양한 변화를 갖게 되는 프로브(probe) 얼굴 영상의 도메인 적응(domain adaptation)을 통해 갤러리 얼굴 영상 스타일과 유사한 새로운 영상 또는 특징을 생성시킴으로써, 갤러리 영상과 프로브 영상 사이의 불일치(mismatch) 문제를 줄이고, 효과적인 인식/매칭(matching)을 수행하는 것이 가능한 도메인 적응 기반 객체 인식 장치 및 그 방법을 제안한다.

[0030] 본 발명의 실시예에 따르면, 제한된 집합의 갤러리 얼굴 영상과 프로브 얼굴 영상을 이용하여 갤러리 얼굴 영상의 스타일을 학습하고, 학습된 모델을 이용하여 프로브 얼굴 영상의 도메인 적응을 통해, 갤러리 얼굴 영상들의 스타일과 유사한 새로운 영상 또는 특징을 생성한다.

[0031] 도메인 적응은 복수의 도메인이 존재할 때 서로 다른 도메인과 유사한 데이터를 생성하거나, 특정 도메인에서 학습된 모델이 다른 도메인에서 사용될 때 효과적으로 작동하게 하는 기술이다.

[0032] 본 발명의 실시예에 따르면, 생성된 프로브 얼굴 영상과 갤러리 얼굴 영상의 특징 추출에 따라 얼굴인식을 수행하게 되며, 갤러리 얼굴 영상 및 프로브 얼굴 영상 사이의 불일치를 줄임으로써 효과적인 얼굴인식이 가능하다.

[0034] 도 1은 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 장치를 나타내는 블록도이다.

[0035] 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 장치는 도메인 적응 기반 객체 인식 프로그램이 저장된 메모리(100) 및 프로그램을 실행시키는 프로세서(200)를 포함하되, 프로세서(200)는 입력 프로브 영상에 대해 도메인 적응 기반으로 갤러리 영상과 유사한 영상 또는 특징으로 생성시키기 위한 생성모델을 학습하고, 갤러리 영상과 프로브 영상의 학습 데이터베이스를 이용하여 객체인식 분류 모델을 학습하여, 입력 프로브 영상을 이용한 객체 인식을 수행하는 것을 특징으로 한다.

[0036] 프로세서(200)는 객체의 특징 정보를 이용한 전처리를 수행하여 학습 데이터베이스를 구축하고, 갤러리 영상 데이터베이스와 갤러리에 등록되지 않은 외부 영상 데이터베이스를 활용하여 전처리를 수행한 결과에 따라 생성모델을 학습한다.

[0037] 프로세서(200)는 입력되는 영상이 학습 데이터베이스에 포함되었는지 여부를 판별하여 영상 소스를 분류하고, 도메인 적응 기반의 새로운 영상 또는 특징을 생성하며, 객체 ID를 판별하여 갤러리 영상의 스타일을 학습한다.

[0038] 프로세서(200)는 학습 데이터베이스에 대해 전처리 수행, 특징 추출 수행에 따라 객체 ID 분류기를 학습시켜, 객체인식 분류 모델을 학습한다.

[0039] 이 때, 객체 ID 분류기를 통해 출력된 결과에 대해 오류 계산을 수행하여, 객체 ID 분류기를 학습시키게 된다.

[0040] 프로세서(200)는 수신된 입력 영상으로부터 객체 영역을 검출하고, 생성모델을 이용하여 입력 영상을 갤러리 영상과 유사한 새로운 영상 또는 특징으로 생성하고, 객체인식 분류 모델을 이용한 특징 추출을 수행하여, 객체의 ID 정보를 획득한다.

[0042] 도 2는 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 장치를 나타내는 블록도이다.

[0043] 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 장치는 모델 학습과 분류 및 매칭에 필요한 얼굴 영상들을 수집하는 데이터 수집부(210)와, 갤러리 얼굴 영상의 스타일을 학습하여 입력 프로브 영상에 대해 도메인 적응을 통해 갤러리 얼굴 영상과 유사한 새로운 얼굴 영상을 생성시키기 위한 생성모델 학습부(220)와, 얼굴 인식 및 매칭을 수행하기 위해 사전에 등록이 필요한 인물 정보를 이용하여 분류모델을 학습하는 분류모델 학습부(230) 및 생성모델과 분류모델을 이용하여 실제 입력 얼굴 영상에 대한 인식을 수행하는 인식 검증부(240)를 포함한다.

[0045] 도 3은 본 발명의 실시예에 따른 데이터 수집부를 나타내는 블록도이다.

[0046] 본 발명의 실시예에 따른 데이터 수집부는 얼굴 영상에 대해 특징점 정보를 이용하여 전처리를 수행하고, 갤러리 얼굴 영상 데이터베이스를 구축한다.

[0047] 도 3을 참조하면, 데이터 수집부는 얼굴 검출기(211), 전처리기(212)를 포함한다.

[0048] 얼굴 검출기(211)는 입력 영상(I)에 대해 얼굴이 존재하는 영역을 검출하고, 전처리기(212)는 검출된 얼굴 영상에 대해 얼굴의 특징점(feature point) 정보를 이용한 얼굴 정렬 또는 밝기 값 정규화와 같은 전처리를 수행하여 갤러리 얼굴 영상 데이터베이스(213)를 구축한다.

[0049] 데이터 수집부는 얼굴인식을 위해 사전에 등록시킬 인물에 대해 오프라인으로 촬영을 통해 갤러리 얼굴 영상 데이터베이스를 구축하거나, 추가적으로 훈련에 필요한 영상을 웹으로부터 확보하는 것이 가능하다.

[0051] 도 4는 본 발명의 실시예에 따른 생성모델 학습부를 나타내는 블록도이다.

[0052] 생성모델 학습부는 갤러리 얼굴 영상 데이터베이스(213)와 외부 얼굴 영상 데이터베이스(214)를 이용하여 얼굴 영상 생성모델을 학습하며, 전처리기(221), 영상소스 분류기(222), 얼굴영상 생성기(223), 얼굴 ID 분류기(224), 오류 계산 및 학습기(225)를 포함한다.

[0053] 생성모델 학습부는 입력된 영상이 갤러리 얼굴 영상 데이터베이스에 포함되는지 여부를 판별하고, 학습된 갤러리 얼굴 영상의 스타일과 유사하게 새로운 얼굴 영상을 생성하고, 입력된 영상의 ID를 판별하여 얼굴 영상 생성모델을 학습한다.

[0054] 전처리기(221)는 사전에 구축된 갤러리 얼굴 영상 데이터베이스(213)와 갤러리에 등록된 얼굴이 아닌 외부 얼굴 영상 데이터베이스(214)를 활용하여 전처리(픽셀 값 정규화, 영상 크기 정규화 등)를 수행한다.

[0055] 본 발명의 실시예에 따르면, 얼굴영상 생성모델(226)은 generative adversarial network 학습 방식으로 학습되는데, 이 모델을 학습하기 위해 영상소스 분류기(222), 얼굴영상 생성기(223), 얼굴 ID 분류기(224)의 3가지 모델이 동시에 학습된다.

[0056] 영상소스 분류기(222)는 입력으로 들어오는 영상이 갤러리 영상 데이터베이스에 포함되는지 여부를 판별하고, 얼굴영상 생성기(223)는 새로운 영상 또는 특징을 생성하는 모델이 되며, 얼굴 ID 분류기(224)는 입력 얼굴 영상의 ID를 판별한다.

[0057] 얼굴 ID 분류기(224)는 입력 얼굴 영상의 ID를 판별함으로써, 얼굴영상 생성 시 ID를 유지하면서도 스타일이 비슷한 영상으로 생성하도록 만드는 것이다.

[0058] 오류 계산 및 학습기(225)는 전술한 3가지의 모델을 통해 출력되는 결과로부터 오류를 계산하고, 반복적으로 학습을 수행하여, 영상소스 분류기(222) 학습을 통해 갤러리 얼굴 영상의 스타일을 학습하는 것과 동시에, 자신의 ID 정보는 잃지 않는 얼굴영상 생성모델(226)을 학습한다.

[0060] 도 5는 본 발명의 실시예에 따른 분류모델 학습부를 나타내는 블록도이다.

[0061] 본 발명의 실시예에 따른 분류모델 학습부는 갤러리 얼굴 영상 데이터베이스(213)를 이용한 전처리 및 특징 추출에 따른 얼굴 ID 분류 결과에 따라 오류 계산을 수행하여, 얼굴 ID 분류기(233)를 학습시키고, 사전에 등록이 필요한 인물 정보를 이용하여 얼굴인식 분류모델(235)을 학습한다.

[0062] 전처리기(231)는 사전에 수집된 갤러리 얼굴 영상 데이터베이스(213)에 대해 전처리를 수행하고, 특징 추출기(232)의 특징추출 후에 얼굴 ID 분류기(233)를 통해 나온 출력을 이용하여, 오류 계산을 통해 얼굴 ID 분류기(233)를 학습시키게 된다.

[0063] 이 때, 딥 네트워크(deep network)의 경우에는 특징 추출기(232) 및 얼굴 ID 분류기(233) 모두 신경망으로 구성되며, 초기 값은 대용량의 얼굴 데이터로 학습된 백본 네트워크(예: VGG Face)를 이용하여 세팅된다.

[0065] 도 6은 본 발명의 실시예에 따른 인식 검증부를 나타내는 블록도이다.

[0066] 인식 검증부의 얼굴 검출기는 비디오 입력으로부터 얻은 각 프레임으로부터 얼굴 영역을 검출하고, 얼굴영상 생성기(243)는 얼굴영상 생성모델(226)을 이용하여 입력된 얼굴 영상이 갤러리 얼굴 영상과 유사하도록 새로운 얼굴 영상을 생성하고, 특징추출 및 매칭기(244)는 얼굴인식 분류모델(235)을 이용하여 특징 추출 및 매칭을 수행하여, ID 정보(245)를 획득한다.

[0068] 도 7은 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 방법을 나타내는 순서도이다.

[0069] 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 방법은 객체 영상을 수집하는 단계(S710)와, 갤러리 영상의 스타일을 학습하여 생성모델을 학습하는 단계(S720)와, 객체 인식을 위해 사전에 등록이 필요한 정보를 이용하여 분류모델을 학습하는 단계(S730) 및 생성모델과 분류모델을 이용하여 영상 내 객체를 인식하는 단계(S740)를 포함한다

[0070] S710 단계는 특징점 정보를 이용하여 객체 영상에 대한 전처리를 수행하고, 갤러리 영상 데이터베이스를 구축한다

[0071] S720 단계는 갤러리 영상 데이터베이스와 외부 영상 데이터베이스를 이용하여, 입력 영상을 갤러리 영상의 스타일에 부합되는 새로운 영상 또는 특징으로 생성하기 위한 생성모델을 학습한다

[0072] S730 단계는 갤러리 영상 데이터베이스를 이용하여 전처리 및 특징 추출을 수행하고, ID 분류 결과에 따른 오류 계산을 수행하여 ID 분류기를 학습시킨다.

[0073] S740단계는 비디오 입력으로부터 얻은 각 프레임으로부터 객체 영역을 검출하고, 생성모델을 이용하여 객체가 갤러리 영상과 유사하도록 새로운 영상 또는 특징을 생성하고, 분류모델을 이용하여 특징을 추출하고 매칭을 수행하여, 객체의 ID 정보를 획득한다.

[0075] 한편, 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 방법은 컴퓨터 시스템에서 구현되거나, 또는 기록매체에 기록될 수 있다. 컴퓨터 시스템은 적어도 하나 이상의 프로세서와, 메모리와, 사용자 입력 장치와, 데이터 통신 버스와, 사용자 출력 장치와, 저장소를 포함할 수 있다. 전술한 각각의 구성 요소는 데이터 통신 버스를 통해 데이터 통신을 한다.

[0076] 컴퓨터 시스템은 네트워크에 커플링된 네트워크 인터페이스를 더 포함할 수 있다. 프로세서는 중앙처리 장치(central processing unit (CPU))이거나, 혹은 메모리 및/또는 저장소에 저장된 명령어를 처리하는 반도체 장치일 수 있다.

[0077] 메모리 및 저장소는 다양한 형태의 휘발성 혹은 비휘발성 저장매체를 포함할 수 있다. 예컨대, 메모리는 ROM 및 RAM을 포함할 수 있다.

[0078] 따라서, 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 방법은 컴퓨터에서 실행 가능한 방법으로 구현될 수 있다. 본 발명의 실시예에 따른 도메인 적응 기반 객체 인식 방법이 컴퓨터 장치에서 수행될 때, 컴퓨터로 판독 가능한 명령어들이 본 발명에 따른 객체 인식 방법을 수행할 수 있다.

[0079] 한편, 상술한 본 발명에 따른 도메인 적응 기반 객체 인식 방법은 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현되는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록 매체로는 컴퓨터 시스템에 의하여 해독될 수 있는 데이터가 저장된 모든 종류의 기록 매체를 포함한다. 예를 들어, ROM(Read Only Memory), RAM(Random Access Memory), 자기 테이프, 자기 디스크, 플래시 메모리, 광 데이터 저장장치 등이 있을 수 있다. 또한, 컴퓨터로 판독 가능한 기록매체는 컴퓨터 통신망으로 연결된 컴퓨터 시스템에 분산되어, 분산방식으로 읽을 수 있는 코드로서 저장되고 실행될 수 있다.

[0081] 이제까지 본 발명의 실시예들을 중심으로 살펴보았다. 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 개시된 실시예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

부호의 설명

[0083] 100: 메모리 200: 프로세서

210: 데이터 수집부 211: 얼굴 검출기

212: 전처리기 213: 갤러리 얼굴 영상 DB

214: 외부 얼굴 영상 DB 220: 생성모델 학습부

221: 전처리기 222: 영상소스 분류기

223: 얼굴영상 생성기 224: 얼굴 ID 분류기

225: 오류 계산 및 학습기 226: 얼굴영상 생성모델

230: 분류모델 학습부 231: 전처리기

232: 특징 추출기 233: 얼굴 ID 분류기

234: 오류 계산 및 학습기 235: 얼굴인식 분류모델

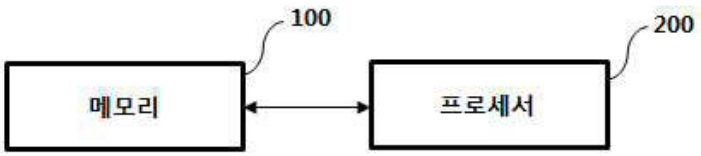
240: 인식 검증부 241: 얼굴 검출기

242: 전처리기 243: 얼굴영상 생성기

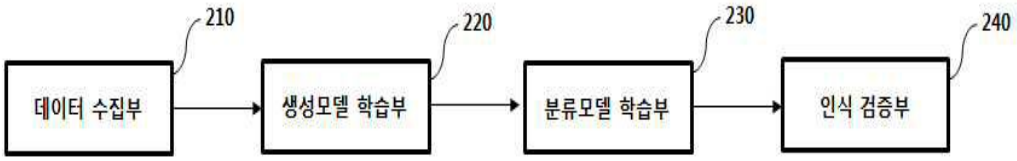
244: 특징추출 및 매칭기 245: ID 정보

도면

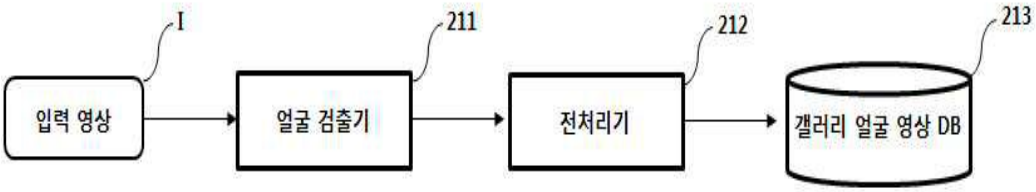
도면1



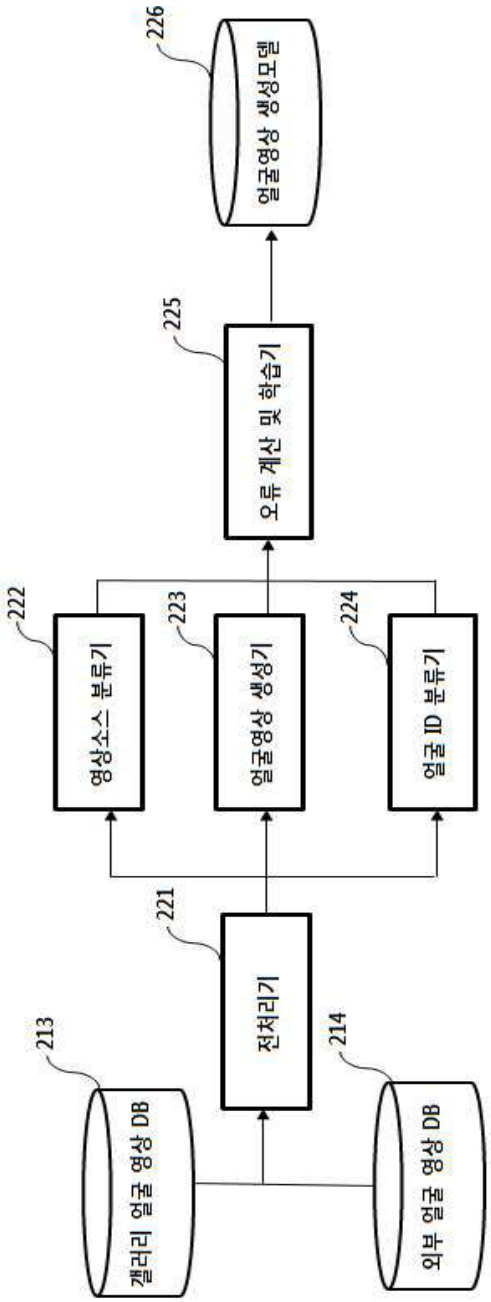
도면2



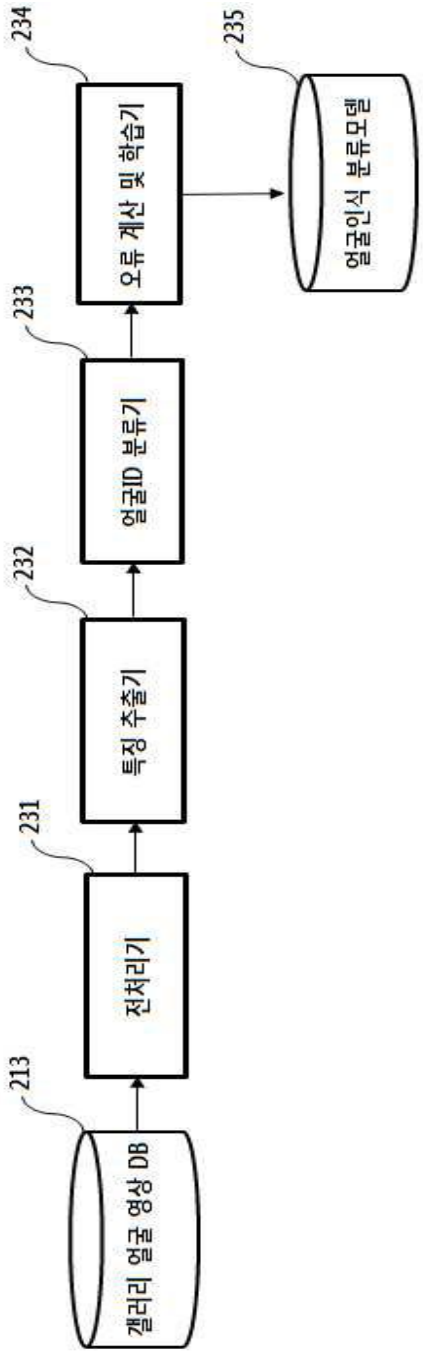
도면3


I
211
212
213
입력 영상
얼굴 검출기
전처리기
갤러리 얼굴 영상 DB

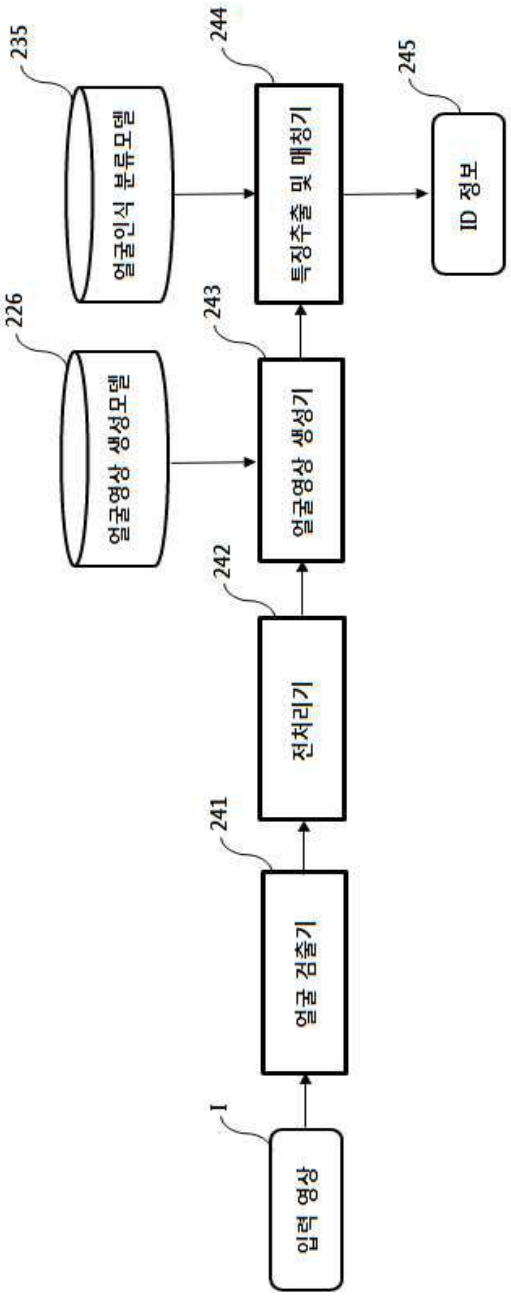
도면4


226
얼굴영상 생성모델
225
오류 계산 및 학습기
222
영상소스 분류기
223
얼굴영상 생성기
224
얼굴 ID 분류기
221
전처리기
213
갤러리 얼굴 영상 DB
214
외부 얼굴 영상 DB

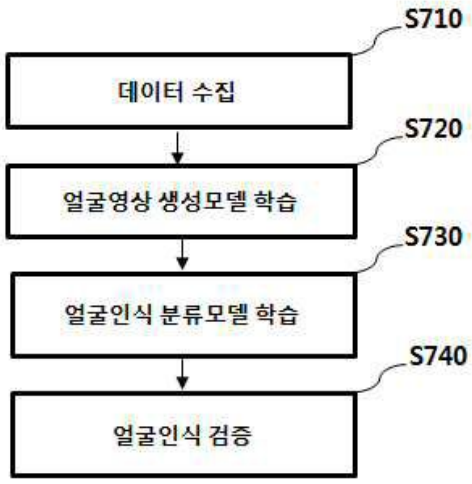
도면5


213
갤러리 얼굴 영상 DB
231
전처리기
232
특징 추출기
233
얼굴ID 분류기
234
오류 계산 및 학습기
235
얼굴인식 분류모델

도면6


235
244
245
얼굴인식 분류모델
특징추출 및 매칭기
ID 정보
226
243
얼굴영상 생성모델
얼굴영상 생성기
242
전처리기
241
얼굴 검출기
I
입력 영상

도면7



[별첨 5]

딥러닝 기반 해상감시영상 내 선박 검출 식별 기술 2.0

ETRI 한국전자통신연구원
www.etri.re.kr
차세대콘텐츠연구본부

목 차

1. 기술의 개요
2. 기술이전 내용 및 범위
3. 경쟁기술과 비교
4. 기술의 사업성
5. 국내외 시장 동향

▣ 기술 개요

학습 콘텐츠 생성

해상 선박/구조물 3차원 모델 데이터 구축
·제한 자료 기반 한국 군함 3차원 모델 개발
·가상 선박 모델 재현 및 이미지 생성 시스템 설계

실사 해상 선박/구조물 데이터 수집 및 정제 기술 개발
·해상 선박/구조물에 대한 실사 영상 구축 및 학습데이터 변환 기술 개발
·학습 콘텐츠 제작을 위한 편집 도구 개발

기상 효과 생성·제거

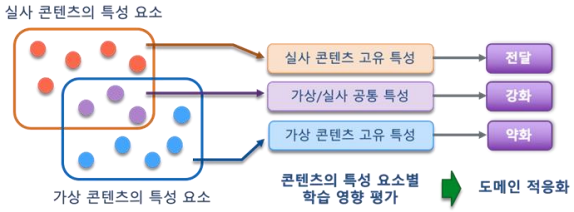
해상 선박/구조물 검출을 위한 영상 전처리 기술 개발
·환경 요건에 의한 영상 콘텐츠 화질 저하 개선 기술 개발

기상/환경요소 부여/제거를 위한 가상 콘텐츠 고유 특성 분석
·가상 기상 및 환경 요소 생성제거 기술 개발

해상구조물/선박 인식

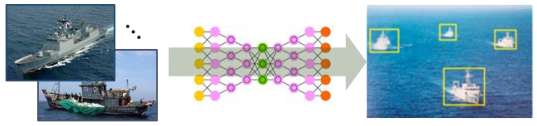
실사 학습데이터 기반 해상 객체 검출/식별 원천기술 개발
·실사 콘텐츠 기반 선박 식별 모델 학습기 개발

기상/환경요소 부여/제거를 위한 가상 콘텐츠 고유 특성 분석
·가상/실사 콘텐츠 도메인별 특성 요소 분석 기술 개발



유사성이 높은 객체간 상세 식별

실사 학습데이터 기반 해상 객체 검출/식별 원천기술 개발
·일반 선박 및 해상 구조물 식별 기술
·원거리 객체 검출을 위한 소형 객체 인식 기술



기술이전 대상 기술

1. 기술의 개요

▣ 딥러닝 기반 해상감시영상 내 선박 검출/식별 기술 2.0

❖ 기술의 목적

- 해군/육군의 해안 감시영상으로부터 선박을 탐지하고 식별하기 위한 기술
- 적국의 주요 군함에 대한 식별에 앞서 일상적인 감시 과정에서 발견되는 선박을 인공지능 기반으로 자동 검출/식별하기 위한 기반 기술

❖ 필요성

- 현재 병역자원(인력, 예산)은 지속적으로 감축되고 있으나 해상을 통한 안보의 위협은 증가하고 있는 상황
- 인공지능 기반의 해상 감시를 통해 해군의 효과적인 해상 감시체계 구축 필요

❖ 학습 데이터 규모와 학습 모델 튜닝을 통한 성능 향상버전

- 검출성능 (11종 객체 → 12종 객체, 79.8% → 82.3%)
- 식별성능 (11종 객체 → 12종 객체, 85.84% → 91.94%)

1. 기술의 개요

▣ 기술 성숙도 (TRL : 5 단계)

구 분	단계	정 의	세 부 설 명
기초 연구 단계	1	기초 이론/실험	▪기초이론 정립 단계
	2	실용 목적의 아이디어 특허 등 개념정립	▪기술개발 개념 정립 및 아이디어에 대한 특허 출원 단계
실험 단계	3	실험실 규모의 기본성능 검증	▪실험실 환경에서 실험 또는 전산 시뮬레이션을 통해 기본성능이 검증될 수 있는 단계 ▪개발하려는 부품/시스템의 기본 설계도면을 확보하는 단계
	4	실험실 규모의 소재/부품/시스템 핵심성능 평가	▪시험샘플을 제작하여 핵심성능에 대한 평가가 완료된 단계 ▪3단계에서 도출된 다양한 결과 중에서 최적의 결과를 선택하려는 단계 ▪컴퓨터 모사가 가능한 경우 최적화를 완료하는 단계
시작품 단계	5	확정된 소재/부품/시스템 시작품 제작 및 성능 평가	▪확정된 소재/부품/시스템의 실험실 시작품 제작 및 성능 평가가 완료된 단계 ▪개발 대상의 생산을 고려하여 설계하나 실제 제작한 시작품 샘플은 1~수개 미만인 단계 ▪경제성을 고려하지 않고 기술의 핵심성능으로만 볼 때, 실제로 판매가 될 수 있는 정도로 목표 성능을 달성한 단계
	6	파일롯 규모 시작품 제작 및 성능 평가	▪파일롯 규모(복수 개~양산규모의 1/10정도)의 시작품 제작 및 평가가 완료된 단계 ▪파일롯 규모 생산품에 대해 생산량, 생산용량, 불량률 등 제시 ▪파일롯 생산을 위한 대규모 투자가 동반되는 단계 ▪생산기업이 수요기업 적용환경에 유사하게 자체 현장테스트를 실시하여 목표 성능을 만족시킨 단계 ▪성능 평가 결과에 대해 가능하면 공인인증 기관의 성적서 확보
실용화 단계	7	신뢰성평가 및 수요기업 평가	▪실제 환경에서 성능 검증이 이루어지는 단계 ▪부품 및 소재개발의 경우 수요업체에서 직접 파일롯 시작품을 현장 평가(성능 및 신뢰성 평가) ▪가능하면 인증기관의 신뢰성 평가 결과 제출
	8	시제품 인증 및 표준화	▪표준화 및 인허가 취득 단계
사업화	9	사업화	▪본격적인 양산 및 사업화 단계 ▪6-시그마 등 품질관리가 중요한 단계

2. 기술이전 내용 및 범위

▣ 기술이전 내용

❖ 기술명: 딥러닝 기반 해상감시영상 내 선박 검출/식별 기술 2.0

- (1 세부기술) 딥러닝 기반 해상감시영상 내 선박추정객체 검출기술 2.0
  - ✓ 해상 영상 내 선박추정 객체의 위치 특정 기능
  - ✓ 위치가 특정된 객체의 정보 데이터 저장 기능
- (2 세부기술) 딥러닝 기반 해상감시영상 내 선박종류 식별기술 2.0
  - ✓ 입력된 선박 영상에 대한 12종의 선종 기준 분류 기능
  - ✓ 분류된 선종에 대한 식별기의 판단 근거 가시화 기능

2. 기술이전 내용 및 범위

▣ 기술이전 범위

❖ 1, 2 세부기술은 소스코드 및 학습 네트워크 형태로 기술이전 가능

대상 기술	기술이전 범위
세부기술 1	• 도커(docker) 기반 개발환경 바이너리 및 도커 이미지 생성 스크립트 • 선박추정객체 검출기 - 학습(training) / 실행(test) 코드 및 학습된 네트워크 • 제공된 기술에 대한 사용설명서
세부기술 2	• 도커(docker) 기반 개발환경 바이너리 및 도커 이미지 생성 스크립트 • 선박식별기 - 학습(training) / 실행(test) 코드 및 학습된 네트워크 • 선박식별 분석용 Grad-CAM 가시화 코드 • 제공된 기술에 대한 사용설명서
기술문서	• 2550-2019-00302 (객체 검출을 위한 학습데이터 생성 방법, 세부기술 1 대상) • 2550-2019-00306 (해상 선박 및 구조물 분류 및 ID, 세부기술 2 대상) • 2550-2019-00784 (요구사항 정의서, 공통) • 2550-2019-00785 (시스템 시험 계획서, 공통) • 2550-2019-00786 (시스템 시험 절차 및 결과서, 공통)
특허	• 영상에 기반하여 선박을 식별하는 방법 및 장치 (2019-0110110) • 객체 검출 방법 및 장치 (2020-0036066) • 촬영 방향 추정을 통한 가상 군함 유사도 비교 기반 군함 함종/함급 식별 방법 (2020-0022249)

2. 기술이전 내용 및 범위

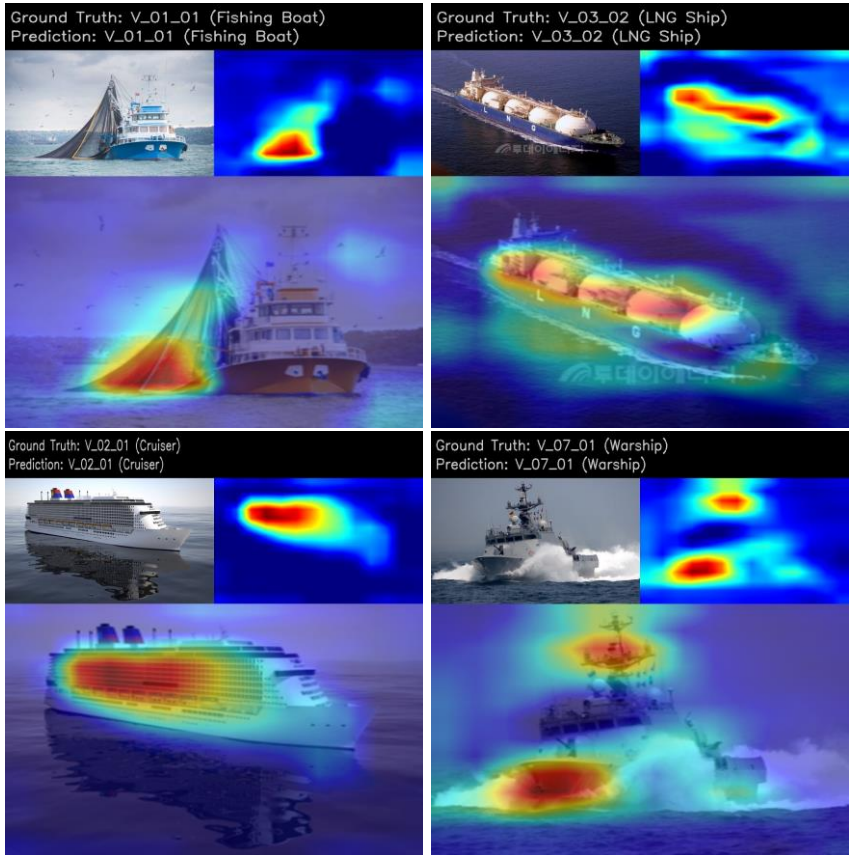
▣ 1 세부기술

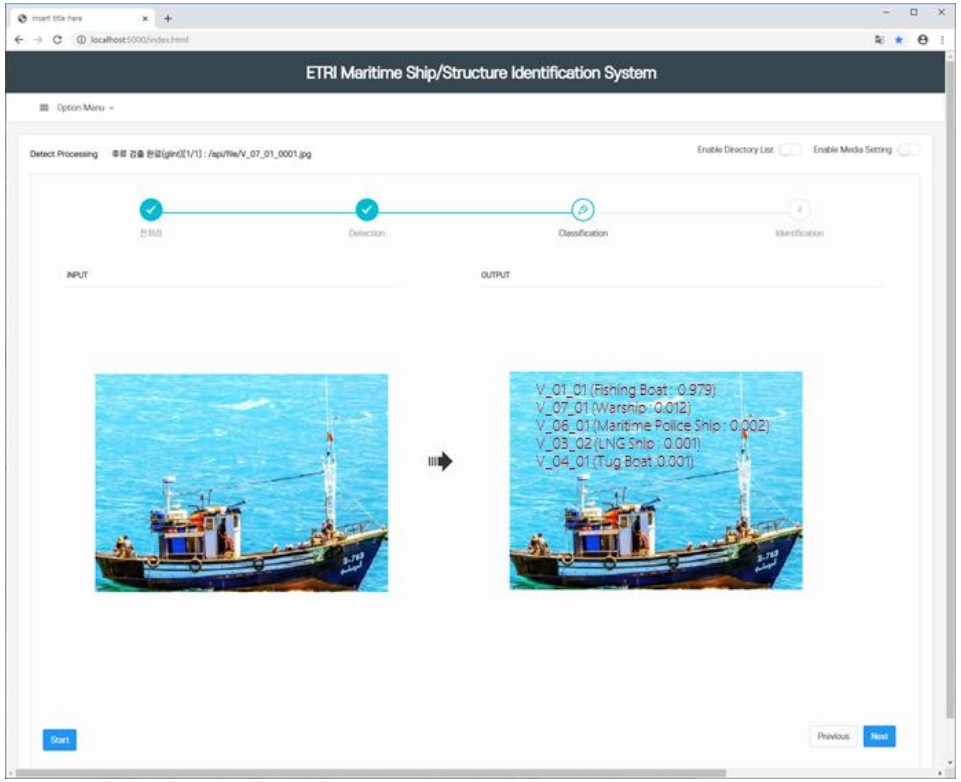
❖ 딥러닝 기반 해상감시영상 내 선박추정객체 검출기술 2.0

2. 기술이전 내용 및 범위

▣ 2 세부기술

❖ 딥러닝 기반 해상감시영상 내 선박종류 식별기술 2.0





2. 기술이전 내용 및 범위

▣ 기술료 조건

- ❖ (1 세부기술) 딥러닝 기반 해상감시영상 내 선박추정객체 검출기술 2.0
- ❖ (2 세부기술) 딥러닝 기반 해상감시영상 내 선박종류 식별기술 2.0
- ❖ 기존 기술 (1.0) 이전 기업은 차액만 지불하고 기술이전 가능
- ❖ 실질기여 공동연구 참여기업은 기술료 20% 감면

구분	실질기여 공동연구 참여기업			일반기업		
	중소기업	중견기업	대기업	중소기업	중견기업	대기업
(1 세부기술) 딥러닝 기반 해상감시영상 내 선박추정객체 검출기술 2.0				40,000	80,000	80,000
(2 세부기술) 딥러닝 기반 해상감시영상 내 선박종류 식별기술 2.0				50,000	100,000	100,000
매출정률사용료(%)				1.25	3.75	5

3. 경쟁기술과 비교

▣ 선박 검출/식별 기술

❖ UNINA (이탈리아)

- SAR 영상 내에서 선박 추정 객체 탐지 기술 개발
- 선박 탐지 성능은 우수하나 선박의 종류나 세부 정보 식별이 불가능

❖ BUCT (중국)

- 다수의 특징점 조합을 통해 항공/선상영상에 대한 선종분류기술 개발
- 유사한 외관의 선박에 대한 식별 불가능

❖ ViNotion社 (네덜란드)

- 항구의 PTZ 카메라를 이용한 연안선박 탐지 기술 개발
- 선박의 크기가 큰 경우 선수/선미를 별개의 선박으로 오인식
- 선미 부분의 파도를 선박의 일부로 오인식

❖ 기존 경쟁기술 대비 개량된 부분

- 세계 최고 수준의 선박 검출률 보유
- 유사한 선박에 대한 선종 식별 지원 (여객선, 화물선 세부 식별 수행 및 군함에 대한 세부 식별 진행으로 기술 고도화 예정)

4. 기술의 사업성

▣ 국내외 시장

❖ 예상 응용 제품 및 서비스

- 선박 식별 교육/훈련 시스템
- 운항 안전 모니터링 시스템
- 지능형 상시 감시 체계

선박 식별 교육/훈련 시스템


운항 안전 모니터링 시스템


지능형 상시 감시 체계

4. 기술의 사업성

▣ 국내외 시장

❖ 사업성

- 해군, 육군, 해경, 항만 등에서 해상 감시를 위한 광학 장비가 이미 운용되고 있으며 여기서 획득한 영상을 활용할 수 있으므로 컴퓨팅 장비를 제외한 추가 비용은 매우 낮음

❖ 기술이전 업체 조건

- 방위산업 분야에 대한 지식/경험 필요

❖ 사업화시 제약 조건

- 레이더, 전파 기반의 해상 감시체계와의 연동을 통한 통합 감시 시스템 개발 필요
- 사용처의 요구와 편의성을 반영한 사용자 인터페이스 개발 필요

5. 국내외 시장 동향

▣ 국내외 동향

❖ 세계 해양안전 및 방산시장 규모 성장 기대

- 세계 해양안전 시장(Maritime safety market) 규모는 2016년 1,671억 달러에서 연 평균 7.2% 성장을 통해 2021년 2,367억 달러 규모로 성장할 것으로 추정 (MarketsandMarkets, 2016)
- 향후 10년간 전자광학 · 적외선 장비 시장 규모는 약 198억 달러, 이 중 약 25% 수준인 49억 달러가 지상/해상용 장비가 차지할 것으로 추정 (세계 방산시장 연감, 2017)

감사합니다.

www.etri.re.kr

♣ 연락처 : 차세대콘텐츠연구본부, 남도원 책·연 (042-860-4994, dwnam@etri.re.kr)

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0119672
(43) 공개일자 2021년10월06일

(51) 국제특허분류(Int. Cl.)
G06T 7/20 (2017.01) *G06N 3/08* (2006.01)
(52) CPC특허분류
G06T 7/20 (2013.01)
G06N 3/08 (2013.01)
(21) 출원번호 10-2020-0036066
(22) 출원일자 2020년03월25일
심사청구일자 2021년01월26일
기술이전 희망 : 기술양도

(71) 출원인
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
문성원
대전광역시 유성구 노은서로210번길 32, 405동 1005호 (지족동, 열매마을아파트4단지)
이지원
대전광역시 유성구 반석서로 98, 602동 2203호 (반석동, 반석마을6단지아파트)
(74) 대리인
팬코리아특허법인

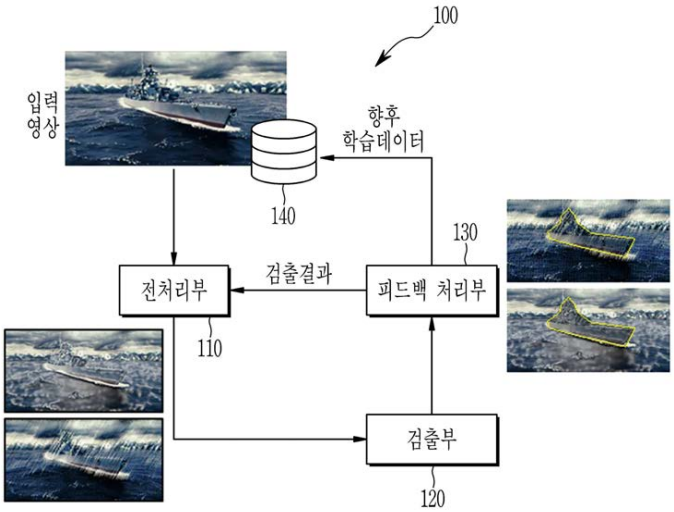
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **객체 검출 방법 및 장치**

(57) 요 약

객체 검출 장치는 객체 검출 성능을 향상시키는 방향으로 학습된 전처리 신경망을 이용해 입력되는 영상에 대해 해당 영상의 통계적 특성을 변화시키는 영상 전처리를 수행하고, 상기 영상 전처리된 영상으로부터 객체를 검출하며, 상기 객체의 검출 결과를 상기 전처리 신경망으로 피드백한다.

대 표 도 - 도1



(52) CPC특허분류
G06T 2207/20084 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	2019-0-00524
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원(IITP)
연구사업명	가상증강현실콘텐츠원천기술개발
연구과제명	영상 내 객체간 관계 분석 기반 해상 선박/구조물 상세 식별 콘텐츠 기술 개발
기 여 율	1/1
과제수행기관명	한국전자통신연구원
연구기간	2020.01.01~2020.12.31

명 세 서

청구범위

청구항 1

객체 검출 장치에서 입력되는 영상으로부터 객체를 검출하는 방법으로서,

객체 검출 성능을 향상시키는 방향으로 학습된 전처리 신경망을 이용해 입력되는 영상에 대해 해당 영상의 통계적 특성을 변화시키는 영상 전처리를 수행하는 단계,

상기 영상 전처리된 영상으로부터 객체를 검출하는 단계, 그리고

상기 객체의 검출 결과를 상기 전처리 신경망으로 피드백하는 단계

를 포함하는 객체 검출 방법.

청구항 2

제1항에서,

상기 검출하는 단계는 객체 검출을 위해 학습된 객체 검출 신경망을 이용하여 상기 영상 전처리된 영상으로부터 상기 객체를 검출하는 단계를 포함하는 객체 검출 방법.

청구항 3

제1항에서,

상기 피드백하는 단계는 상기 객체의 검출 결과를 토대로 상기 객체 검출 성능을 향상시키는 방향으로 상기 영상의 통계적 특성이 변화되도록 상기 전처리 신경망을 학습시키는 단계를 포함하는 객체 검출 방법.

청구항 4

제1항에서,

상기 검출 결과는 검출된 객체의 클래스 및 상기 검출된 객체에 대한 손실값을 포함하고,

상기 피드백하는 단계는 상기 손실값을 줄이는 방향으로 상기 전처리 신경망을 학습시키는 단계를 포함하는 객체 검출 방법.

청구항 5

제1항에서,

상기 통계적 특성은 색상값, 채도값, 밝기값, 화이트밸런스, 확률분포 자체 중 적어도 하나를 포함하는 객체 검출 방법.

청구항 6

제1항에서,

상기 영상은 해상에서 촬영된 영상을 포함하고, 상기 객체는 선박을 포함하는 객체 검출 방법.

청구항 7

입력된 영상으로부터 객체를 검출하는 객체 검출 장치로서,

신경망 기반 객체 검출에 최적화된 형태로 영상을 변조시키도록 학습된 전처리 신경망을 이용하여 입력되는 영상을 변조하여 출력하는 전처리부,

상기 전처리부로부터 출력된 영상으로부터 상기 신경망 기반 객체 검출을 수행하는 검출부, 그리고

상기 검출부의 객체 검출 결과를 상기 전처리부로 전달하는 피드백 처리부

를 포함하는 객체 검출 장치.

청구항 8

제7항에서,

상기 전처리부는 상기 입력되는 영상과 상기 입력되는 영상으로부터 검출된 객체 검출 결과를 이용하여, 상기 신경망 기반 객체 검출 성능을 향상시키는 방향으로 상기 전처리 신경망을 재학습시키는 객체 검출 장치.

청구항 9

제7항에서,

상기 전처리 신경망은 상기 영상을 변조시키기 위해 상기 영상의 통계적 특성을 변화시키는 객체 검출 장치.

청구항 10

제7항에서,

상기 검출 결과는 검출된 객체의 클래스 및 상기 검출된 객체에 대한 손실값을 포함하고,

상기 전처리부는 상기 손실값을 줄이는 방향으로 상기 전처리 신경망을 재학습시키는 객체 검출 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 객체 검출 방법 및 장치에 관한 것으로, 특히 다양한 기상효과가 존재하는 해상 영상을 사람의 육안이 아닌 선박 검출을 위한 기계 학습에 적합한 형태로 전처리하여 선박 검출 성능을 향상시킬 수 있는 객체 검출 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 운행 중인 선박이나 해안 초소, 관제 센터 등에서는 교통 통제나 해양 사고 지원 등 여러 가지 목적을 위해 주변 선박을 검출하는 방법이 필요하다. 이를 위해 일정 크기 이상의 선박에 대해 선박자동식별시스템(AIS, automatic identification system), 소형 선박에 대해 위치발신장치(V-pass) 등의 통신 기반 식별 방법을 적용하고 있으나 기기의 고장, 어장을 숨기기 위한 고의적인 미사용 등으로 민간 선박 정보를 획득하지 못하는 경우가 있다. 또한 적국 및 타국 군함의 경우 이러한 통신 기반으로 정보를 얻기는 어렵다. 이러한 경우 합성 개구 레이더, 위성 영상 등을 활용하는 기술이 제시되었으나 해당 정보의 획득에 오랜 시간이 걸려 해상 광학 영상에 대한 선박 검출이 필요하다.

[0003] 해상 환경은 육상과 달리 해무, 후류, 수면 반사광 등의 환경요소가 많으며 데이터의 획득 또한 어렵다. 해양 영상의 경우 해무 제거를 통해 영상의 화질을 개선하고자 하는데, 대체적으로 해무의 특성을 분석하여 해무와 운광을 분리하고 이에 따른 밝기 보상을 하는 방식으로 해무를 제거한다. 하지만 영상 내에 밝기가 극단적으로 차이 나는 경우 왜곡이 심해지는 문제가 있으며 또한 사람의 육안으로 보는 것에 적합한 형태의 전처리 기술이므로 해상 광학 영상내의 객체를 검출하기 위한 기술로는 부적합하다. 또한 전통적인 영상처리 기법을 활용하여 해상 영상을 전처리하는 방법이 제안되었는데, 대체적으로 단순히 기상효과를 제거하는 방법에만 초점을 맞추고 있다.

[0004] 디노이징, 블러처리, 안개제거와 같은 영상 전처리 기법의 경우 대부분 인간의 육안에 최적화된 출력을 위한 것이며, 이러한 전처리 기술들은 기계학습 성능을 높이기 위해 활용하기에는 어려움이 있다. 이는 기존 대부분의 전처리 기술이 가지고 있는 문제이기도 하다. 또한 학습데이터 부족을 해결하기 위해 영상에 기하학적인 변화를 가하거나 색상을 바꾸는 등의 데이터 증강 기법이 있으나 이 방법 또한 인간의 직관에 의존하고 증강기법의 사각(화이트밸런스 등) 변화에 취약하다.

[0005] 또한 많은 데이터 중 학습 효과가 높은 데이터를 선별적으로 학습하는 능동학습(Active learning) 기법이 유의미한 결과를 보이는 것을 통해 기계학습에 유효한 학습데이터가 존재하는 것을 확인하였으나 이는 기존 데이터

중 일부를 선별적으로 학습하는 기술이다.

발명의 내용

해결하려는 과제

[0006] 본 발명이 해결하려는 과제는 데이터 확보가 어려운 해상 환경에서 촬영된 영상으로부터 객체 검출 성능을 높일 수 있는 객체 검출 방법 및 장치를 제공하는 것이다.

과제의 해결 수단

[0007] 본 발명의 한 실시 예에 따르면, 객체 검출 장치에서 입력되는 영상으로부터 객체를 검출하는 방법이 제공된다. 객체 검출 방법은 객체 검출 성능을 향상시키는 방향으로 학습된 전처리 신경망을 이용해 입력되는 영상에 대해 해당 영상의 통계적 특성을 변화시키는 영상 전처리를 수행하는 단계, 상기 영상 전처리된 영상으로부터 객체를 검출하는 단계, 그리고 상기 객체의 검출 결과를 상기 전처리 신경망으로 피드백하는 단계를 포함한다.

[0008] 상기 검출하는 단계는 객체 검출을 위해 학습된 객체 검출 신경망을 이용하여 상기 영상 전처리된 영상으로부터 상기 객체를 검출하는 단계를 포함할 수 있다.

[0009] 상기 피드백하는 단계는 상기 객체의 검출 결과를 토대로 상기 객체 검출 성능을 향상시키는 방향으로 상기 영상의 통계적 특성이 변화되도록 상기 전처리 신경망을 학습시키는 단계를 포함할 수 있다.

[0010] 상기 검출 결과는 검출된 객체의 클래스 및 상기 검출된 객체에 대한 손실값을 포함하고, 상기 피드백하는 단계는 상기 손실값을 줄이는 방향으로 상기 전처리 신경망을 학습시키는 단계를 포함할 수 있다.

[0011] 상기 통계적 특성은 색상값, 채도값, 밝기값, 화이트밸런스, 확률분포 자체 중 적어도 하나를 포함할 수 있따.

[0012] 상기 영상은 해상에서 촬영된 영상을 포함하고, 상기 객체는 선박을 포함할 수 있다.

[0013] 본 발명의 다른 한 실시 예에 따르면, 입력된 영상으로부터 객체를 검출하는 객체 검출 장치가 제공된다. 객체 검출 장치는 전처리부, 검출부, 그리고 피드백 처리부를 포함한다. 상기 전처리부는 신경망 기반 객체 검출에 최적화된 형태로 영상을 변조시키도록 학습된 전처리 신경망을 이용하여 입력되는 영상을 변조하여 출력한다. 상기 검출부는 상기 전처리부로부터 출력된 영상으로부터 상기 신경망 기반 객체 검출을 수행한다. 그리고 상기 피드백 처리부는 상기 검출부의 객체 검출 결과를 상기 전처리부로 전달한다.

[0014] 상기 전처리부는 상기 입력되는 영상과 상기 입력되는 영상으로부터 검출된 객체 검출 결과를 이용하여, 상기 신경망 기반 객체 검출 성능을 향상시키는 방향으로 상기 전처리 신경망을 재학습시킬 수 있다.

[0015] 상기 전처리 신경망은 상기 영상을 변조시키기 위해 상기 영상의 통계적 특성을 변화시킬 수 있다.

[0016] 상기 검출 결과는 검출된 객체의 클래스 및 상기 검출된 객체에 대한 손실값을 포함하고, 상기 전처리부는 상기 손실값을 줄이는 방향으로 상기 전처리 신경망을 재학습시킬 수 있다.

발명의 효과

[0017] 본 발명의 실시 예에 의하면, 선박 검출 성능의 향상시키는 방향으로 학습된 신경망을 이용해 입력되는 해상 광학 영상의 통계적 특성을 변화시키는 영상 전처리를 수행한 후, 전처리된 영상에 대해 선박 검출을 시도함으로써, 선박 검출 성능을 향상시킬 수 있다.

[0018] 즉, 본 발명의 실시 예에 의하면, 인간의 직관이나 육안으로 보이는 정성적 결과에 대한 판단에 의존하지 않고 선박 검출 성능의 향상만을 목적으로 할 수 있는 영상 전처리 기법이 적용되기 때문에, 적용 현장에 최적화된 전처리 기법을 데이터 누적에 따라 자동으로 도출해낼 수 있다. 이는 기존의 통신/레이더로 검출/식별이 불가능하였던 선박을 광학 영상을 통해 검출하는 것에 활용 가능한 기술로 해상감시 응용에 사용될 수 있다.

도면의 간단한 설명

[0019] 도 1은 본 발명의 한 실시 예에 따른 객체 검출 장치를 나타낸 도면이다.

도 2는 본 발명의 실시 예에 따른 전처리 신경망의 학습 방법을 나타낸 도면이다.

도 3은 본 발명의 실시 예에 따른 객체 검출 방법을 나타낸 흐름도이다.

도 4는 본 발명의 다른 실시 예에 따른 객체 검출 장치를 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0020] 아래에서는 첨부한 도면을 참고로 하여 본 발명의 실시 예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시 예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0021] 명세서 및 청구범위 전체에서, 어떤 부분이 어떤 구성 요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성 요소를 더 포함할 수 있는 것을 의미한다.

[0022] 이제 본 발명의 실시 예에 따른 객체 검출 방법 및 장치에 대하여 도면을 참고로 하여 상세하게 설명한다.

[0023] 도 1은 본 발명의 한 실시 예에 따른 객체 검출 장치를 나타낸 도면이다.

[0024] 도 1을 참고하면, 객체 검출 장치(100)는 전처리부(110), 검출부(120) 및 피드백 처리부(130)를 포함한다. 객체 검출 장치(100)는 저장부(140)를 더 포함할 수 있다.

[0025] 전처리부(110)는 입력되는 영상을 전처리하고, 전처리된 영상을 검출부(120)로 전달한다. 본 발명의 한 실시 예에 따른 객체는 선박일 수 있으나, 사람 또는 구조물 등 이에 한정되지 않는다. 또한 입력되는 영상은 해양 환경에서 촬영된 해양 영상일 수 있으나, 이에 한정되지 않는다. 해양 환경에서 촬영된 영상에 적용되는 기존의 디노이징(denoising), 블러링(blurring) 처리, 안개제거와 같은 영상 전처리 기법은 대부분 인간의 육안에 최적화된 영상 출력을 위한 방법으로, 이러한 영상 전처리 기법들은 기계학습 성능을 높이기 위해 활용되기에는 어려움이 있다.

[0026] 본 발명의 실시 예에 따른 전처리부(110)는 기계학습 모델에 해당하는 신경망을 이용한 객체 검출에 최적화되도록 영상의 통계적 특성을 변화시키는 영상 전처리 과정을 수행한다. 영상의 통계적 특성은 예를 들면, 픽셀의 색상값, 픽셀의 밝기값, 픽셀의 채도값, 픽셀의 특성값에 대한 평균값, 색상 분포, 화이트 밸런스(White balance), 특성값에 대한 확률분포 등을 포함할 수 있으며, 전처리부(110)는 입력되는 영상에 대해 이들 중 적어도 하나의 요소를 변화시켜 신경망을 이용한 객체 검출에 최적화시킨다. 전처리부(110)는 신경망을 이용한 객체 검출에 최적화되도록 영상의 통계적 특성을 변화시키는 영상 전처리 자체를 학습시킨 신경망을 이용한다. 아래에서는 영상 전처리를 위해 학습된 신경망을 전처리 신경망이라 한다. 기존의 스타일 변화(Style transfer) 등의 기법에 사용된 적대적 생성 모델(Generative Adversarial Networks)은 목표하는 분포와 유사한 분포를 만드는 신경망을 사용하였다면, 본 발명의 실시 예에 따른 전처리부(110)에서는 입력되는 영상의 통계적 특성을 신경망을 이용한 객체 검출 성능 향상에 유리한 형태로 변화시키는 전처리 신경망을 사용함으로써, 입력되는 영상에서 특징점을 찾아내는 기존의 전처리 기법과는 달리 입력되는 영상을 신경망을 이용한 객체 검출에 최적화되도록 변조하여 출력한다.

[0027] 즉, 전처리부(110)에서 입력되는 영상의 통계적 특성을 신경망을 이용한 객체 검출 성능 향상에 유리한 형태로 변화시킨다는 것은 입력되는 영상을 사람의 육안으로 보이는 것에 적합한 형태로 변화시키는 것을 의미하는 것이 아니라, 객체 검출을 위한 신경망에 적합한 형태가 되도록 입력되는 영상을 전처리하는 것을 의미하며, 입력되는 영상과 완전히 다른 영상으로 변조하는 것을 포함할 수 있다.

[0028] 전처리부(110)는 학습 영상과 학습 영상에 대한 신경망을 이용한 객체 검출 결과를 이용하여, 신경망을 이용한 객체 검출에 최적화되도록 해당 영상의 통계적 특성이 변하도록 전처리 신경망을 학습시켜, 객체 검출을 위한 신경망에 특화된 전처리 신경망을 생성한다. 또한 전처리부(110)는 피드백 처리부(130)를 통해 피드백되는 검출부(120)의 검출 결과를 이용하여 객체 검출을 위한 신경망의 객체 검출 성능을 향상시키는 방향으로 전처리 신경망을 재학습시킬 수 있다. 전처리부(110)는 검출부(120)의 검출 결과를 토대로 전처리 신경망의 학습 방향을 결정하고 전처리 신경망의 파라미터 등을 업데이트할 수 있다. 전처리부(110)는 검출부(120)의 검출 결과의 양성 또는 음성에 따라 전처리 신경망의 학습 방향을 결정할 수 있다. 전처리부(110)는 검출부(120)의 검출된 객체에 대한 손실 값을 최소화하는 방향으로 전처리 신경망을 학습시킬 수 있다.

[0029] 이와 같이, 선박 검출을 위한 신경망에 특화된 전처리 신경망을 이용한 영상 전처리를 통해 검출부(120)의 객체 검출 성능을 향상시킬 수 있다.

[0030] 검출부(120)는 전처리부(110)로부터 전처리된 영상에 대해 신경망 기반의 선박 검출 기법을 적용하여 선박 검출을 수행하고, 선박 검출 성능을 측정한다. 검출부(120)는 Cascade R-CNN(Region-based Convolutional Neural Network)과 같은 신경망을 사용하여 선박 검출을 수행할 수 있으며, 그 외에도 다양한 기계 학습을 이용한 선박 검출 기법을 적용할 수 있다. 검출부(120)는 선박 검출 과정에서 객체 클래스(class)와 검출된 객체에 대한 손실(loss) 값을 측정할 수 있다. 객체 클래스는 식별의 대상이 되는 객체의 종류를 의미하며, 사람, 선박이나 구조물 등으로 구별될 수 있으며, 필요에 따라 좀 더 세밀하게 구별될 수도 있다. 예를 들면, 선박 같은 경우, 상선, 군함, 어선 등으로 객체 클래스가 설정될 수 있다. 또한 검출된 객체에 대한 손실값은 신경망의 출력 값과 실제 객체 클래스 값의 차이를 나타낸다. 이때 검출부(120)는 전처리부(110)의 전처리 신경망의 적용 여부에 따른 객체 클래스와 검출된 객체에 대한 손실 값을 측정할 수 있으며, 이를 통해 적용된 전처리 신경망을 이용한 영상 전처리 기법의 유효성 여부를 판별할 수 있다.

[0031] 피드백 처리부(130)는 검출부(120)의 검출 결과를 전처리부(110)로 전달한다. 또한 피드백 처리부(130)는 검출부(120)의 검출 결과를 입력 영상에 대응하여 저장부(140)에 저장할 수 있다. 이렇게 저장된 데이터는 추후 학습 데이터로 활용되어, 검출부(120)의 신경망이 업데이트될 수 있다.

[0032] 도 2는 본 발명의 실시 예에 따른 전처리 신경망의 학습 방법을 나타낸 도면이다.

[0033] 도 2를 참고하면, 전처리부(110)는 입력 영상들과 입력 영상들에 대한 실제 객체 검출 결과를 전처리 신경망의 학습 데이터로 사용한다.

[0034] 전처리부(110)는 입력 영상과 입력 영상에 대한 객체 검출 결과를 수신한다(S210).

[0035] 전처리부(110)는 입력 영상과 이 입력 영상에 대한 객체 검출 결과를 토대로 전처리 신경망의 학습 방향을 결정한다(S220). 전처리부(110)는 실제 객체 검출 결과의 양성 또는 음성에 따라 전처리 신경망의 학습 방향을 결정할 수 있다. 양성은 객체가 제대로 검출된 것을 나타내고, 음성은 그렇지 않은 것을 나타낸다.

[0036] 전처리부(110)는 결정된 학습 방향을 토대로 입력된 영상을 신경망 기반 객체 검출에 최적화된 형태로 변환시키는 전처리 신경망을 학습시킨다(S230). 예를 들면, 객체 검출 결과의 손실값이 기존보다 작아졌다면, 객체 검출 성능이 향상되는 방향으로 학습이 된 것이고, 객체 검출 결과의 손실값이 기존보다 커졌다면, 객체 검출 성능이 나빠진 방향으로 학습이 된 것이라 할 수 있다. 전처리부(110)는 입력 영상과 이 입력 영상에 대한 실제 객체 검출 결과를 토대로 객체 검출 성능을 향상시키는 방향으로 전처리 신경망을 학습시킬 수 있다. 객체 검출 성능을 향상시키는 방향으로 전처리 신경망을 학습시키는 방법으로, 전처리부(110)는 영상 내의 객체 좌표의 실제 값(Ground truth)에 해당하는 바운딩 박스와 예측된 바운딩 박스의 손실을 최소화 하는 방향, 객체의 센터 좌표의 손실을 최소화 하는 방향, 세그맨테이션 값에 대한 IOU(Intersection over Union)를 최대화 하는 방향 등으로 전처리 신경망을 학습시킬 수 있다.

[0037] 이렇게 학습된 전처리 신경망에서는 영상이 입력되면, 입력되는 영상을 변조하여, 최종적으로 검출부(120)에서 전처리된 영상을 이용한 객체 검출 성능이 향상되도록 한다.

[0038] 도 3은 본 발명의 실시 예에 따른 객체 검출 방법을 나타낸 흐름도이다.

[0039] 도 3을 참고하면, 전처리부(110)는 객체를 검출하기 위한 입력 영상을 수신하면(S310), 학습된 전처리 신경망의 영상 전처리 기법을 이용하여 입력 영상의 통계적 특성을 변화시켜 출력한다(S320).

[0040] 검출부(120)는 전처리부(110)로부터 출력된 영상으로부터 신경망 기반 객체 검출 기법을 이용하여 객체를 검출한다(S330).

[0041] 검출부(120)는 객체 검출 결과를 피드백 처리부(130)를 통해 전처리부(110)로 전달한다. 검출부(120)는 검출된 객체 클래스 및 검출된 객체에 대한 손실 값 등을 객체 검출 결과로서 피드백 처리부(130)를 통해 전처리부(110)로 전달할 수 있다.

[0042] 전처리부(110)는 입력 영상과 이 입력 영상에 대한 객체 검출 결과를 토대로 전처리 신경망을 학습시킬 수 있다(S340). 즉, 전처리부(110)는 입력 영상과 이 입력 영상에 대한 객체 검출 결과를 토대로 신경망 기반 객체 검출 성능이 향상되도록, 전처리 신경망을 업데이트시킬 수 있다.

[0043] 도 4는 본 발명의 다른 실시 예에 따른 객체 검출 장치를 나타낸 도면으로, 도 1 내지 도 3을 참고하여 설명한 객체 검출 장치 및 방법 중 적어도 일부를 수행하는 데 사용할 수 있는 시스템을 나타낸다.

[0044] 도 4를 참고하면, 객체 검출 장치(400)는 프로세서(410), 메모리(420), 저장 장치(430) 및 입출력

(input/output, I/O) 인터페이스(440)를 포함한다.

[0045] 프로세서(410)는 중앙 처리 유닛(central processing unit, CPU)이나 기타 칩셋, 마이크로프로세서 등으로 구현될 수 있다.

[0046] 메모리(420)는 동적 랜덤 액세스 메모리(dynamic random access memory, DRAM), 램버스 DRAM(rambus DRAM, RDRAM), 동기식 DRAM(synchronous DRAM, SDRAM), 정적 RAM(static RAM, SRAM) 등의 RAM과 같은 매체로 구현될 수 있다.

[0047] 저장 장치(430)는 하드 디스크(hard disk), CD-ROM(compact disk read only memory), CD-RW(CD rewritable), DVD-ROM(digital video disk ROM), DVD-RAM, DVD-RW 디스크, 블루레이(blu-ray) 디스크 등의 광학 디스크, 플래시 메모리, 다양한 형태의 RAM과 같은 영구 또는 휘발성 저장 장치로 구현될 수 있다.

[0048] I/O 인터페이스(440)는 프로세서(410) 및/또는 메모리(420)가 저장 장치(430)에 접근할 수 있도록 한다. 또한 I/O 인터페이스(440)는 외부 예를 들면, 사용자와의 인터페이스를 제공할 수 있다.

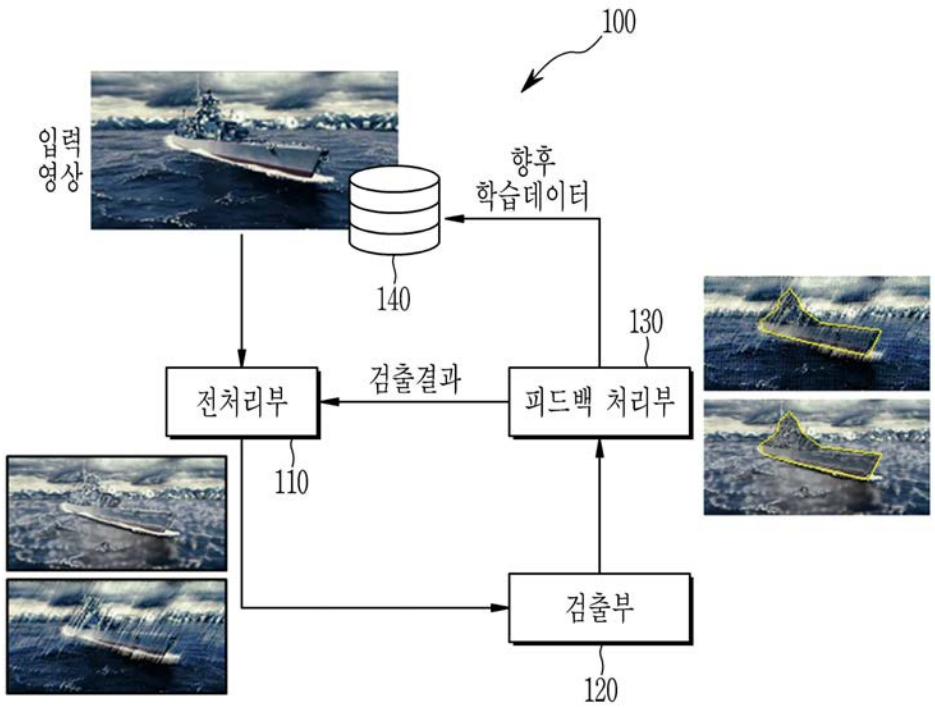
[0049] 메모리(420) 또는 저장 장치(430)는 저장부(140)를 포함할 수 있다.

[0050] 프로세서(410)는 도 1 내지 도 3에서 설명한 객체 검출 기능을 수행할 수 있으며, 전처리부(110), 검출부(120) 및 피드백 처리부(130) 중 적어도 일부의 기능을 구현하기 위한 프로그램 명령을 메모리(420)에 로드시켜, 도 1 내지 도 3을 참고로 하여 설명한 동작이 수행되도록 제어할 수 있다. 그리고 이러한 프로그램 명령은 저장 장치(430)에 저장되어 있을 수 있으며, 또는 네트워크로 연결되어 있는 다른 시스템에 저장되어 있을 수 있다.

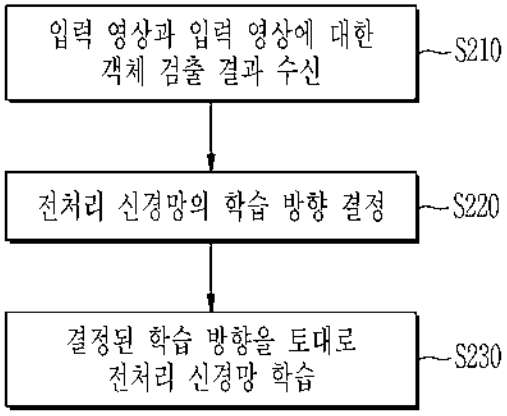
[0051] 이상에서 본 발명의 실시 예에 대하여 상세하게 설명하였지만 본 발명의 권리 범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리 범위에 속하는 것이다.

도면

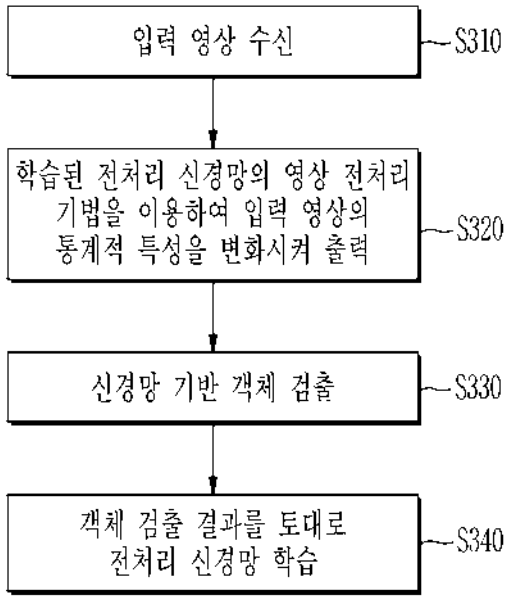
도면1



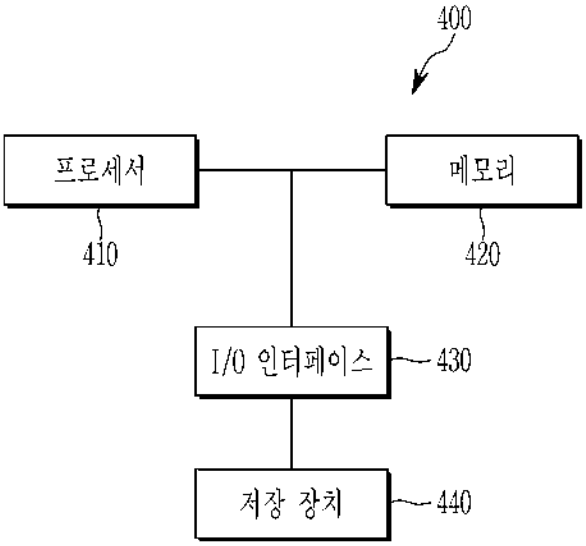
도면2


입력 영상과 입력 영상에 대한 객체 검출 결과 수신
S210
전처리 신경망의 학습 방향 결정
S220
결정된 학습 방향을 토대로 전처리 신경망 학습
S230

도면3


입력 영상 수신
S310
학습된 전처리 신경망의 영상 전처리 기법을 이용하여 입력 영상의 통계적 특성을 변화시켜 출력
S320
신경망 기반 객체 검출
S330
객체 검출 결과를 토대로 전처리 신경망 학습
S340

도면4



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0028941
(43) 공개일자 2021년03월15일

(51) 국제특허분류(Int. Cl.)
G06K 9/00 (2006.01) *G06K 9/62* (2006.01)
(52) CPC특허분류
G06K 9/00624 (2013.01)
G06K 9/00201 (2013.01)
(21) 출원번호 10-2019-0110110
(22) 출원일자 2019년09월05일
심사청구일자 없음

(71) 출원인
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
이지원
대전광역시 유성구 반석서로 98, 602동 2203호
남도원
대전광역시 유성구 대덕대로 598, 301호
(뒷면에 계속)
(74) 대리인
특허법인이상

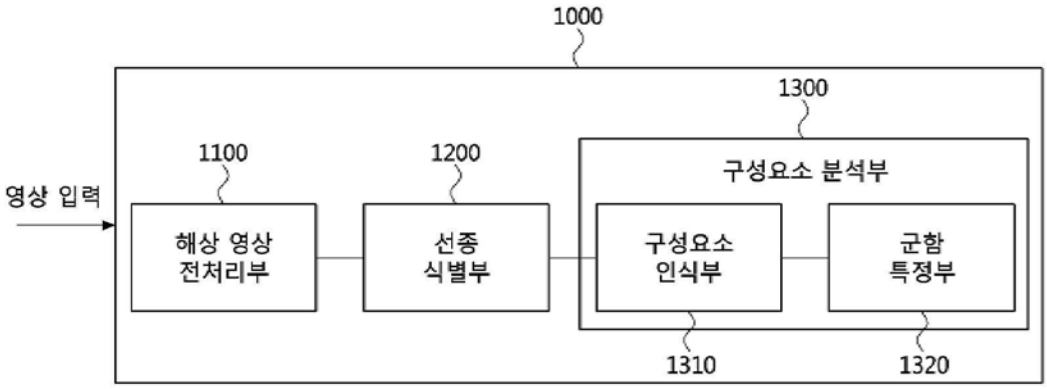
전체 청구항 수 : 총 1 항

(54) 발명의 명칭 **영상에 기반하여 선박을 식별하는 방법 및 장치**

(57) 요 약

본 발명은 입력된 영상에서 객체를 식별하는 데 방해가 되는 요소를 제거하는 단계; 선박의 종류를 분류하도록 학습된 제 1 딥 러닝 모델에 기반하여 상기 영상에 포함된 선박의 종류를 식별하는 단계; 상기 식별된 선박이 군함인 경우 영상 분할 기법을 활용하여 군함을 구성요소 단위로 분류하도록 학습된 제 2 딥 러닝 모델을 기초로 상기 군함을 구성요소 단위로 분할하여 식별하는 단계; 상기 군함의 구성요소의 종류와 위치에 기반한 3차원 패턴 매칭을 이용하여 상기 군함의 종류 및 함급(Ship-Class)을 식별하는 단계; 및 상기 식별된 군함의 종류 및 함급을 상기 3차원 패턴 매칭 결과에 따른 유사도가 높은 순서대로 사용자에게 제시하는 단계를 포함하는, 영상 기반 선박 식별 방법을 개시한다.

대 표 도 - 도1



(52) CPC특허분류
G06K 9/627 (2013.01)
G06N 20/00 (2019.01)
G06T 7/11 (2017.01)

(72) 발명자
문성원
대전광역시 유성구 노은서로210번길 32, 405동 1005호
유원영
대전광역시 유성구 어은로 57, 138동 1301호
윤기송
대전광역시 유성구 용산2로 30, 106동 303호
이정수
세종특별자치시 달빛1로 206, 905동 2001호

이 발명을 지원한 국가연구개발사업

과제고유번호	2019-0-00524
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	가상증강현실콘텐츠원천사업
연구과제명	영상 내 객체간 관계 분석 기반 해상 선박/구조물 상세 식별 콘텐츠 기술 개발
기 여 율	1/1
과제수행기관명	한국전자통신연구원
연구기간	2019.04.01 ~ 2019.12.31

명 세 서

청구범위

청구항 1

입력된 영상에서 객체를 식별하는 데 방해가 되는 요소를 제거하는 단계;

선박의 종류를 분류하도록 학습된 제 1 딥 러닝 모델에 기반하여 상기 영상에 포함된 선박의 종류를 식별하는 단계;

상기 식별된 선박이 군함인 경우 영상 분할 기법을 활용하여 군함을 구성요소 단위로 분류하도록 학습된 제 2 딥 러닝 모델을 기초로 상기 군함을 구성요소 단위로 분할하여 식별하는 단계;

상기 군함의 구성요소의 종류와 위치에 기반한 3차원 패턴 매칭을 이용하여 상기 군함의 종류 및 함급(Ship-Class)을 식별하는 단계; 및

상기 식별된 군함의 종류 및 함급을 상기 3차원 패턴 매칭 결과에 따른 유사도가 높은 순서대로 사용자에게 제시하는 단계를 포함하는, 영상 기반 선박 식별 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 선박을 식별하는 방법 및 장치에 관한 것으로, 더욱 상세하게는 광학 해상 영상으로부터 선박을 선종 수준까지 식별하고, 식별된 선박이 군함인 경우 구체적으로 선박을 식별하는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 운행 중인 선박이나 해안 초소, 관제 센터 등에서는 교통을 통제하거나 해양 사고를 지원하는 등 여러가지 목적을 위해 주변 선박의 정보, 위치 및 속도 등을 확인하기 위한 방법이 필요하다. 이를 위해 무선 통신을 통하여 해당 선박의 정보, 위치 및 속도 등을 송수신할 수 있는 선박 자동 식별 시스템(Automatic identification system; AIS)이나 브이패스(V-Pass)와 같은 시스템이 사용되고 있으나, 통신 기기의 부재 혹은 고장, 전시 상황에서 상대방에게 발각되지 않도록 기기를 의도적으로 끄는 경우 등과 같이 무선 통신만으로는 주변 선박에 대한 정보를 충분히 획득하지 못하는 문제점이 존재한다. 따라서, 이를 보완하기 위해 선상에서 촬영된 영상에 기반하여 주변 선박을 인식하는 선행기술이 존재한다.

[0003] 예를 들어, 선박을 외형 수준에 따라 6종으로 분류하고 딥 러닝을 통해 입력된 영상에서 선박의 종류를 인식하는 기술이 있다. 또한, 기존 선행기술은 주간 광학영상 및 야간 적외선 영상을 토대로 컨볼루션 신경망(Convolutional Neural Network; CNN)을 학습시켜 임의의 선박 영상에 대하여 87.4% 수준의 정확도를 갖는 분류기 및 다수의 특징점 조합을 통해 영상에서 특징점을 추출한 것에 기반하여 선박을 분류하고 85% 수준의 정확도를 갖는 분류기를 제안한 바 있다.

[0004] 또한, 선박 자체를 인식하고 해당 선박과 자선과의 거리, 속도 및 방향 등 위험 상황 여부를 추론하여 영상을 통해 선박 간에 충돌을 방지하는 등과 같은 위험 상황 회피에 목적이 있는 선행기술도 존재한다.

[0005] 일반적인 경우에는 선종을 분류하는 것만으로도 충분하나, 군사적인 목적과 같은 특별한 경우에는 선급 분류도 필요하다. 예를 들어, 군함이라는 선종을 식별한다고 하더라도 군함 내에는 구축함, 순양함, 초계함 등 목적에 따라 세부적으로 함종이 나누어지고 구축함 내에도 광개토대왕급, 이순신급, 세종대왕급 구축함으로 나눌 수 있으며, 나누어진 급수에 따라 다른 무기 체계를 갖추고 있기 때문에 군사 전략에 따라 각 함급에 따른 대응이 달라지므로 급수에 대한 식별도 필요하다. 다만, 선행기술들은 영상에 기반하여 인식된 선박이 어떤 종류인지를 구분할 수 있으나, 선종 내에 급수에 따라 선박을 세부 식별하지 못하는 문제가 있다.

발명의 내용

해결하려는 과제

[0007] 상기와 같은 문제점을 해결하기 위한 본 발명의 목적은, 측면에서 촬영된 영상에 기반하여 선박의 선종을 인식하고, 인식된 선종이 군함인 경우 해당 군함의 구조 분석을 통해 군함의 종류 및 함급(Ship-Class)을 식별하는 방법을 제공하는 데 있다.

과제의 해결 수단

[0008] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 영상 기반 선박 식별 방법은, 입력된 영상에서 객체를 식별하는 데 방해가 되는 요소를 제거하는 단계; 선박의 종류를 분류하도록 학습된 제 1 딥 러닝 모델에 기반하여 상기 영상에 포함된 선박의 종류를 식별하는 단계; 상기 식별된 선박이 군함인 경우 영상 분할 기법을 활용하여 군함을 구성요소 단위로 분류하도록 학습된 제 2 딥 러닝 모델을 기초로 상기 군함을 구성요소 단위로 분할하여 식별하는 단계; 상기 군함의 구성요소의 종류와 위치에 기반한 3차원 패턴 매칭을 이용하여 상기 군함의 종류 및 함급(Ship-Class)을 식별하는 단계; 및 상기 식별된 군함의 종류 및 함급을 상기 3차원 패턴 매칭 결과에 따른 유사도가 높은 순서대로 사용자에게 제시하는 단계를 포함할 수 있다.

발명의 효과

[0009] 본 발명의 일 실시예에 따르면, 측면 해상 영상에서 딥 러닝 및 패턴 매칭 기법을 단계적으로 활용하여 선박의 선종을 인식하고, 해당 선종이 군함인 경우 군함의 종류 및 함급(Ship-Class)까지 식별할 수 있는 장점을 가진다.

[0010] 본 발명의 일 실시예에 따르면, 유사한 해상 선박을 딥 러닝 및 패턴 매칭 기법을 혼합하여 단계적인 인식을 통해 최종적으로 선박을 식별하기 때문에 유사한 객체를 딥 러닝을 통해 단계적으로 식별할 때 발생할 수 있는 오류 누적 문제를 해결할 수 있고, 신뢰성 있는 식별 정보는 국방 분야의 전술적 응용에 사용될 수 있는 장점을 가진다.

도면의 간단한 설명

[0011] 도1은 본 발명의 일 실시예에 따른 영상에 기반하여 선박을 식별하는 장치의 블록 구성도이다.

도2는 본 발명의 다른 실시예에 따른 영상에 기반하여 선박을 식별하는 장치의 블록 구성도이다.

도3은 본 발명의 일 실시예에 따른 영상에 기반하여 선박을 식별하는 방법을 설명하기 위한 개념도이다.

도4는 본 발명의 일 실시예에 따른 영상에 기반하여 선박을 식별하면서 발생하는 오류를 최소화하는 방법을 설명하기 위한 개념도이다.

도5는 해상 영상 전처리부에서 선박을 식별하는 데 방해되는 여러 요소를 나타낸 도면이다.

도6은 선종 식별부에서 딥 러닝에 기반하여 선박의 종류를 식별하는 방법을 설명하기 위한 개념도이다.

도7은 구성요소 분석부에서 군함의 종류 및 함급을 식별하는 방법을 설명하기 위한 개념도이다.

도8은 본 발명의 일 실시예에 따른 영상에 기반하여 선박을 식별하는 방법의 동작 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0012] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였다.

[0013] 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는 데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. "및/또는"이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[0014] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에

직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0015] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0016] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0018] 이하, 본 발명에 따른 바람직한 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0019] 도1은 본 발명의 일 실시예에 따른 영상에 기반하여 선박을 식별하는 장치의 블록 구성도이다.

[0020] 도 1을 참조하여 살펴보면, 본 발명의 일 실시예에 따른 영상에 기반하여 선박을 식별하는 장치(1000)는 해상 영상 전처리부(1100), 선종 식별부(1200), 구성요소 인식부(1310) 및 군함 특정부(1320)를 포함하는 구성요소 분석부(1300)로 구성될 수 있다.

[0021] 여기서, 해상 영상 전처리부(1100)는 영상에서 해상 객체를 식별할 때 방해가 되는 수면 반사광, 해무 등과 같이 해상 상황에서 발생할 수 있는 방해 요소를 제거할 수 있다.

[0022] 또한, 선종 식별부(1200)는 해상 영상 전처리부에 의해 전처리된 영상으로부터 해상 객체를 식별하여 선박의 종류를 인식할 수 있다. 여기서, 선박의 종류를 분류하도록 학습된 딥 러닝 모델에 기반하여 선박의 종류를 식별할 수 있다. 또한, 선박의 종류는 어선, 컨테이너선, 군함, 여객선 및 요트를 포함할 수 있다.

[0023] 또한, 구성요소 분석부(1300)는 군함을 구성요소 단위로 분해 인식하는 구성요소 인식부(1310)와 구성요소 간 구조 및 관계를 토대로 최종적으로 입력된 영상으로부터 군함의 종류 및 함급을 도출하는 군함 특정부(1320)를 포함할 수 있다.

[0024] 여기서, 구성요소 인식부(1310)는 선종 식별부에 의해 식별된 선박의 종류가 군함인 경우 군함의 구성요소를 세부적으로 분할하여 인식할 수 있다. 특히, 영상 분할 기법을 활용하여 군함을 구성요소 단위로 분류하도록 학습된 딥 러닝 모델에 기반하여 군함을 구성요소 단위로 분할하여 식별할 수 있다. 여기서, 군함의 구성요소는 몸체, 포, 레이더, 하우스를 포함할 수 있다.

[0025] 또한, 군함 특정부(1320)는 구성요소 인식부에 의해 인식된 군함의 구성요소를 구성요소 간 관계나 구조를 분석하여 군함의 종류 및 함급을 특정할 수 있다. 여기서, 군함의 구성요소의 종류와 위치에 기반한 3차원 패턴 매칭을 이용하여 군함의 종류 및 함급을 식별할 수 있다. 또한, 군함의 종류는 수송함, 구축함, 호위함, 초계함, 상륙함 및 고속함을 포함할 수 있고, 함급은 세종대왕급, 충무공 이순신급, 광개토 대왕급 및 KDDX를 포함할 수 있다. 또한, 군함 특정부에서 특정된 군함을 3차원 패턴 매칭 결과에 따른 유사도가 높은 순서대로 사용자에게 제시할 수 있다.

[0027] 도2는 본 발명의 다른 실시예에 따른 영상에 기반하여 선박을 식별하는 장치의 블록 구성도이다.

[0028] 도2를 참조하면, 본 발명의 일 실시예에 따른 영상에 기반하여 선박을 식별하는 장치(1000)는 프로세서(1010) 및 프로세서를 통해 실행되는 적어도 하나의 명령 및 명령 수행의 결과를 저장하는 메모리(1020) 및 네트워크와 연결되어 통신을 수행하는 송수신 장치(1030)를 포함할 수 있다.

[0029] 선박 식별 장치(1000)는 또한, 입력 인터페이스 장치(1040), 출력 인터페이스 장치(1050), 저장 장치(1060) 등을 더 포함할 수 있다. 선박 식별 장치(1000)에 포함된 각각의 구성 요소들은 버스(Bus)(1070)에 의해 연결되어

서로 통신을 수행할 수 있다.

[0030] 프로세서(1010)는 메모리(1020) 및 저장 장치(1060) 중에서 적어도 하나에 저장된 프로그램 명령(program command)을 실행할 수 있다. 프로세서(1010)는 중앙 처리 장치(central processing unit, CPU), 그래픽 처리 장치(graphics processing unit, GPU), 또는 본 발명의 실시예들에 따른 방법들이 수행되는 전용의 프로세서를 의미할 수 있다. 메모리(1020) 및 저장 장치(1060) 각각은 휘발성 저장 매체 및 비휘발성 저장 매체 중에서 적어도 하나로 구성될 수 있다. 예를 들어, 메모리(1020)는 읽기 전용 메모리(read only memory, ROM) 및 랜덤 액세스 메모리(random access memory, RAM) 중에서 적어도 하나로 구성될 수 있다.

[0031] 저장 장치(1060)는 또한, 선박의 종류를 식별하기 위해 외형의 유사성에 따라 선박을 분류한 데이터에 기반하여 딥 러닝을 수행한 결과 및 군함의 구성요소를 식별하기 위해 군함을 구성요소에 따라 분류한 데이터에 기반하여 딥 러닝을 수행한 결과를 저장할 수 있다.

[0032] 여기서, 적어도 하나의 명령은 입력된 영상에서 객체를 식별하는 데 방해가 되는 요소를 제거하도록 하는 명령; 선박의 종류를 분류하도록 학습된 제 1 딥 러닝 모델에 기반하여 상기 영상에 포함된 선박의 종류를 식별하도록 하는 명령; 상기 식별된 선박이 군함인 경우 영상 분할 기법을 활용하여 군함을 구성요소 단위로 분류하도록 학습된 제 2 딥 러닝 모델을 기초로 상기 군함을 구성요소 단위로 분할하여 식별하도록 하는 명령; 상기 군함의 구성요소의 종류와 위치에 기반한 3차원 패턴 매칭을 이용하여 상기 군함의 종류 및 함급(Ship-Class)을 식별하도록 하는 명령; 상기 식별된 군함의 종류 및 함급을 상기 3차원 패턴 매칭 결과에 따른 유사도가 높은 순서대로 사용자에게 제시하도록 하는 명령을 포함할 수 있다.

[0034] 도3은 본 발명의 일 실시예에 따른 영상에 기반하여 선박을 식별하는 방법을 설명하기 위한 개념도이다.

[0035] 도3을 참조하여 살펴보면, 다양한 고도에서 촬영된 영상에 기반하여 해상구조물, 일반 선박 및 군함에 따라 객체를 식별하는 목표의 수준이 다름을 알 수 있다. 예를 들어, 일반적인 선박은 어선, 컨테이너선, 군함, 여객선 및 요트 등 선박의 종류까지만 식별하더라도 여러 응용에 활용이 가능하지만, 군함의 경우 그 용도에 따라 수송함, 구축함, 호위함, 초계함, 상륙함, 및 고속함 등 여러 종류로 나누어질 수 있고, 하나의 군함 안에서도 그 크기나 배수량에 따라 세종대왕급, 충무공 이순신급, 광토대왕급 및 한국형 차기 구축함(KDDX) 등 여러 함급으로 나누어질 수 있다.

[0036] 또한, 함급에 따라 서로 다른 무기 체계를 갖추고 있기 때문에 이에 대한 적절한 대응을 위해 군함의 종류 및 함급까지 식별을 할 필요가 있다. 따라서, 본 발명에 따르면 다양한 고도에서 촬영된 영상을 통해 선박으로 1차적으로 식별하고, 식별된 선박이 군함인 경우 구성 요소를 세부적으로 인식하고, 구조를 분석하여 해당 군함의 종류 및 함급까지 식별할 수 있다.

[0038] 도4는 본 발명의 일 실시예에 따른 영상에 기반하여 선박을 식별하는 방법의 오류를 최소화하기 위한 방법을 설명하기 위한 개념도이다.

[0039] 도 4를 참조하여, 본 발명에서 3차원 패턴 매칭을 통하여 최종적으로 식별된 군함의 후보를 사용자에게 제시하는 이유를 살펴보면, 먼저, 4-단계에 걸쳐서 딥 러닝을 통하여 선박의 종류, 군함의 종류 및 함급을 식별할 경우 정확도의 오류가 단계적으로 누적되어 최종적으로 정확도가 크게 하락될 수 있는 바, 본 발명에 따르면 오류 누적 문제를 최소화하기 위해 2-단계에 걸쳐서 딥 러닝을 통하여 선박을 식별할 수 있다. 특히, 2-단계의 딥 러닝 모델 중 하나의 모델은 오류가 일정 수준 발생하여 선박을 인식하는데 큰 영향을 주지 않는 선박의 구성요소를 식별하는 모델일 수 있다.

[0040] 또한, 딥 러닝 모델을 통하여 군함의 종류 및 함급을 식별하는 경우 현재 기술로는 딥 러닝 네트워크가 영상에서 객체의 어떤 특징을 인식하여 유사한 객체로 분류하는지 정확한 분석이 이루어지지 않았고, 3차원 패턴 매칭을 통해 유사도를 측정하는 방식이 일반적으로 사람이 비슷한 객체를 구분하는 방식과 유사하고, 군함의 종류 및 함급의 식별 결과는 군사적이고 전술적인 목적으로 활용되는 바 사용자에게 N개의 군함 후보를 제시하여 숙련된 전문가에게 판단하게 하는 것이 안전하므로 본 발명에 따르면 최종적으로 식별된 군함의 후보를 사용자에게 제시함으로써 사용자가 직접 군함을 결정하도록 할 수 있다.

[0042] 도5는 해상 영상 전처리부에서 선박을 식별하는 데 방해되는 여러 요소를 나타낸 도면이다.

[0043] 도5를 참조하면, 해상 영상 전처리부(1100)는 일반적인 영상 처리 기법을 이용하여 영상에서 해상 객체를 식별할 때 방해가 되는 해상 요소를 제거하거나, 식별 대상에서 제외하기 위해 마킹하는 과정을 거칠 수 있다. 제거 대상이 되는 해상 요소는 수면 반사광, 해무 및 우적(비)이 있을 수 있고, 식별 대상에서 제외하기 위해 마킹하는 대상이 되는 것은 후류(선박이 움직일 때 뒤에 나타나는 선미파) 및 육지(섬) 등을 미리 식별하여 식별 대상에서 제외시킴으로써 이후 선박의 종류를 판단함에 있어 오인식률을 크게 감소시킬 수 있다.

[0044] 여기서, 영상 처리 기법은 컴퓨터를 이용하여 영상을 생성하고 처리하고 영상을 해석, 인식하는 영상과 관련된 모든 분야를 의미할 수 있다. 따라서, 사용자는 흐린 영상을 보다 선명하게 볼 수 있거나 영상이 훼손된 경우 다시 원래 영상으로 복원하는 등 영상에서 필요한 정보만을 추출하여 얻을 수 있다.

[0045] 또한, 영상 처리 기법에 따라 영상 획득시 주위 환경의 영향으로 영상이 흐리거나 너무 어두울 경우 혹은 잡영이 많이 석인 경우 사용자가 원하는 영상을 얻기 위해 영상을 조작할 수 있다.

[0046] 또한, 영상 처리 기법에 따라 영상 조작에 의해 보정된 영상에서 특징을 찾아낼 수 있고, 사람의 눈으로 식별이 불가능한 미세한 영상물의 차이점을 발견하고 다른 영상과 비교분석하며 영상의 특징을 찾아 영상을 인식할 수 있다.

[0048] 도6은 선종 식별부에서 딥 러닝에 기반하여 선박의 종류를 식별하는 방법을 설명하기 위한 개념도이다.

[0049] 선종 식별부(1200)는 해상 영상 전처리부(1100)에서 전처리된 영상을 입력으로 하여 영상에서 선박의 종류를 인식할 수 있다. 따라서, 선박의 종류를 식별하기 위해서는 선종을 외형적인 구조나 특성에 기반하여 분류하는 과정이 선행되어야 한다. 예를 들어, 한국의 선박 안전법의 선종 분류 체계를 1차적으로 따르면서, 외형적으로 특수성이 있거나 유사성이 높은 선박을 분류하여 데이터를 수집하여 딥 러닝을 수행한 모델을 통해 선박의 종류를 식별할 수 있다. 이 때, 영상에 기반하여 객체를 식별하도록 설계된 AlexNet, VGGNet, ResNet 등을 기초로 전이 학습(Transfer Learning)을 수행한 모델이 선종 식별부(1200)의 딥 러닝 모델로 사용될 수 있다.

[0050] 여기서, 선종 식별부(1200)에서 인식된 선박의 종류가 군함이 아닌 경우 그 결과를 사용자에게 제시하고 선박 식별을 종료할 수 있다. 다만, 인식된 선박이 군함인 경우 계속해서 군함의 구성요소에 따라 구체적으로 선박을 식별할 수 있다.

[0052] 도7은 구성요소 분석부에서 군함의 종류 및 함급을 식별하는 방법을 설명하기 위한 개념도이다.

[0053] 도 7을 참조하면, 구성요소 인식부(1310)는 영상 분할(Segmentation) 기법을 통하여 군함을 구성요소 단위로 분할할 수 있다. 여기서, 군함을 구성요소 단위로 분할하기 위해 PSPNet, DeepLab 등의 딥 러닝 기술이 활용될 수 있으며, 군함을 군함의 주요 구성요소인 몸체, 포, 레이더, 하우스 등으로 구분하여 식별할 수 있다. 또한, 원거리 또는 소형의 군함도 인식할 수 있다. 예를 들어, 도7을 참조하여 본 발명의 원리를 설명하면, 구성요소 인식부를 통해 해상 객체를 양망기(Power Block), 레이더(Radar), 헬리콥터(Helicopter), 소형보트(Net Skiff), 퍼스 윈치(Purse Winch) 및 스피드 보트(Speed)로 분류하여 식별할 수 있다.

[0054] 또한, 군함 특정부(1320)는 구성요소 인식부(1310)으로부터 인식된 군함의 구성요소들의 종류와 위치를 토대로 군함 구조 데이터베이스에서 3차원 패턴 매칭을 통하여 가장 유사도가 높은 군함의 종류와 함급을 식별할 수 있다. 예를 들어, 도7을 참조하여 본 발명의 원리를 설명하면, 군함 특정부는 상기 양망기, 레이더, 헬리콥터, 소형보트, 퍼스 윈치 및 스피드 보트 간 구조 및 관계를 분석하여 가장 유사한 해상 객체를 식별할 수 있다. 여기서, 3차원 패턴 매칭의 결과에 따라 가장 유사도가 높은 군함을 유사도 순으로 N개를 사용자에게 제시할 수 있다.

[0056] 도8은 본 발명의 일 실시예에 따른 영상에 기반하여 선박을 식별하는 방법의 동작 순서도이다.

[0057] 본 발명의 일 실시예에 따른 영상에 기반하여 선박을 식별하는 방법은 먼저, 입력된 영상에서 영상 처리 기법을 이용하여 해상 객체를 식별하는데 방해가 되는 요소를 제거하는 단계(S810)를 포함할 수 있다. 여기서, 영상 처리 기법은 영상에서 해상 객체를 식별할 때 방해가 되는 해상 요소를 제거하거나, 식별 대상에서 제외하기 위해

마킹하는 과정을 거칠 수 있다. 제거 대상이 되는 해상 요소는 수면 반사광, 해무 및 우적(비)이 있을 수 있고, 식별 대상에서 제외하기 위해 마킹하는 대상이 되는 것은 후류(선박이 움직일 때 뒤에 나타나는 선미파) 및 육지(섬) 등을 미리 식별하여 식별 대상에서 제외시킴으로써 이후 선박의 종류를 판단함에 있어 오인식률을 크게 감소시킬 수 있다.

[0058] 이어서, 본 발명의 일 실시예에 따른 선박 식별 방법은, 선박의 종류를 분류하도록 학습된 제 1 딥 러닝 모델에 기반하여 상기 영상에 포함된 선박의 종류를 식별하는 단계(S820)를 포함할 수 있다. 여기서, 선박의 종류를 식별하기 위해 한국의 선박 안전법의 선종 분류 체계를 1차적으로 따르면서, 외형적으로 특수성이 있거나 유사성이 높은 선박을 분류하여 데이터를 수집하여 딥 러닝을 수행한 모델을 통해 선박의 종류를 식별할 수 있다.

[0059] 이어서, 본 발명의 일 실시예에 따른 선박 식별 방법은, 상기 식별된 선박이 군함인 경우 영상 분할 기법을 활용하여 군함을 구성요소 단위로 분류하도록 학습된 제 2 딥 러닝 모델을 기초로 상기 군함을 구성요소 단위로 분할하여 식별하는 단계(S830)를 포함할 수 있다. 여기서, 군함의 구성요소에 따라 학습된 딥 러닝 모델을 통해 군함을 구성요소 단위로 분할하여 식별할 수 있다.

[0060] 이어서, 본 발명의 일 실시예에 따른 선박 식별 방법은, 상기 군함의 구성요소의 종류와 위치에 기반한 3차원 패턴 매칭을 이용하여 상기 군함의 종류 및 함급(Ship-Class)을 식별하는 단계(S840)를 포함할 수 있다.

[0061] 이어서, 본 발명의 일 실시예에 따른 선박 식별 방법은, 상기 식별된 군함의 종류 및 함급을 상기 3차원 패턴 매칭 결과에 따른 유사도가 높은 순서대로 사용자에게 제시하는 단계를 포함(S850)할 수 있다.

[0063] 본 발명의 실시예에 따른 방법의 동작은 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 프로그램 또는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의해 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치를 포함한다. 또한 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어 분산 방식으로 컴퓨터로 읽을 수 있는 프로그램 또는 코드가 저장되고 실행될 수 있다.

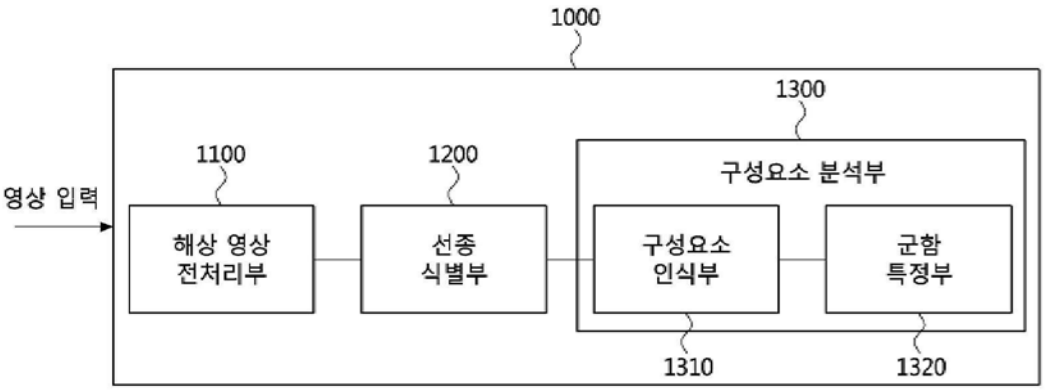
[0064] 또한, 컴퓨터가 읽을 수 있는 기록매체는 롬(rom), 램(ram), 플래시 메모리(flash memory) 등과 같이 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치를 포함할 수 있다. 프로그램 명령은 컴파일러(compiler)에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터(interpreter) 등을 사용해서 컴퓨터에 의해 실행될 수 있는 고급 언어 코드를 포함할 수 있다.

[0065] 본 발명의 일부 측면들은 장치의 문맥에서 설명되었으나, 그것은 상응하는 방법에 따른 설명 또한 나타낼 수 있고, 여기서 블록 또는 장치는 방법 단계 또는 방법 단계의 특징에 상응한다. 유사하게, 방법의 문맥에서 설명된 측면들은 또한 상응하는 블록 또는 아이템 또는 상응하는 장치의 특징으로 나타낼 수 있다. 방법 단계들의 몇몇 또는 전부는 예를 들어, 마이크로프로세서, 프로그램 가능한 컴퓨터 또는 전자 회로와 같은 하드웨어 장치에 의해(또는 이용하여) 수행될 수 있다. 몇몇의 실시예에서, 가장 중요한 방법 단계들의 하나 이상은 이와 같은 장치에 의해 수행될 수 있다.

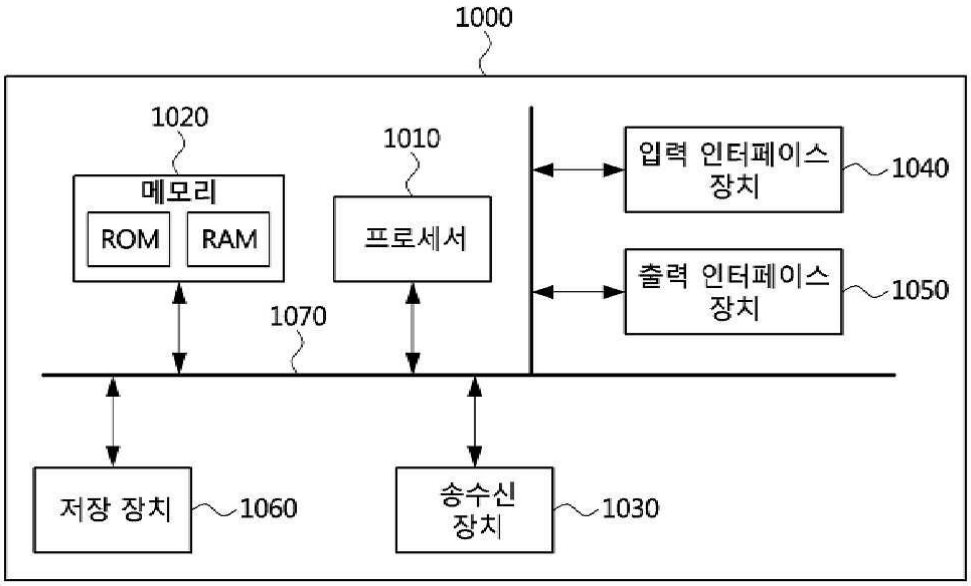
[0066] 이상 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

도면

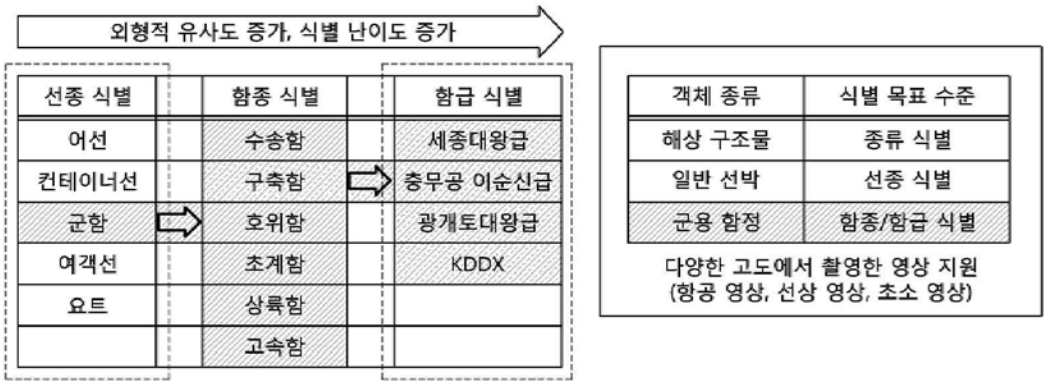
도면1



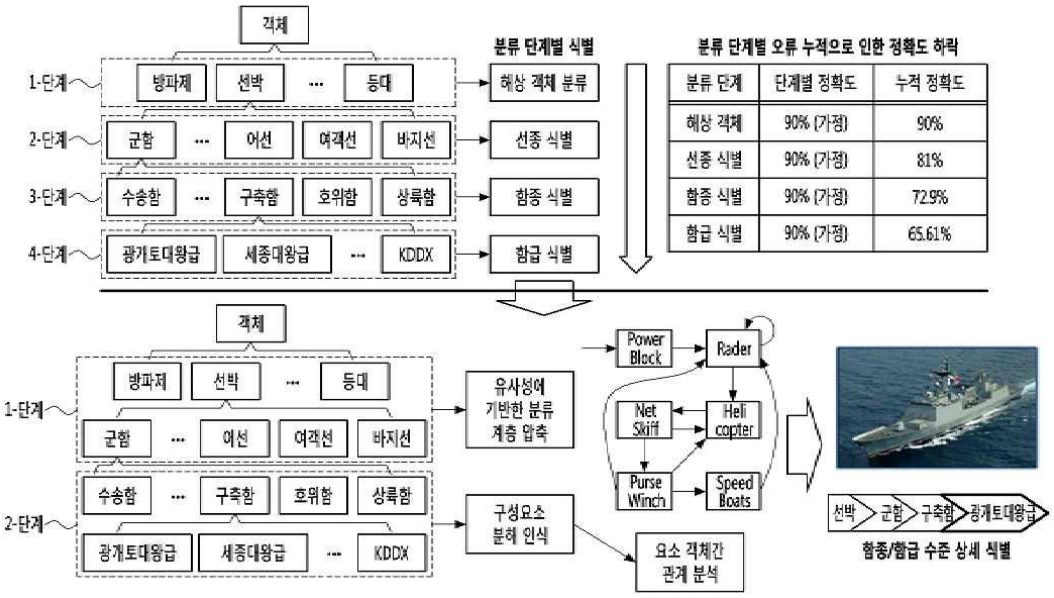
도면2



도면3



도면4

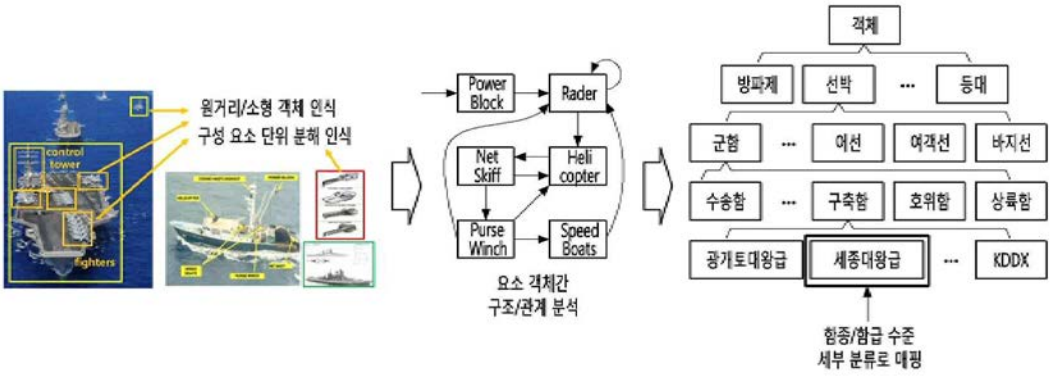


도면5

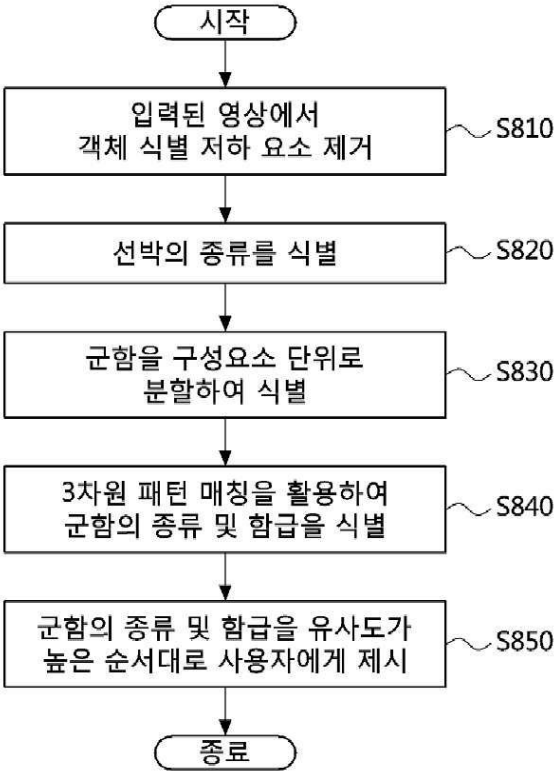
도면6

1차 분류	2차 분류	식별 ID
어선		V_01_01
여객선	크루즈	V_02_01
	페리	V_02_02
화물선	컨테이너선	V_03_01
	LNG선	V_03_02
	기타 화물선	V_03_03
예인선		V_04_01
부선		V_05_01
해경선		V_06_01
군함		V_07_01
기타		V_08_01

도면7


원거리/소형 객체 인식
구성 요소 단위 분해 인식
control tower
fighters
Power Block
Rader
Net Skiff
Heli copter
Purse Winch
Speed Boats
요소 객체간
구조/관계 분석
객체
방파제
선박
...
등대
군함
...
여선
여객선
바지선
수송함
...
구축함
호위함
상륙함
광개토대왕급
세종대왕급
...
KDDX
함종/함급 수준
세부 분류로 대핑

도면8


시작
입력된 영상에서
객체 식별 저하 요소 제거
S810
선박의 종류를 식별
S820
군함을 구성요소 단위로
분할하여 식별
S830
3차원 패턴 매칭을 활용하여
군함의 종류 및 함급을 식별
S840
군함의 종류 및 함급을 유사도가
높은 순서대로 사용자에게 제시
S850
종료

태블릿 환경에서 구동이 가능한 GAN 기반의 고화질 얼굴 편집 기술

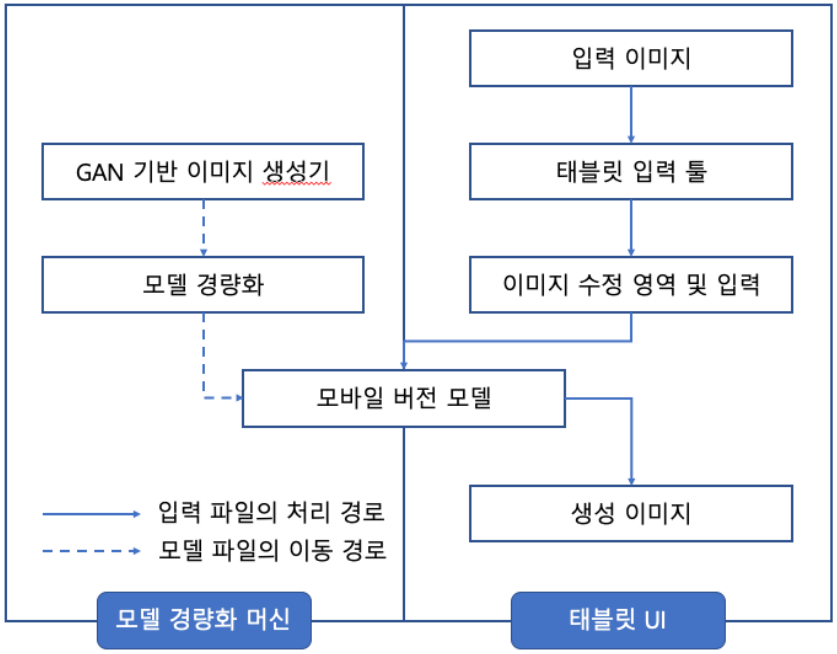
목 차

1. 기술의 개요
2. 기술이전 내용 및 범위
3. 경쟁기술과 비교
4. 기술의 사업성
 - 활용분야 및 기대효과
5. 국내외 시장 동향

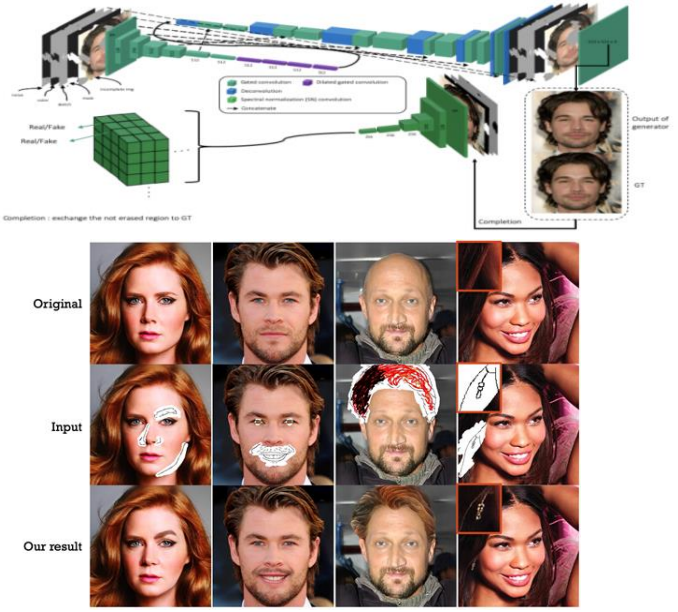
1. 기술의 개요

▣ 태블릿 환경에서 사용자의 스케치 입력에 따라 얼굴을 편집하고 고화질의 얼굴 이미지를 생성 복원하는 기술

본 기술이전은 얼굴 편집 및 복원 기술로서, 고화질의 얼굴 이미지 편집과 복원을 위한 GAN 모델 생성 기술, 태블릿 환경에서 동작이 가능하도록 모델을 변환하는 기술, 이미지 편집 및 결과 출력을 위한 UI를 포함하여 구성됨



< 본 기술의 전체적인 구조도 >



< 본 기술에 활용된 딥러닝 네트워크 구조 및 실행 결과 >

※ 본 기술 우수성은 SC-FEGAN 논문으로 ICCV 2019에서 발표

2. 기술이전 내용 및 범위

▣ 기술이전 내용

- ❖ 태블릿 환경에서 구동 가능한 GAN 기반의 얼굴 이미지 편집 엔진
  - ❖ 고화질 얼굴 이미지 편집 및 복원을 위한 GAN 모델 생성 기술
  - ❖ 태블릿 환경에서 구동이 가능하도록 모델을 변환해주는 기술
  - ❖ 이미지 수정 및 결과 출력을 위한 초기 태블릿 UI, 데스크톱 UI 모듈 (GUI 기반의 시각화된 출력 인터페이스 제공)

▣ 기술이전 범위

- ❖ 기술문서
- ❖ 관련 소스 코드, 샘플 데이터 및 구동 프로그램
- ❖ 공개 데이터에 학습된 모델
- ❖ 시험 절차서 및 결과서

2. 기술이전 내용 및 범위

▣ 기술 개발 현황

❖ 기술성숙도(TRL : Technology Readiness Level) 단계 : (6)단계

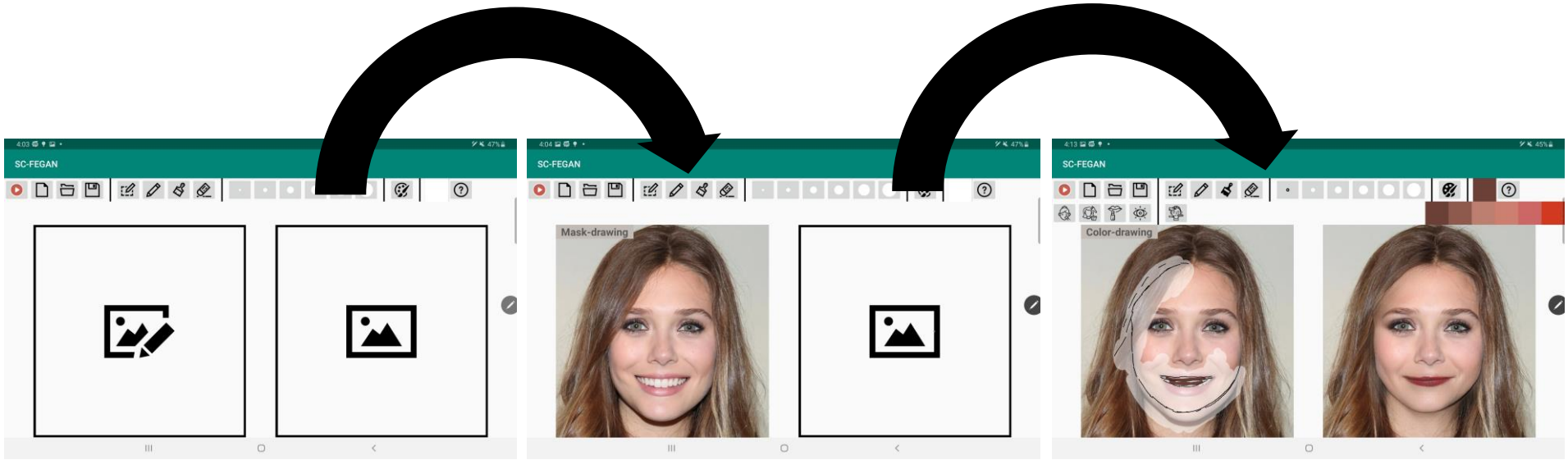
구 분	단계	정 의	세 부 설 명
기초 연구 단계	1	기초 이론/실험	▪기초이론 정립 단계
	2	실용 목적의 아이디어 특허 등 개념정립	▪기술개발 개념 정립 및 아이디어에 대한 특허 출원 단계
실험 단계	3	실험실 규모의 기본성능 검증	▪실험실 환경에서 실험 또는 전산 시뮬레이션을 통해 기본성능이 검증될 수 있는 단계 ▪개발하려는 부품/시스템의 기본 설계도면을 확보하는 단계
	4	실험실 규모의 소재/부품/시스템 핵심성능 평가	▪시험샘플을 제작하여 핵심성능에 대한 평가가 완료된 단계 ▪3단계에서 도출된 다양한 결과 중에서 최적의 결과를 선택하려는 단계 ▪컴퓨터 모사가 가능한 경우 최적화를 완료하는 단계
시작품 단계	5	확정된 소재/부품/시스템 시작품 제작 및 성능 평가	▪확정된 소재/부품/시스템의 실험실 시작품 제작 및 성능 평가가 완료된 단계 ▪개발 대상의 생산을 고려하여 설계하나 실제 제작한 시작품 샘플은 1~수개 미만인 단계 ▪경제성을 고려하지 않고 기술의 핵심성능으로만 볼 때, 실제로 판매가 될 수 있는 정도로 목표 성능을 달성한 단계
	6	파일롯 규모 시작품 제작 및 성능 평가	▪파일롯 규모(복수 개~양산규모의 1/10정도)의 시작품 제작 및 평가가 완료된 단계 ▪파일롯 규모 생산품에 대해 생산량, 생산용량, 불량률 등 제시 ▪파일롯 생산을 위한 대규모 투자가 동반되는 단계 ▪생산기업이 수요기업 적용환경에 유사하게 자체 현장테스트를 실시하여 목표 성능을 만족시킨 단계 ▪성능 평가 결과에 대해 가능하면 공인인증 기관의 성적서 확보
실용화 단계	7	신뢰성평가 및 수요기업 평가	▪실제 환경에서 성능 검증이 이루어지는 단계 ▪부품 및 소재개발의 경우 수요업체에서 직접 파일롯 시작품을 현장 평가(성능 및 신뢰성 평가) ▪가능하면 인증기관의 신뢰성 평가 결과 제출
	8	시제품 인증 및 표준화	▪표준화 및 인허가 취득 단계
사업화	9	사업화	▪본격적인 양산 및 사업화 단계 ▪6-시그마 등 품질관리가 중요한 단계

3. 경쟁기술과 비교

▣ 기술의 주요 특징

❖ <u>태블릿 환경에서 구동이 가능한 GAN 기반의 고화질 얼굴 편집 기술</u>

- 태블릿 환경에서의 고화질의 얼굴 이미지 편집 및 복원 지원
- 사용자의 자유로운 입력을 반영하여 고화질의 얼굴 이미지를 편집하고 특별한 툴 없이 GAN 기술을 활용하여 고품질의 얼굴 이미지 생성 복원이 가능



4. 기술의 사업성

▣ 활용 분야

예상 제품 / 서비스	예상 수요자
얼굴 이미지 편집 시스템	- 성형외과 업체 - 모바일 어플리케이션 공급 업체
얼굴 이미지 복원 시스템	- 영상 분석 사업자 - CCTV 분석 사업자

▣ 기대 효과

❖ 본 기술은 사람의 얼굴 이미지를 PC 및 모바일 환경에서 특정한 툴 없이 자유롭게 수정이 가능하고 이를 위한 기초 UI를 제공. 본 기술은 기학습된 얼굴에 대한 이미지와 비슷한 이미지를 생성할 수 있으므로, 사업화에 필요한 추가적인 기술 개발이 필요함

❖기대 활용처

- 1. 모바일 어플리케이션 : 모바일 환경에서도 간단한 조작으로 높은 성능으로 얼굴 이미지를 수정하여 보여주는 분야. 성형외과 등에서 성형 전후 얼굴 비교를 통한 모바일 홍보 어플리케이션으로 활용 가능

- 2. CCTV 몽타주 분석: CCTV에서 범인과 같은 특정 사람의 얼굴을 다시 복원하거나, 몽타주를 스케치하여 자연스러운 얼굴을 생성하는 방법으로 활용 가능

5. 국내외 시장 동향

▣ 시장전망

❖ 이미지/영상 분석 SW 관련 국내 시장은 2017년 605억원에서 2021년 1,444억원으로 연 평균 15.59% 정도의 성장세, 세계 시장은 2016년 86.7억 달러에서 2021년 170.9억달러로 연평균 약 11.97% 정도의 성장세를 전망

출처: Intelligent Video Analytics-Global intelligent video analytics market(2017-2021)

(단위 : 억달러, 억원)

관련 제품 /서비스	시장	1차년도 (2017)	2차년도 (2018)	3차년도 (2019)	4차년도 (2020)	5차년도 (2021)	합계
지능형 영상 분석	해외	97.7	110.8	127.1	146.7	170.9	653.2
	국내	698	815	966	1,167	1,443	5,091

(출처 :Intelligent Video Analytics-Global intelligent video analytics market(2017-2021))

* 국내 시장은 아시아 지역 시장의 합에서 20% 정도의 시장 규모로 산정, (환률 1$ = 1,100원)

감사합니다.

www.etri.re.kr

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0107742
(43) 공개일자 2020년09월16일

(51) 국제특허분류(Int. Cl.)
G06T 11/60 (2006.01) *G06T 11/00* (2006.01)
G06T 11/20 (2006.01)
(52) CPC특허분류
G06T 11/60 (2013.01)
G06T 11/001 (2013.01)
(21) 출원번호 10-2019-0130282
(22) 출원일자 2019년10월18일
심사청구일자 없음
(30) 우선권주장
1020190026006 2019년03월06일 대한민국(KR)

(71) 출원인
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
조영주
경기도 화성시 향남읍 행정죽전로2길 42-9 201호
박종열
대전광역시 중구 서문로 96, 203동 1503호 (문화동, 센트럴파크2단지아파트)
배유석
대전광역시 유성구 관평1로 12, 704동 1401호 (관평동, 대덕테크노밸리7단지아파트)
(74) 대리인
특허법인지명

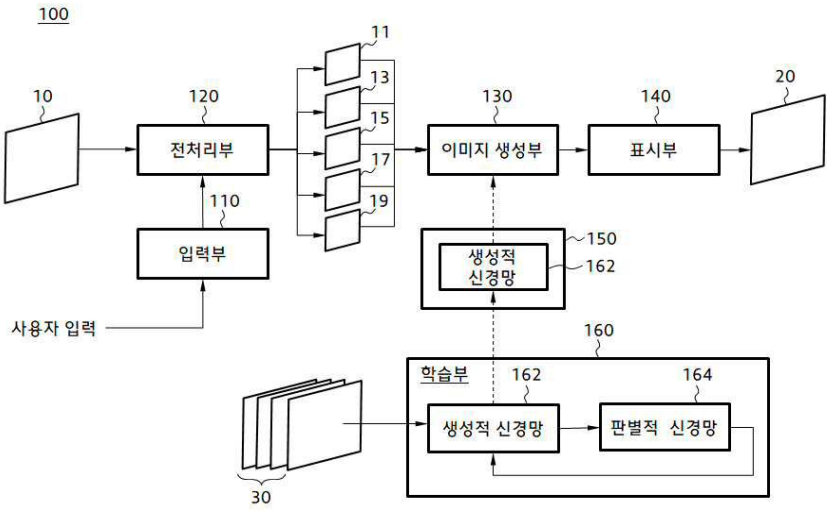
전체 청구항 수 : 총 16 항

(54) 발명의 명칭 **이미지 수정 시스템 및 이의 이미지 수정 방법**

(57) 요 약

본 발명의 이미지 수정 방법은 원본 이미지에 대해 전처리 과정을 수행하여 상기 원본 이미지 내에서 지워진 영역만을 포함하는 마스크(mask) 이미지를 생성하는 단계; 생성적 적대 신경망(Generative Adversarial Networks)을 이용하여 상기 마스크 이미지 내에서 상기 지워진 영역에 합성될 이미지를 예측하는 단계; 및 상기 예측된 이미지를 상기 원본 이미지 내에서 상기 지원진 영역에 합성하여 새로운 이미지를 생성하는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06T 11/20 (2013.01)
G06T 2207/20081 (2013.01)
G06T 2207/20084 (2013.01)
G06T 2207/30201 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 2014-3-00123
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 2019년 RnD 재발견프로젝트
연구과제명 (딥뷰-1세부) 실시간 대규모 영상 데이터 이해·예측을 위한 고성능 비주얼 디스커버리 플랫폼 개발
기 여 율 1/1
과제수행기관명 한국전자통신연구원
연구기간 2019.01.01 ~ 2019.12.31

명 세 서

청구범위

청구항 1

원본 이미지에 대해 전처리 과정을 수행하여 상기 원본 이미지 내에서 지워진 영역만을 포함하는 마스크(mask) 이미지를 생성하는 단계;

생성적 적대 신경망(Generative Adversarial Networks)을 이용하여 상기 마스크 이미지 내에서 상기 지워진 영역에 합성될 이미지를 예측하는 단계; 및

상기 예측된 이미지를 상기 상기 원본 이미지 내에서 상기 지원진 영역에 합성하여 새로운 이미지를 생성하는 단계

를 포함하는 이미지 수정 방법.

청구항 2

제1항에서,

상기 생성적 적대 신경망은 서로 적대적인(Adversarial) 관계를 유지하도록 학습된 생성적 신경망(Generative neural network)과 판별적 신경망(Discriminative neural network)을 포함하고,

상기 예측하는 단계는,

상기 생성적 신경망을 이용하여 상기 합성될 이미지를 예측하는 단계인 것인 이미지 수정 방법.

청구항 3

제1항에서,

상기 마스크 이미지를 생성하는 단계 이전에, 대량의 훈련 데이터를 이용하여 상기 생성적 적대 신경망을 학습시키는 단계를 더 포함하고,

상기 대량의 훈련 데이터는,

훈련용 원본 이미지들, 상기 지워진 영역만을 포함하는 훈련용 마스크 이미지들, 상기 지워진 영역에 다양한 모양이 스케치된 훈련용 스케치 이미지들, 상기 지워진 영역에 다양한 색깔이 채워진 훈련용 색깔 이미지들을 포함하는 것인 이미지 수정 방법.

청구항 4

제1항에서,

상기 마스크 이미지를 생성하는 단계 이전에, 대량의 훈련 데이터를 이용하여 상기 생성적 적대 신경망을 학습시키는 단계를 더 포함하고,

상기 생성적 적대 신경망을 학습시키는 단계는,

상기 대량의 훈련 데이터를 이용하여 생성적 신경망(Generative neural network)을 학습시키는 단계; 및

상기 생성적 신경망과 적대적인(Adversarial) 관계를 유지하도록 판별적 신경망(Discriminative neural network)을 학습시키는 단계를 포함하고,

상기 예측하는 단계는,

학습이 완료된 상기 생성적 신경망을 이용하여 상기 합성될 이미지를 예측하는 단계인 것인 이미지 수정 방법.

청구항 5

제4항에서,

상기 생성적 신경망을 학습시키는 단계는,

상기 판별적 신경망으로부터 출력되는 손실(L_G)을 이용하여 상기 생성적 신경망을 학습시키는 단계이고,

상기 손실(L_G)은,

상기 대량의 훈련 데이터에 포함된 훈련용 마스크 이미지 내에서 지워진 영역의 픽셀값(M)과 상기 원본 이미지와 상기 새로운 이미지 사이의 픽셀 차이값($I_{gen}-I_{gt}$)을 포함하는 것인 이미지 수정 방법.

청구항 6

제4항에서,

상기 생성적 신경망을 학습시키는 단계는,

상기 대량의 훈련 데이터에 포함된 훈련용 마스크 이미지 내에서 지워진 영역의 픽셀값(M)과 상기 원본 이미지와 상기 새로운 이미지 사이의 픽셀 차이값($I_{gen}-I_{gt}$)을 이용하여, 상기 생성적 신경망에 포함된 게이트 컨볼루션 레이어(gated convolution layer)를 학습시키는 단계인 것인 이미지 수정 방법.

청구항 7

원본 이미지에 대해 전처리 과정을 수행하여 상기 원본 이미지 내에서 사용자 입력에 의해 지워진 영역만을 포함하는 마스크(mask) 이미지, 상기 사용자 입력에 의해 상기 지워진 영역에 스케치된 모양만을 포함하는 스케치 이미지 및 상기 사용자 입력에 의해 상기 지워진 영역에 칠해진 색깔만을 포함하는 색깔 이미지를 생성하는 전처리부;

생성적 적대 신경망(Generative Adversarial Networks)을 이용하여, 상기 마스크 이미지, 상기 스케치 이미지 및 상기 색깔 이미지로부터 상기 지워진 영역에 합성될 이미지를 예측하고, 상기 예측된 이미지를 상기 지워진 영역에 합성하여 상기 원본 이미지로부터 새로운 이미지를 생성하는 이미지 생성부; 및

상기 새로운 이미지를 표시하는 표시부

를 포함하는 이미지 수정 시스템.

청구항 8

제7항에서,

상기 생성적 적대 신경망은 서로 적대적인(Adversarial) 관계를 유지하도록 학습된 생성적 신경망(Generative neural network)과 판별적 신경망(Discriminative neural network)을 포함하고,

상기 이미지 생성부는,

상기 생성적 신경망을 이용하여 상기 합성될 이미지를 예측하는 것인 이미지 수정 시스템.

청구항 9

제7항에서,

대량의 훈련 데이터를 이용하여 상기 생성적 적대 신경망을 학습시키는 학습부; 및

상기 학습부에 의해 학습이 완료된 상기 생성적 적대 신경망을 저장하는 저장부를 더 포함하고,

상기 대량의 훈련 데이터는,

훈련용 원본 이미지들, 상기 지워진 영역만을 포함하는 훈련용 마스크 이미지들, 상기 지워진 영역에 다양한 모양이 스케치된 훈련용 스케치 이미지들, 상기 지워진 영역에 다양한 색깔이 채워진 훈련용 색깔 이미지들을 포함하는 것인 이미지 수정 시스템.

청구항 10

제9항에서,

상기 학습부는,

서로 적대적인(Adversarial) 관계를 유지하도록 생성적 신경망(Generative neural network)과 판별적 신경망(Discriminative neural network)을 포함하는 상기 생성적 적대 신경망을 학습시키고,

상기 저장부는,

학습이 완료된 상기 생성적 신경망만이 저장되는 것인 이미지 수정 시스템.

청구항 11

제10항에서,

상기 학습부는,

상기 판별적 신경망으로부터 출력되는 손실(L_G)을 감소시키는 방향으로 상기 생성적 신경망을 학습시키고,

상기 손실(L_G)은,

상기 대량의 훈련 데이터에 포함된 훈련용 마스크 이미지 내에서 지워진 영역의 픽셀값(M)과 상기 원본 이미지와 상기 새로운 이미지 사이의 픽셀 차이값($I_{gen}-I_{gt}$)을 포함하는 것인 이미지 수정 시스템.

청구항 12

제10항에서,

상기 학습부는,

상기 대량의 훈련 데이터에 포함된 훈련용 마스크 이미지 내에서 지워진 영역의 픽셀값(M)과 상기 원본 이미지와 상기 새로운 이미지 사이의 픽셀 차이값($I_{gen}-I_{gt}$)을 이용하여, 상기 생성적 신경망에 포함된 게이트 컨볼루션 레이어(gated convolution layer)를 학습시키는 것인 이미지 수정 시스템.

청구항 13

서로 적대적인 관계에 있는 생성적 신경망과 판별적 신경망으로 구성된 생성적 적대 신경망(Generative Adversarial Networks)을 학습시키는 단계;

학습이 완료된 상기 생성적 신경망을 저장부에 저장하는 단계;

전치치 과정을 통해 원본 이미지로부터 상기 원본 이미지 내에서 사용자 입력에 의해 지워진 영역만을 포함하는 마스크(mask) 이미지, 상기 사용자 입력에 의해 상기 지워진 영역에 스케치된 모양만을 포함하는 스케치 이미지 및 상기 사용자 입력에 의해 상기 지워진 영역에 칠해진 색깔만을 포함하는 색깔 이미지를 생성하는 단계;

상기 저장부에 저장된 생성적 신경망을 이용하여, 상기 마스크 이미지, 상기 스케치 이미지 및 상기 색깔 이미지로부터 상기 지워진 영역에 합성될 이미지를 예측하는 단계; 및

상기 예측된 이미지를 상기 지워진 영역에 합성하여 상기 원본 이미지로부터 새로운 이미지를 생성하는 단계

를 포함하는 이미지 수정 방법.

청구항 14

제13항에서,

상기 학습시키는 단계는,

서로 적대적인(Adversarial) 관계를 유지하도록 생성적 신경망(Generative neural network)과 판별적 신경망(Discriminative neural network)을 학습시키는 단계이고,

상기 생성적 신경망을 학습시키는 단계는,

대량의 훈련 데이터에 포함된 훈련용 마스크 이미지 내에서 지워진 영역의 픽셀값(M)과 상기 원본 이미지와 상기 새로운 이미지 사이의 픽셀 차이값($I_{gen}-I_{gt}$)을 이용하여, 상기 생성적 신경망에 포함된 게이트 컨볼루션 레이

어(gated convolution layer)를 학습시키는 단계인 것인 이미지 수정 방법.

청구항 15

제13항에서,

상기 지워진 영역에 합성될 이미지를 예측하는 단계에서,

상기 합성될 이미지는,

상기 지워진 영역에 스케치된 모양에 따라 예측된 사실적인 이미지인 것인 이미지 수정 방법.

청구항 16

제13항에서,

상기 원본 이미지는 얼굴 이미지인 것인 이미지 수정 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 이미지를 수정하기 위한 기술에 관한 것으로, 더욱 상세하게는 얼굴 이미지를 수정하기 위한 기술에 관한 것이다.

배 경 기 술

[0003] 최근 SNS 를 통해 이미지와 같은 다양한 정보를 공유하는 사람들이 증가하는 추세이고, 이러한 추세에 따라, 이미지 편집 프로그램이나 어플리케이션에 대한 관심이 높아지고 있다.

[0004] 종래의 이미지 편집 프로그램(또는 이미지 편집 툴)이나 이와 관련된 어플리케이션은 주로 이미지의 픽셀 값을 조절하는 방식으로 이미지를 수정한다. 사실적으로 수정된 이미지는 이미지 편집 프로그램을 다루는 사용자의 숙련도에 따라 결정된다.

[0005] 따라서, 이미지 편집 프로그램에 대해 전문적인 지식이나 경험이 적은 일반 사용자가 이미지 편집 프로그램을 이용하여 이미지를 수정할 경우, 그 결과물은 사실적이지 못하고, 어색한 이미지일 가능성이 높다.

[0006] 이미지 편집과 관련해, 이미지에 대해 3D 모델링을 수행하여 획득한 3D 모델을 엔진(이하, 3D 모델 엔진)을 이용하여 편집하는 방식도 있다. 그러나, 이러한 방식은 3D 모델을 구현하는 3D 모델 엔진이 필요하며 이러한 3D 모델 엔진은 다양하고, 사용하는 3D 모델 엔진마다 결과물의 완성도나 완성에 필요한 지식이 천차만별로 다르다.

[0007] 또한 각 3D 모델 엔진에 대한 사용법을 사용자가 충분히 숙지하고 있어야 하는 한계가 존재한다. 즉, 기존의 방법들은 단순한 이미지 입력만으로 이미지를 사실적으로 수정하는 것이 매우 어렵다.

발명의 내용

해결하려는 과제

[0009] 본 발명은 얼굴 이미지와 사용자 입력을 이용하여 사전에 학습시킨 신경망을 이용하여, 기존 방식에 비해 쉽고 빠르게 사실적인 합성 이미지를 제공할 수 있는 얼굴 이미지 수정 시스템 및 그 방법을 제공하는 데 목적이 있다.

[0010] 즉, 본 발명에서는 사용자가 쉽고 직관적으로 얼굴 이미지를 수정하고 사실적인 결과물을 얻을 수 있도록 하는 시스템과 방법을 제시한다. 사실적인 결과물을 얻기 위하여 전문적인 지식이나 경험을 필요로 하는 것이 아닌 누구나 쉽게 얼굴 이미지를 수정하도록 하는 것이 본 발명의 목적이다.

[0011] 본 발명의 전술한 목적들 및 그 이외의 목적과 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부된 도면과

함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다.

과제의 해결 수단

[0013] 상술한 목적을 달성하기 위한 본 발명의 일측면에 따른 이미지 수정 방법은 원본 이미지에 대해 전처리 과정을 수행하여 상기 원본 이미지 내에서 지워진 영역만을 포함하는 마스크(mask) 이미지를 생성하는 단계; 생성적 적대 신경망(Generative Adversarial Networks)을 이용하여 상기 마스크 이미지 내에서 상기 지워진 영역에 합성될 이미지를 예측하는 단계; 및 상기 예측된 이미지를 상기 원본 이미지 내에서 상기 지원진 영역에 합성하여 새로운 이미지를 생성하는 단계를 포함한다.

[0014] 본 발명의 다른 측면에 따른 이미지 수정 시스템은, 원본 이미지에 대해 전처리 과정을 수행하여 상기 원본 이미지 내에서 사용자 입력에 의해 지워진 영역만을 포함하는 마스크(mask) 이미지, 상기 사용자 입력에 의해 상기 지워진 영역에 스케치된 모양만을 포함하는 스케치 이미지 및 상기 사용자 입력에 의해 상기 지워진 영역에 칠해진 색깔만을 포함하는 색깔 이미지를 생성하는 전처리부;

[0015] 생성적 적대 신경망(Generative Adversarial Networks)을 이용하여, 상기 마스크 이미지, 상기 스케치 이미지 및 상기 색깔 이미지로부터 상기 지워진 영역에 합성될 이미지를 예측하고, 상기 예측된 이미지를 상기 지워진 영역에 합성하여 상기 원본 이미지로부터 새로운 이미지를 생성하는 이미지 생성부; 및 상기 새로운 이미지를 표시하는 표시부를 포함한다.

[0016] 본 발명의 또 다른 측면에 따른 이미지 수정 방법은, 서로 적대적인 관계에 있는 생성적 신경망과 판별적 신경망으로 구성된 생성적 적대 신경망(Generative Adversarial Networks)을 학습시키는 단계; 학습이 완료된 상기 생성적 신경망을 저장부에 저장하는 단계; 전처리 과정을 통해 원본 이미지로부터 상기 원본 이미지 내에서 사용자 입력에 의해 지워진 영역만을 포함하는 마스크(mask) 이미지, 상기 사용자 입력에 의해 상기 지워진 영역에 스케치된 모양만을 포함하는 스케치 이미지 및 상기 사용자 입력에 의해 상기 지워진 영역에 칠해진 색깔만을 포함하는 색깔 이미지를 생성하는 단계; 상기 저장부에 저장된 생성적 신경망을 이용하여, 상기 마스크 이미지, 상기 스케치 이미지 및 상기 색깔 이미지로부터 상기 지워진 영역에 합성될 이미지를 예측하는 단계; 및 상기 예측된 이미지를 상기 지워진 영역에 합성하여 상기 원본 이미지로부터 새로운 이미지를 생성하는 단계를 포함한다.

발명의 효과

[0018] 본 발명에 의하면, 적대적 신경망을 이용하여 얼굴 이미지를 수정함으로써, 별도의 이미지 툴에 대한 전문적인 지식이나 경험 없이 사용자가 원하는 방식으로 얼굴 이미지를 빠르고 간편하게 할 수 있다.

[0019] 기존의 이미지 수정 프로그램은 얼굴을 가늘게 만들거나 눈을 키우거나 하기 위하여 각각에 해당하는 툴이 따로 존재하고 이를 잘 활용하기 위해서는 사용자의 많은 경험이 필요했다.

[0020] 하지만, 본 발명에서 제공하는 시스템은 별도의 이미지 툴이 필요하지 않고, 마스크, 스케치 또는 색깔을 입력 정보로 사용하여 사용자가 원하는 방향으로 얼굴 이미지를 수정할 수 있다.

도면의 간단한 설명

[0022] 도 1은 본 발명의 실시 예에 따른 얼굴 이미지 수정 시스템의 블록도.

도 2는 도 1에 도시한 전처리부로부터 이미지 생성부로 입력되는 다수의 입력 이미지를 설명하기 위한 도면.

도 3 내지 도 7은 본 발명의 실시 예에 따른 사용자 인터페이스의 화면 구성을 나타내는 도면들.

도 8은 본 발명에 적용되는 생성적 적대 신경망(GANs)의 전체 네트워크 구조를 나타내는 도면.

도 9는 본 발명의 실시 예에 따른 이미지 수정 방법을 보여주는 흐름도.

도 10 내지 12는 본 발명의 이미지 수정 방법에 따라 원본 이미지를 수정한 결과 이미지들의 예시한 도면들.

발명을 실시하기 위한 구체적인 내용

[0023] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시 예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.

[0024] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 변경, 균등물, 또는 대체물을 포함한다.

[0025] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만, 예를 들어 본 발명의 개념에 따른 권리 범위로부터 이탈되지 않은 채, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.

[0026] 본 명세서에서 사용한 용어는 단지 특정한 실시예들을 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설시된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0028] 이하, 실시 예들을 첨부된 도면을 참조하여 상세하게 설명한다. 그러나, 특허출원의 범위가 이러한 실시 예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0030] 도 1은 본 발명의 실시 예에 따른 이미지 수정 시스템의 블록도이다.

[0031] 도 1을 참조하면, 본 발명의 실시 예에 따른 이미지 수정 시스템(100)는 이미지 수정 프로그램과 같은 툴에 대한 전문적인 지식이나 충분한 경험이 없는 사용자가 이미지를 원하는 방식으로 빠르고 간편하게 수정 할 수 있다.

[0032] 본 실시 예에 따른 이미지 수정 시스템이 얼굴 이미지를 수정하는 것으로 한정한다. 따라서, 이하에서는 이미지 수정 시스템을 '얼굴 이미지 수정 시스템'으로 지칭한다. 그러나 본 발명이 얼굴 이미지 수정에 제한적으로 적용되는 것은 아니며, 차량 디자인, 건축 디자인, 가전 제품 디자인 등과 관련된 모든 종류의 이미지 수정에 적용될 수 있다.

[0033] 사용자가 간편하고 빠르게 얼굴 이미지를 수정하는 시스템을 설계하기 위해, 얼굴 이미지 수정 시스템(100)은 입력부(110), 전치리부(120), 이미지 생성부(130), 표시부(140), 저장부(150) 및 학습부(160)를 포함한다.

[0035] **입력부(110)**

[0036] 입력부(110)는 사용자 입력을 레이어 생성부(120)로 전달하는 구성으로, 키보드, 마우스, 터치 패드, 터치 패널과 같은 하드웨어 수단과 이러한 하드웨어 수단과 연동하도록 프로그래밍된 소프트웨어 수단(이하, '사용자 인터페이스'라 함)을 포함한다.

[0037] 사용자는 입력부(110)를 통해 다음과 같은 작업을 수행할 수 있다.

[0038] - 원본 얼굴 이미지(Original face image, 10)에서 사용자가 수정하고자 하는 영역을 지우는 작업.

[0039] - 원본 이미지(10)에서 지워진 영역(erased area) 내에 사용자가 수정하고자 하는 모양(또는 형태)으로 스케치하는 작업. 여기서, 스케치는 그 모양을 알아볼 수 있는 수준을 의미하며, 전문가 수준의 스케치는 불필요하다.

[0040] - 스케치한 모양을 사용자가 수정하고자 하는 컬러로 색칠하는 작업

[0041] 실시 예에 따르면, 사용자가 수정하고자 하는 영역은 원본 얼굴 이미지에서 눈 영역, 귀 영역, 입 영역, 코 영

역, 헤어 영역 등을 포함한다.

[0042] 실시 예에 따르면, 사용자가 수정하고자 하는 모양(또는 형태)은 눈 영역에 착용하는 안경, 귀 영역에 착용하는 귀걸이, 입술 모양, 헤어 스타일 등을 포함한다.

[0043] 실시 예에 따르면, 사용자가 수정하고자 하는 컬러는 눈동자의 색깔, 귀걸이의 색깔, 입술의 색깔, 헤어 색깔 등을 포함한다.

[0044] 입력부(110)는, 상기의 작업들에 대응하는 사용자 입력에 따라, 원본 이미지에서 지워진 영역을 나타내는 제1 입력 값, 상기 지워진 영역에 스케치된 모양(또는 형태)를 나타내는 제2 입력 값 및 사용자가 수정하고자 하는 색깔을 나타내는 제3 입력 값을 생성하여 이를 전처리부(120)로 입력한다.

[0046] **전처리부(120)**

[0047] 전처리부(120)는 상기 입력부(110)로부터 입력된 제1 내지 제3 입력 값에 따라 원본 얼굴 이미지를 전처리하여, 전처리 된 다수의 입력 이미지를 생성한다. 생성된 다수의 입력 이미지는 이미지 생성부(130)로 입력되는 정보로 사용된다.

[0048] 도 2는 본 발명의 실시 예에 따른 전처리부로부터 이미지 생성부로 입력되는 다수의 입력 이미지를 설명하기 위한 도면이다.

[0049] 도 2를 참조하면, 다수의 입력 이미지는 상기 제1 입력 값에 따라 사용자에 의해 지워진 영역을 포함하는 원본 얼굴 이미지(11), 상기 제1 입력 값에 따라 상기 지워진 영역만을 포함하는 마스크 이미지(mask image)(13), 상기 제2 입력 값에 따라 사용자에 의해 상기 지워진 영역에 스케치된 모양만을 나타내는 스케치 이미지(15), 상기 제3 입력 값에 따라 상기 지워진 영역에 나타나는 색깔만을 포함하는 색깔 이미지(17) 및 상기 지워진 영역의 가우시안 노이즈만을 나타내는 노이즈 이미지(19)를 포함한다.

[0051] **이미지 생성부(130)**

[0052] 다시 도 1을 참조하면, 이미지 생성부(130)는 사전에 학습된 신경망을 이용하여 상기 전처리부(120)로부터 입력된 상기 다수의 입력 이미지를 분석하고, 그 분석 결과에 따라 상기 원본 얼굴 이미지(10) 내에서 상기 지워진 영역에 사용자가 합성하고자 하는 사실적인 이미지를 예측하여, 상기 예측된 사실적인 이미지를 상기 원본 얼굴 이미지(10)의 지워진 영역에 자동으로 합성한다.

[0053] 실시 예에 따르면, 상기 사전에 학습된 신경망은 생성적 적대 신경망(Generative Adversarial Networks, GANs)일 수 있다. 이에 따라, 이미지 생성부(130)는 생성적 적대 신경망(GANs)을 이용하여 원본 얼굴 이미지를 수정한다.

[0054] 생성적 적대 신경망(GANs)은 2014년 몬트리올 대학의 Ian Goodfellow와 Yoshua Bengio를 포함하는 연구자들이 작성한 논문에서 소개된 심층 신경망 기술이다.

[0055] 생성적 적대 신경망(GANs)에 대한 상세한 설명은 상기 논문으로 대신하고, 본 명세서에서는 생성적 적대 신경망(GANs)의 개략적인 동작 원리와 본 발명에 적용되도록 수정 및 변경된 부분에 대해서만 설명하기로 한다.

[0056] 생성적 적대 신경망(GANs)은 생성기(Generator)로 불릴 수 있는 생성적 신경망(162, Generative neural network)과 판별기(Discriminator)로 불릴 수 있는 판별적 신경망(164, Discriminative neural network)으로 구성된 심층 신경망 구조를 갖는다.

[0057] 생성적 신경망(162)은 판별적 신경망(164)으로 전달하는 새로운 이미지를 생성하고, 판별적 신경망(164)은 생성적 신경망(162)로부터 입력된 새로운 이미지의 진위 여부를 판단한다.

[0058] 생성적 신경망(162)은 새로운 이미지를 생성하고, 판별적 신경망(164)는 생성적 신경망(162)으로부터 입력되는 새로운 이미지의 진위를 판별하고, 그 판별 결과를 생성적 신경망(162)의 입력으로 피드백한다.

[0059] 이때, 생성적 신경망(162)은 자신이 생성한 새로운 이미지를 판별적 신경망(164)이 진짜 이미지(real image)로 판별하도록 이미지 생성 과정을 학습하고, 반대로 판별적 신경망(164)은 생성적 신경망(162)으로부터 입력된 새로운 이미지를 가짜 이미지로 판별하도록 이미지 판별 과정을 학습한다.

[0060] 이처럼 생성적 신경망(162)과 판별적 신경망(164)은 제로섬 게임처럼 서로 반대되는 목적 함수 또는 손실 함수를 통해 학습된다. 즉, 생성적 신경망(162)과 판별적 신경망(164)은 서로 적대적인(Adversarial) 관계를 유지하도록 학습되고 진화한다.

[0061] 이미지 생성부(130)는 생성적 신경망(162)과 판별적 신경망(164) 중에서 학습이 완료된 생성적 신경망(162)만을 이용하여 원본 얼굴 이미지에서 지워진 영역에 합성되는 사실적인 이미지를 생성한다.

[0063] **저장부(150)**

[0064] 저장부(150)에는 이미지 생성부(130)가 사용하는 생성적 신경망(162)이 저장된다. 생성적 신경망(162)은 생성 알고리즘(Generative Algorithms)이라는 용어로 대체될 수 있다. 유사하게, 판별적 신경망(164)은 판별 알고리즘(Discriminative Algorithms)이라는 용어로 대체될 수 있다. 저장부(150)는 휘발성 메모리 및 비휘발성 메모리로 구현될 수 있다.

[0066] **학습부(160)**

[0067] 학습부(160)는 사전에 수집된 대용량의 훈련 데이터를 이용하여 생성적 적대 신경망(GANs), 즉, 생성적 신경망(162)과 판별적 신경망(164)을 학습시킨다.

[0068] 여기서, 대용량의 훈련 데이터는 대용량의 훈련용 얼굴 이미지, 서로 다른 위치에서 지워진 영역을 포함하도록 구성된 대용량의 훈련용 얼굴 이미지, 상기 지워진 영역만을 나타내는 대용량의 훈련용 마스크 이미지, 상기 지워진 영역에 다양한 모양이 스케치된 대용량의 훈련용 스케치 이미지, 상기 지워진 영역에 서로 다른 색깔이 채워진 대용량의 훈련용 색깔 이미지 및 대용량의 노이즈 이미지를 포함한다.

[0069] 이미지 생성부는 학습이 완료된 생성적 신경망(162)과 판별적 신경망(164) 중에서 생성적 신경망(162)만을 이용하므로, 학습부(160)는 학습이 완료된 생성적 신경망(162)과 판별적 신경망(164) 중에서 생성적 신경망(162)을 저장부(150)에 저장함으로써, 이미지 생성부(130)는 저장부(150)에 저장된 생성적 신경망(162)을 이용할 수 있게 된다.

[0071] **표시부(140)**

[0072] 표시부(140)는 이미지 생성부(130)에서 생성한 이미지를 표시하는 구성으로, 상기 원본 얼굴 이미지(10) 내에서 사용자 입력에 따라 지워진 영역에 사용자가 합성하고자 하는 사실적인 이미지가 자동으로 합성된 합성 이미지를 표시한다. 이러한 표시부(140)는 LCD, LED, OLE 등일 수 있다.

[0073] 또한 표시부(140)는 사용자 인터페이스의 화면 구성을 표시한다.

[0075] 도 3 내지 도 7은 본 발명의 실시 예에 따른 사용자 인터페이스의 화면 구성을 나타내는 도면들이다.

[0076] 사용자 인터페이스는 사용자에게 원본 얼굴이미지의 수정에 필요한 입력, 즉, 이미지 생성부로 입력되는 입력 정보를 생성하기 위한 환경을 제공한다.

[0077] 이를 위해, 사용자 인터페이스의 화면 구성은 원본 얼굴 이미지 및 원본 얼굴 이미지에 대한 진행 과정이 표시되는 영역(31)과 수정된 결과(합성 이미지)가 표시되는 영역(32)을 포함한다.

[0078] 또한, 사용자 인터페이스의 화면 구성은 입력 정보 생성과 관련된 다수의 아이콘 형태의 버튼(33, 34, 35, 36 및 37)을 포함한다.

[0079] 먼저, 도 3을 참조하면, 버튼(33)은 원본 이미지의 불러오기 기능을 제공한다. 사용자가 버튼(33)을 터치 또는 클릭하면, 원본 이미지가 영역(31)에 표시된다.

[0080] 도 4를 참조하면, 버튼(34)은 사용자가 원본 얼굴 이미지(10)에서 수정할 영역을 지우는 기능을 제공한다. 예를 들면, 사용자가 버튼(34)을 터치 또는 클릭하면, 지우개 형태의 아이콘 도구가 생성되고, 지우개 형태의 아이콘 도구를 이용하여 원본 이미지(10)에서 수정할 부분을 지운다. 도 4는 사용자가 원본 얼굴 이미지(10)에서 수정

하고자 하는 영역이 눈 영역, 코 영역, 입술 영역인 경우를 도시한 것이다.

[0081] 도 5를 참조하면, 버튼(35)은 사용자가 원본 얼굴 이미지(10)에서 지워진 영역에 수정하고자 하는 모양을 스케치하는 기능을 제공한다. 예를 들면, 사용자가 버튼(35)을 클릭 또는 터치하면, 펜 형태의 아이콘 도구가 생성되고, 펜 형태의 아이콘 도구를 이용하여 원본 얼굴 이미지(10)에서 지워진 영역에 수정하고자 하는 모양을 스케치한다. 이때, 스케치는 그 모양을 알아볼 수 있는 수준을 의미하며, 전문가 수준의 스케치는 불필요하다.

[0082] 도 6을 참조하면, 버튼(36)은 스케치한 모양을 사용자가 수정하고자 하는 컬러로 색칠하는 기능을 제공한다. 예를 들면, 사용자가 버튼(36)을 클릭 또는 터치하면, 붓 형태의 아이콘 도구가 생성되고, 붓 형태의 아이콘 도구를 이용하여 스케치한 모양 또는 그 주변에 사용자가 결정한 색깔을 칠한다. 도 6에서는 눈동자의 색깔을 사용자가 정한 색깔로 칠한 경우를 도시한 것이다.

[0083] 도 7을 참조하면, 버튼(37)은 수정된 결과를 보여주는 기능을 제공한다. 사용자가 버튼(37)을 클릭 또는 터치하면, 영역(32)에 상기 원본 얼굴 이미지(10) 내에서 사용자 입력에 따라 지워진 영역에 사용자가 합성하고자 하는 사실적인 이미지가 자동으로 합성된 합성 이미지가 표시된다. 도 7에서는 생성적 신경망(162)의 이미지 생성 과정에서 수정된 코 형상 이미지, 입술 형상 이미지 및 눈동자 색상 이미지가 원본 얼굴 이미지에 합성된 예를 도시한 것이다.

[0084] 도 7에서는 생성적 신경망(162)이 마스크 작업(원본 얼굴 이미지에서 사용자가 수정하고자 하는 영역을 지우는 작업), 스케치 작업 및 색깔 작업에 따라 생성된 모든 입력 정보를 이용하여 합성된 이미지를 생성한 예를 도시하고 있으나, 합성된 이미지를 생성하기 위해 모든 입력 정보가 필요한 것은 아니다.

[0085] 예를 들면, 생성적 신경망(162)이 마스크 작업에 따라 생성된 오직 하나의 입력 정보, 즉, 마스크 이미지만을 이용하여 합성된 이미지를 생성할 수도 있다. 이는 생성적 신경망(162)이 대용량의 훈련 데이터, 즉, 대용량의 마스크 이미지, 대용량의 스케치 이미지 및 대용량의 색깔 이미지를 모두 이용하여 학습된 것이기 때문이다.

[0086] 이것은 원본 얼굴 이미지(10)에서 사용자가 수정하고자 하는 영역의 위치를 나타내는 마스크 이미지는 반드시 생성적 신경망(162)의 입력으로 사용되어야 함을 의미하기도 한다. 즉, 마스크 이미지는 사용자 수정하고자 하는 의도를 생성적 신경망(162)에게 알리는 최소한 정보로 해석할 수 있다.

[0087] 물론, 생성적 신경망(162)으로 입력되는 정보의 량이 많을수록 사용자가 수정하고자 하는 의도에 가장 부합하는 수정된 얼굴 이미지가 생성될 확률은 당연히 가장 높을 것이다.

[0089] 도 8은 본 발명에 적용되는 생성적 적대 신경망(GANs)의 전체 네트워크 구조를 나타내는 도면이다.

[0090] 전술한 바와 같이, 생성적 신경망(162)을 학습시키기 위해서는 생성적 긴경망(162)과 판별적 신경망(164)으로 구성된 전체 네트워크(GANs)를 같이 학습시켜야 한다.

[0091] 학습이 완료되면, 얼굴 이미지의 수정에는 생성적 신경망(162)만이 사용되므로, 생성적 신경망(162)의 속도 향상과 수정된 얼굴 이미지의 사실성을 높이기 위해, 생성적 적대 신경망(GANs)은, 도 8에 도시된 바와 같은 구조로 구성될 수 있다.

[0092] 도 8에 도시된 바와 같이, 생성적 신경망(162)은 U-net 구조로 이루어져 있으며 게이트 컨볼루션 레이어(Gated convolution layer)들, 확장된 게이트 컨볼루션 레이어(Dilated gated convolution layer)들, 디컨볼루션 레이어(Deconvolution layer)들을 포함한다. 각 컨볼루션 레이어는 채널 수에 따라 서로 다른 사이즈의 육면체 형상으로 도시된다.

[0093] 생성적 신경망(162)은 일반적인 컨볼루션 레이어(convolution layer)가 아니라 게이트 컨볼루션 레이어(gated convolution layer)를 사용함에 특징이 있다. 일반적인 컨볼루션 레이어는 이전단의 컨볼루션 레이어로부터 입력된 특징값에 대해 다른 특징값을 출력한다. 이에 반해, 게이트 컨볼루션 레이어는 이전단의 게이트 컨볼루션 레이어로부터 입력된 특징에 대해 다른 특징값과 마스크 이미지에 대한 특징값을 출력한다. 즉, 일반적인 컨볼루션 레이어는 하나의 데이터를 출력하고, 게이트 컨볼루션 레이어는 두 개의 데이터를 출력하는 점에서 차이가 있다. 게이트 컨볼루션 레이어에서 출력되는 2개의 데이터 중 하나의 데이터는 마스크 이미지의 특징값이고, 이 마스크 이미지의 특징값은 인접하지 않은 다른 게이트 컨볼루션 레이어로 입력된다. 도 8에서 점선의 화살표는 현재의 게이트 컨볼루션 레이어가 출력하는 마스크 이미지의 특징값이 인접하지 않은 다른 게이트 컨볼루션 레이어로의 입력을 나타낸 것이다.

[0094] 판별적 신경망(164)은 스펙트럼 정규화(Spectral normalization)가 적용된 게이트 컨볼루션 레이어(gated convolution layer)로 이루어진 patchGAN 형태이다. 판별적 신경망(164)을 학습시키는 과정은 데이터셋(dataset)에 대하여 일반적인 생성적 적대 신경망(GANs)을 학습시키는 방법과 동일하다.

[0095] 다만, 판별적 신경망(164)의 학습에 사용되는 손실 변수(Loss)는 일반적인 손실(Loss)와는 차이가 있다. 여기서, 손실(Loss)는 원본 얼굴 이미지와 생성적 신경망(162)이 생성한 새로운 얼굴 이미지와의 차이를 의미한다.

[0096] 본 발명의 실시 예에 따른 생성적 신경망(162)을 학습시키는데 사용되는 파라미터(L_G)는 아래의 수학식 1과 같고, 판별적 신경망(164)을 학습시키는데 사용되는 파라미터(L_D)는 아래의 수학식 2와 같다.

수학식 1

[0098]
$$L_G = L_{per-pixel} + \sigma L_{percept} + \beta L_{G\_SN} + \gamma(L_{sytle}(I_{gen}) + L_{sytle}(I_{comp})) + \upsilon L_{tv} + \epsilon \mathbb{E}\left[D(I_{gt})^2\right]$$

수학식 2

[0099]
$$L_D = \mathbb{E}[1 - D(I_{gt})] + \mathbb{E}[1 + D(I_{comp})] + \theta L_{GP}$$

[0101] 파라미터 L_G는 생성적 신경망(162)의 레이어(layer)를 학습시키는 손실(loss)이다. 여기서, 생성적 신경망(162)의 레이어(layer)는 도 8에 도시된 바와 같이, 다수의 게이트 컨볼루션 레이어(Gated convolution layer), 다수의 확장된 게이트 컨볼루션 레이어(Dilated gated convolution layer) 및 다수의 디컨볼루션 레이어(Deconvolution layer)을 포함한다.

[0102] 파라미터 L_D는 판별적 신경망(164)의 레이어(layer)를 학습시키는 손실(loss)이다. 여기서, 판별적 신경망(164)의 레이어(layer)는 다수의 스펙트럼 정규화(Spectral Normalization: SN) 컨볼루션 레이어를 포함한다.

[0104] 수학식 1에서 $L_{per-pixel}$은 아래의 수학식 3과 같다.

수학식 3

[0105]
$$L_{per-pixel} = \frac{1}{N_{I_{gt}}} \|M \odot (I_{gen} - I_{gt})\|_1 + \alpha \frac{1}{N_{I_{gt}}} \|(1 - M) \odot (I_{gen} - I_{gt})\|_1$$

[0106] 여기서, M은 마스크 이미지에 포함된 지워진 영역을 나타내는 1채널의 픽셀값으로서, 예를 들어, '1'일 수 있다. 이때, 마스크 이미지에서 지워진 영역을 제외한 나머지 영역의 각 픽셀값은 '0'이다. I_{gen}은 생성적 신경망(162)이 지워진 영역을 갖는 원본 얼굴 이미지, 상기 지워진 영역만을 갖는 마스크 이미지, 스케치 이미지,

색깔 이미지 및 노이즈 이미지를 입력받아서 생성한 새로운 얼굴 이미지, 즉, 원본 얼굴 이미지에서 지워진 영역이 다른 이미지로 채워진 이미지를 의미하고, I_{gen}는 생성적 신경망(162)이 생성한 새로운 얼굴 이미지를 3차원 벡터 공간에서 표현한 3차원 벡터값일 수 있다. I_{gt}는 지워진 영역이 없는 원본 얼굴 이미지를 의미하고, I_{gt}는 원본 얼굴 이미지를 3차원 벡터 공간에서 표현한 3차원 벡터값일 수 있다. $L_{per-pixel}$은 생성적 신경망에서 생성한 새로운 얼굴 이미지와 원본 얼굴 이미지 간의 거리 L1를 계산하는 파라미터 이다. 거리 L은 새로운 얼굴 이미지의 픽셀과 상기 픽셀에 대응하는 원본 얼굴 이미지의 픽셀 간의 거리이다. α는 원본 얼굴 이미지 내에서 지워진 영역에 대한 손실(loss)를 강화시키기 위한(줄이기 위한) 가중치(weight)이다. ⊙는 요소 별 곱셈 연산(Element-wise multiplication)을 나타내는 기호이고, $N_{I_{gt}}$는 이미지 크기에 따른 총 픽셀 수로서, 예컨대, 786432(= 512 x 512 x 3)일 수 있다.

[0108] 수학식 1에서 $L_{percept}$는 아래의 수학식 4와 같다.

수학식 4

$$L_{percept} = \sum_{q} \frac{\|\Theta_q(I_{gen}) - \Theta_q(I_{gt})\|_1}{N_{\Theta_q(I_{gt})}} + \sum_{q} \frac{\|\Theta_q(I_{comp}) - \Theta_q(I_{gt})\|_1}{N_{\Theta_q(I_{gt})}}$$

[0109]

[0110] $L_{percept}$는 스타일 손실(style-loss)로 활용되는 손실(loss) 중 하나이다. Θ_q는 대량의 훈련 데이터 셋에 대한 이미지 분류(classification)을 위하여 학습된 생성적 신경망의 q번째 레이어(layer)의 특징을 의미한다. 즉, $L_{percept}$는 기존의 학습된 네트워크를 활용하여 픽셀 단위로 계산된 특징에 대한 손실(loss)이다. $N_\Theta_q(I_gt)$에서 N은 픽셀 총 개수이고, Θ_q는 생성적 신경망과는 다른 기존의 이미지 분류를 위해 대량의 이미지를 학습한 인공신경망의 q번째 특징맵이다. 즉, $N_\Theta_q(I_gt)$는 원본 얼굴 이미지를 이미지 분류 인공신경망에 통과시켜서 얻은 q번째 특징맵의 총 픽셀 수이다. $\|\sim\sim\|_1$는 L1 거리에 대한 손실(loss)로서, 모든 픽셀 차이의 절대값의 합이다. I_{comp}는 지워진 영역을 갖는 원본 얼굴 이미지와 생성적 신경망(162)에 의해 생성된 새로운 얼굴 이미지(I_{gen})에서 상기 지워진 영역에 대응하는 부분만을 추출한 이미지를 합성한 합성 이미지이다. 즉, I_{gen}는 생성적 신경망(162)의 예측 및/또는 추론 과정을 통해 생성된 이미지를 의미하는 것인 반면, I_{comp}는 예측 및/또는 추론 과정이 아니라 지워진 영역을 갖는 원본 이미지와 생성적 신경망(162)에 의해 생성된 이미지(I_{gen})로부터 추출된 이미지(상기 지워진 영역에 대응하는 이미지)를 합성하는 합성 과정을 통해 생성된 이미지를 의미한다. I_{comp}는 3차원 벡터 공간에서 표현되는 3차원 벡터 값일 수 있다.

[0111] 수학식 1에서 $Lstyle(I_{gen})$은 아래의 수학식 5와 같다.

수학식 5

$$L_{sytle}(I) = \sum_{q} \frac{1}{C_q C_q} \left\| \frac{(G_q(I) - G_q(I_{gt}))}{N_q} \right\|_1$$

[0112]

[0113] L_{style}은 대량의 얼굴 데이터 셋에 대하여 이미지 분류(classification)를 위하여 학습된 네트워크의 q번째 layer

특징에 대한 손실(loss)이다. 이러한 L_{style}을 계산하기 위해 Gram matrix가 활용될 수 있다. C_q는 q번째 레이어의 채널 수이고, N_q는 q번째 레이어의 전체 픽셀 수이고, G_q는 q번째 레이어의 Gram matrix 값을 나타낸다.

[0115] 추가로, 생성적 신경망(162)을 학습시키데 분산 손실(variance loss)이 이용된다. 분산 손실(variance loss)은 판별적 신경망(164)이 생성적 신경망(162)로부터 입력된 새로운 얼굴 이미지 내의 모드 픽셀을 1 pixel만큼 강제로 이동시켜서 획득한 새로운 얼굴 이미지와 원본 얼굴 이미지 간의 차이를 나타내는 손실(loss)이다. 이러한 분산 손실(variance loss)은 아래의 수학식 6과 같다.

수학식 6

[0116]
$$L_{tv-col} = \sum_{(i,j)\in R} \frac{\left\| I_{comp}^{i,j+1} - I_{comp}^{i,j} \right\|_1}{N_{comp}}$$

$$L_{tv-row} = \sum_{(i,j)\in R} \frac{\left\| I_{comp}^{i+1,j} - I_{comp}^{i,j} \right\|_1}{N_{comp}}$$

[0117] 분산 손실(variance loss)은 L_{tv-col}, L_{tv-row}를 포함하고, 위의 수학식 6에서 아래 첨자 comp는 원본 얼굴 이미지에서 지워진 영역이 사용자 입력에 따른 이미지로 채워진 이미지를 의미한다. N_{comp}는 생성적 신경망(162)이 생성한 새로운 얼굴 이미지의 전체 픽셀 수이다.

[0119] 생성적 신경망(162)은 위와 같은 분산 손실(variance loss)을 이용하여 블러링에 대하여 강인하도록 학습된다.

[0120] 마지막으로 판별적 신경망(164)이 훈련 데이터에 수렴하지 못하도록 gradient penalty term(L_{GP})이 아래와 같이 추가된다. 여기서, 판별적 신경망(164)이 훈련 데이터에 수렴하지 못하게 한다는 의미는 판별적 신경망(164)이 생성적 신경망(162)보다 더 높은 레벨로 학습되어서는 안됨을 의미하는 것이다. 즉, 생성적 신경망(162)과 판별적 신경망(164)은 서로 적대적인 관계를 유지하도록 동일한 레벨로 학습되어야 함을 의미한다.

수학식 7

[0122]
$$L_{GP} = \mathbb{E}\left[\left(\left\| \nabla_U D(U) \odot M \right\|_2 - 1 \right)^2 \right]$$

[0124] 이상 설명한 바와 같이, 파라미터 L_G와 파라미터 L_D를 감소시키는 방향으로 생성적 신경망(162)과 판별적 신경망(164)을 각각 학습시키는 과정이 완료되면, 이후, 이미지 생성부(130)는 생성적 신경망(162)만을 이용하여 새로운 얼굴 이미지를 생성한다.

[0125] 생성적 신경망(162)은 크기가 가볍기 때문에, 새로운 얼굴 이미지를 생성하는데 걸리는 시간은 일반적인 CPU를 기준으로 2초 이내이다.

[0127] 도 9는 본 발명의 실시 예에 따른 이미지 수정 방법을 보여주는 흐름도이다.

[0128] 도 9를 참조하면, S910에서, 생성적 적대 신경망(160)을 학습시키는 학습 과정이 수행된다. 이 학습 과정은 서로 적대적인(Adversarial) 관계를 유지하도록 생성적 신경망과 판별적 신경망을 학습시키는 과정이다. 아래에서 설명하겠지만, 생성적 신경망(162)과 판별적 신경망(164)의 학습이 완료되면, 생성적 신경망만을 이용하여 원본 이미지에 합성될 이미지가 예측되고, 예측된 이미지와 원본 이미지가 합성된다. 생성적 신경망(162)을 학습시키기 위해, 사전에 수집된 대량의 훈련 데이터가 이용된다. 대량의 훈련 데이터는 훈련용 원본 이미지들, 상기 지워진 영역만을 포함하는 훈련용 마스크 이미지들, 상기 지워진 영역에 다양한 모양이 스케치된 훈련용 스케치 이미지들, 상기 지워진 영역에 다양한 색깔이 채워진 훈련용 색깔 이미지들을 포함한다. 추가로, 상기 지워진 영역의 노이즈를 나타내는 노이즈 이미지가 더 포함될 수 있다. 판별적 신경망(164)과 적대적 관계를 유지하도록, 생성적 신경망(162)을 학습시키는데, 상기 판별적 신경망(164)으로부터 출력되는 손실(L_G)이 이용될 수 있다. 생성적 신경망(162)을 학습시키는 과정은 상기 손실(L_G)을 줄이는 방향으로 학습시키는 과정이다. 이때, 상기 손실(L_G)은, 상기 대량의 훈련 데이터에 포함된 훈련용 마스크 이미지 내에서 지워진 영역의 픽셀값(M)과 상기 원본 이미지와 상기 새로운 이미지 사이의 픽셀 차이값($I_{gen}-I_{gt}$)을 포함한다. 이에 대한 설명은 전술한 수학식 1 및 3에 대한 설명으로 대신한다. 또한 생성적 신경망(162)을 학습시키는 과정은 상기 대량의 훈련 데이터에 포함된 훈련용 마스크 이미지 내에서 지워진 영역의 픽셀값(M)과 상기 원본 이미지와 상기 새로운 이미지 사이의 픽셀 차이값($I_{gen}-I_{gt}$)을 이용하여, 상기 생성적 신경망에 포함된 게이트 컨볼루션 레이어(gated convolution layer)를 학습시키는 과정일 수 있다.

[0129] 이어, S920에서, 생성적 적대 신경망(160)에 대한 학습 과정이 완료되면, 즉, 생성적 신경망(162)과 판별적 신경망(164)의 학습이 완료되면, 학습이 완료된 생성적 신경망(162)을 저장부(150)에 저장하는 과정이 수행된다. 이렇게 함으로써, 이미지 생성부(130)는 생성적 신경망(162)의 접근이 가능하다.

[0130] 이어, S930에서, 원본 이미지(10)에 대한 전처리 과정이 수행되고, 이러한 전처리 과정을 통해 원본 이미지(10)로부터 상기 학습이 완료된 생성적 신경망(162)으로 입력되는 정보가 생성된다. 생성적 신경망(162)으로 입력되는 정보는 상기 원본 이미지 내에서 사용자 입력에 의해 지워진 영역만을 포함하는 마스크(mask) 이미지, 상기 사용자 입력에 의해 상기 지워진 영역에 스케치된 모양만을 포함하는 스케치 이미지 및 상기 사용자 입력에 의해 상기 지워진 영역에 칠해진 색깔만을 포함하는 색깔 이미지를 포함한다. 이때, 생성적 신경망(162)으로 입력되는 정보는 상기 원본 이미지 내에서 사용자 입력에 의해 지워진 영역만을 포함하는 마스크(mask) 이미지로만 구성될 수도 있다. 이처럼 본 발명에서는 사용자에 의해 지워진 영역의 위치 정보만을 포함하고 있는 마스크 이미지만을 생성적 신경망(162)에 입력하는 것만으로도 원본 이미지의 일부를 사용자 의도에 맞게 수정할 수 있다.

[0131] 이어, S940에서, 이미지 생성부(130)가 학습이 완료된 상기 생성적 적대 신경망, 즉, 학습이 완료된 생성적 신경망(162)을 이용하여 상기 마스크(mask) 이미지, 스케치 이미지 및 색깔 이미지를 분석하여, 그 분석 결과에 따라 상기 원본 이미지 내에서 지워진 영역에 합성될 이미지를 예측하는 과정이 수행된다.

[0132] 이어, S950에서, 이미지 생성부(130)가 상기 예측된 이미지를 상기 원본 이미지 내에서 상기 지워진 영역에 합성하여 새로운 이미지를 생성하고, 표시부(140)가 상기 생성된 새로운 이미지를 표시하여, 사용자에게 제공된다.

[0134] 도 10 내지 12는 본 발명의 이미지 수정 방법에 따라 생성된 합성 이미지들의 예시한 도면들이다.

[0135] 도 10의 (A) 및 (B)에 도시한 바와 같이, 본 발명에 따르면, 원본 얼굴 이미지에서 입 모양, 눈 형상을 수정하여 새로운 이미지를 자동으로 생성할 수 있다. 또한, 도 10의 (C)에 도시한 바와 같이, 원본 얼굴 이미지에서 안경을 지워서 새로운 이미지를 자동으로 생성할 수 있다. 여기서 중요한 점은 생성적 신경망의 입력 정보(Free-form input)를 생성하기 위해 지워진 영역에 스케치되는 모양은 전문가 수준의 스케치를 필요로 하지 않는 점이다. 즉, 본 발명은 대량의 훈련 데이터를 이용하여 사전에 학습시킨 생성적 신경망을 이용하기 때문에, 마스크 이미지로부터 지워진 영역의 위치 정보와 일반인 수준으로 스케치된 모양만으로 원본 얼굴 이미지에 합성될 이미지의 추론과 예측이 가능하다. 이것은 마스크 이미지, 스케치 이미지 또는 색깔 이미지와 같은 단순한 이미지 입력만을 이용하여 원본 얼굴 이미지를 사실적으로 수정할 수 있음을 의미한다.

[0136] 또한, 본 발명에 따르면, 도 11의 (A)에 도시된 바와 같이, 헤어 스타일의 일부를 수정하거나 (B)에 도시된 바와 같이, 헤어 스타일의 전부를 수정할 수 있다.

[0137] 또한, 본 발명에 따르면, 도 12의 (A)에 도시된 바와 같이, 원본 얼굴 이미지에서 귀 영역의 일부 영역을 지우고, 그 지운 영역에 사용자가 간단하게 스케치한 귀걸이 모양을 생성적 신경망의 입력 정보로 구성하면, 원본 얼굴 이미지에 귀걸이 이미지가 사실적으로 합성된 이미지를 자동으로 생성할 수 있다.

[0138] 또한, 도 12의 (B)에 도시된 바와 같이, 원본 얼굴 이미지에 포함된 귀걸이를 다른 형상의 귀걸이로 수정하는 것도 가능할 것이다.

[0140] 이상에서 설명된 이미지 수정 시스템은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다.

[0141] 예를 들어, 실시 예들에서 설명된 전처리부(120), 이미지 생성부(130), 학습부(160)와 같은 구성요소들은, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 그래픽 프로세서(GPU), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서로 구현될 수 있다.

[0142] 또한, 이미지 수정 시스템은 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 이미지 수정 시스템은 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다.

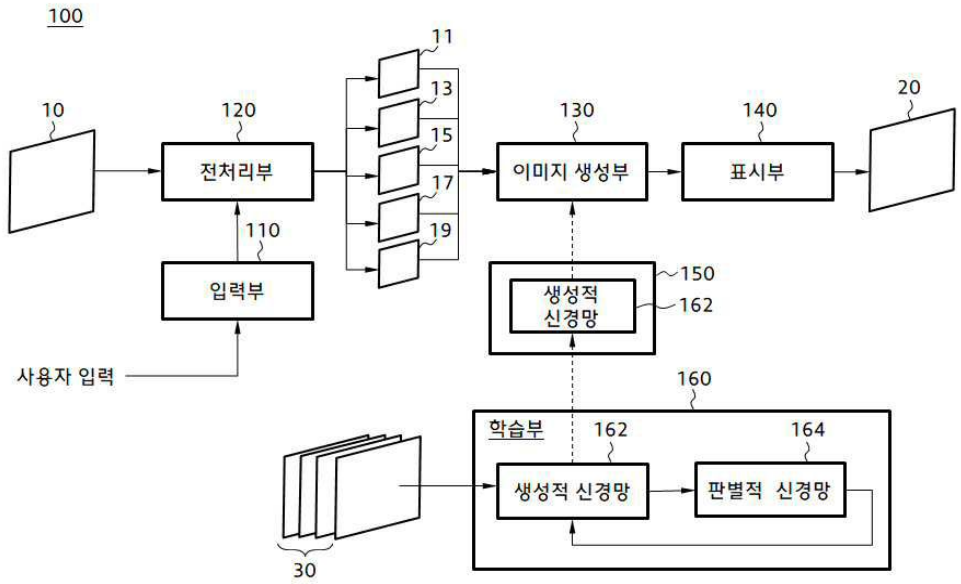
[0143] 실시 예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0144] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

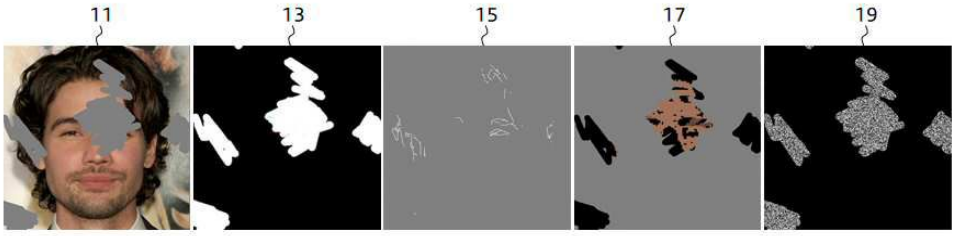
[0145] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

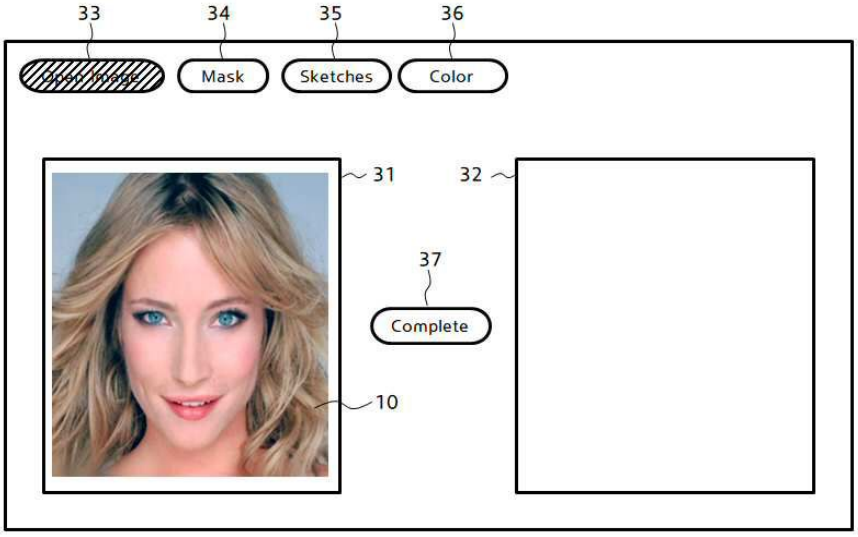
도면1



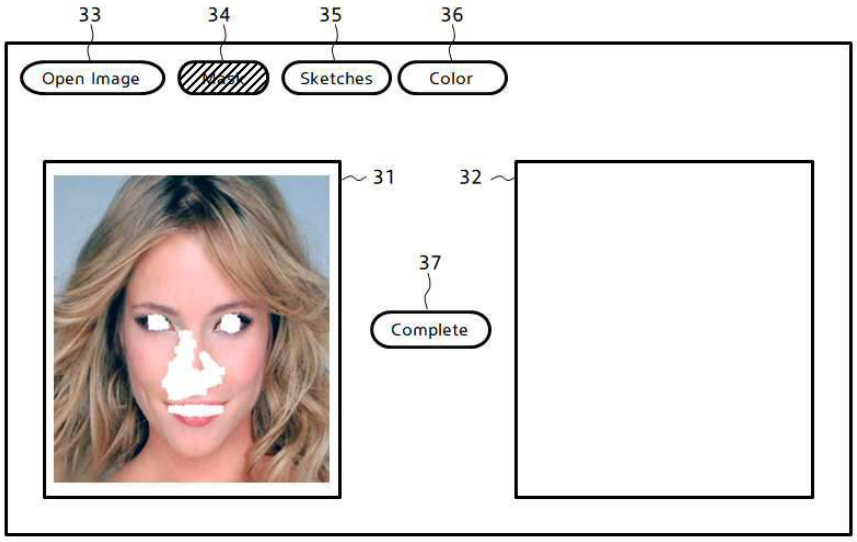
도면2



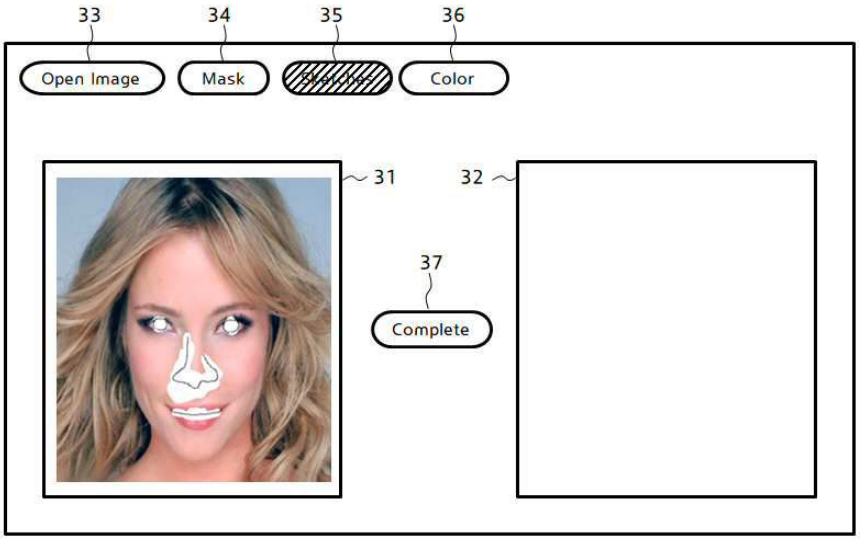
도면3



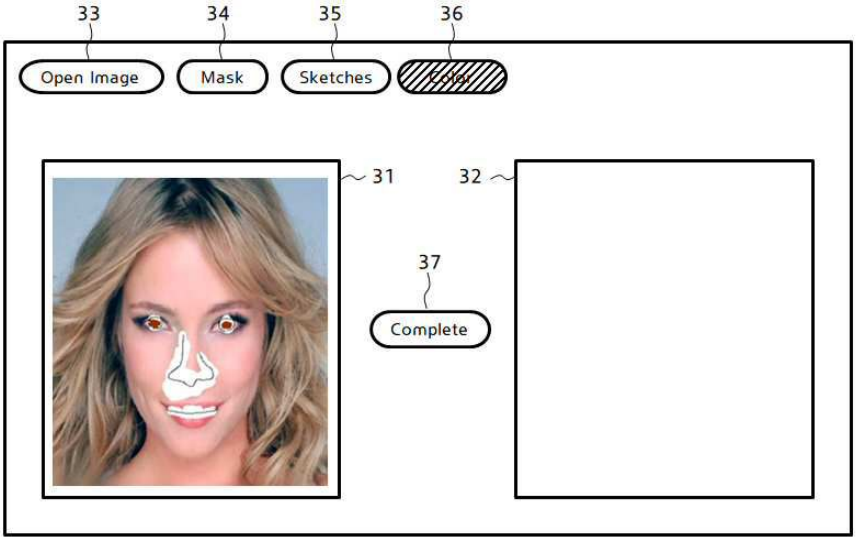
도면4


33
34
35
36
Open Image
Mask
Sketches
Color
31
32
37
Complete

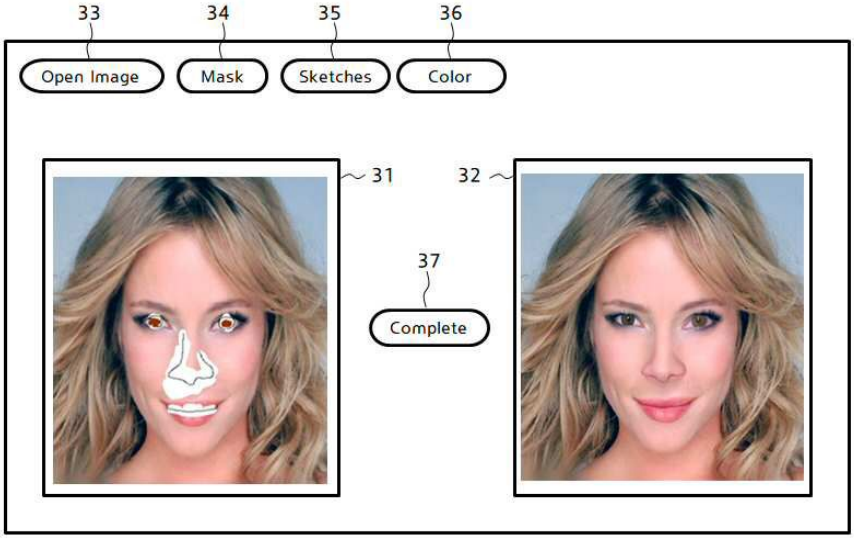
도면5


33
34
35
36
Open Image
Mask
Sketches
Color
31
32
37
Complete

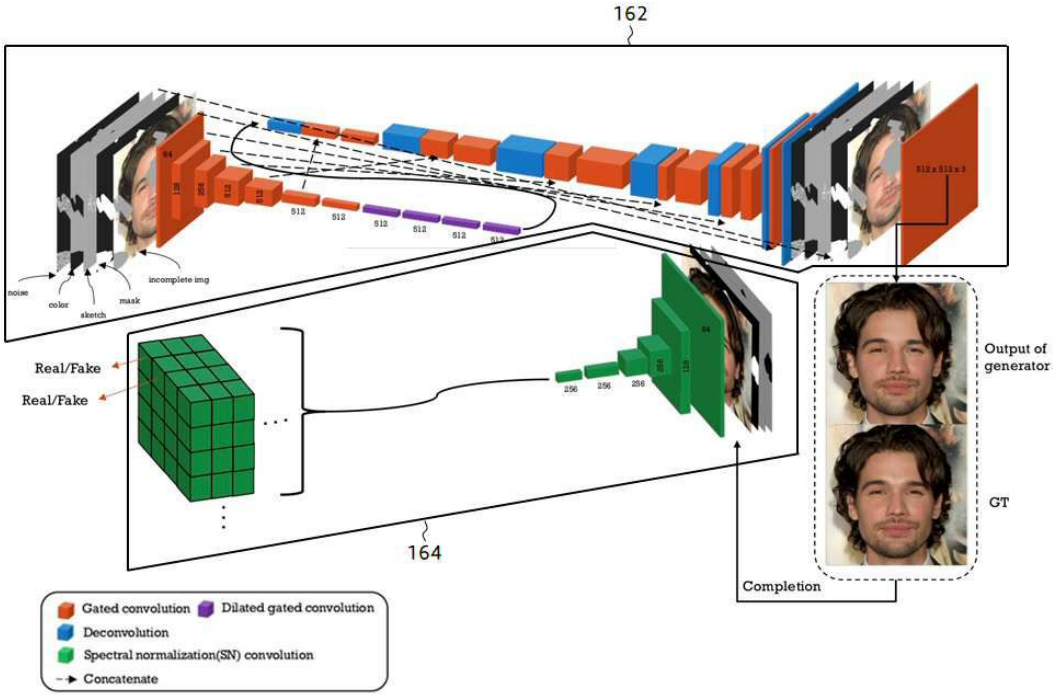
도면6


33
34
35
36
Open Image
Mask
Sketches
Color
31
32
37
Complete

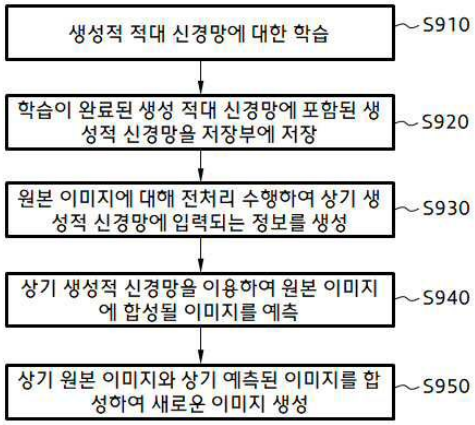
도면7


33
34
35
36
Open Image
Mask
Sketches
Color
31
32
37
Complete

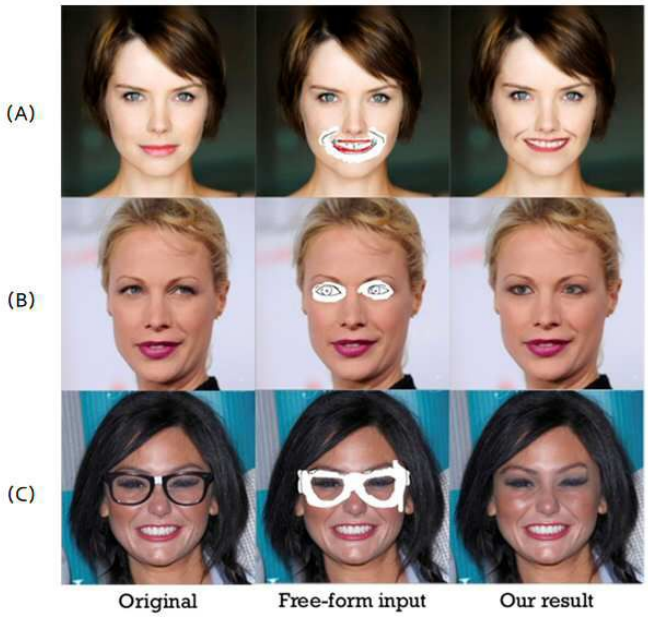
도면8


162
noise
color
sketch
mask
incomplete img
Real/Fake
Real/Fake
164
Output of generator
GT
Completion
Gated convolution
Dilated gated convolution
Deconvolution
Spectral normalization(SN) convolution
Concatenate

도면9

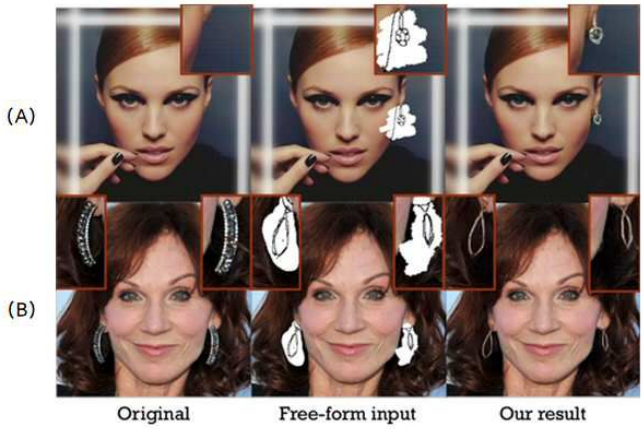

생성적 적대 신경망에 대한 학습
S910
학습이 완료된 생성 적대 신경망에 포함된 생성적 신경망을 저장부에 저장
S920
원본 이미지에 대해 전처리 수행하여 상기 생성적 신경망에 입력되는 정보를 생성
S930
상기 생성적 신경망을 이용하여 원본 이미지에 합성될 이미지를 예측
S940
상기 원본 이미지와 상기 예측된 이미지를 합성하여 새로운 이미지 생성
S950

도면10

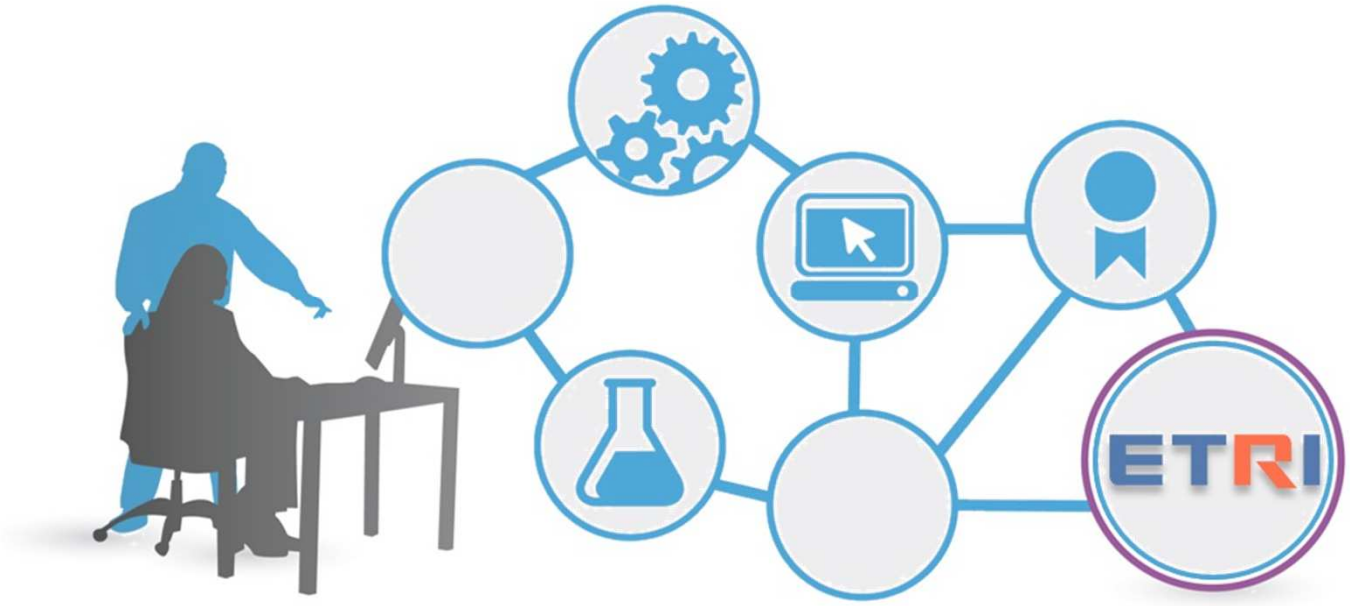


도면11

도면12



딥러닝 기반 영상 메타데이터 생성 기술



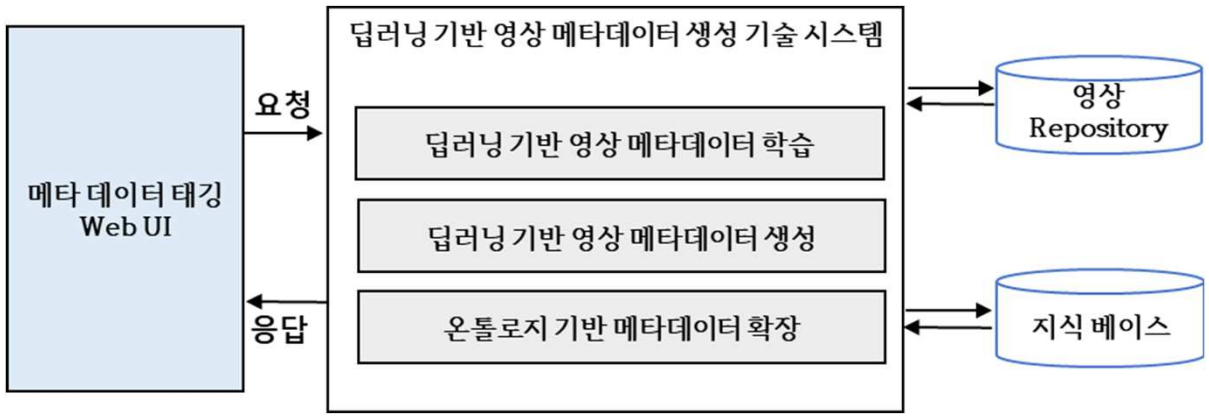
목 차

1. 기술의 개요
2. 기술이전 내용 및 범위
3. 기술 현황
4. 기술의 사업성
5. 국내외 시장 동향

1. 기술의 개요

딥러닝 기반 영상 메타데이터 생성 기술

- 본 기술은 사용자가 업로드한 영상에 대해, 시스템이 해당 영상을 분석하여 임베딩 벡터 기반의 메타데이터를 생성하는데 목적이 있음.
- 본 기술은 영상에서 특징 정보 (객체, 행위, 장소, 시간)를 추출하여 영상 메타데이터 생성 모델을 학습하는 기능을 제공함.
- 본 기술은 학습된 모델을 기반으로 임베딩 벡터 기반의 메타데이터와 영상을 설명하는 지문(영상 description)을 생성하는 기능을 제공함.
- 본 기술은 온톨로지 기반으로 크라우드 소싱을 활용하여 메타데이터 정보를 확장할 수 있는 기능을 제공함.

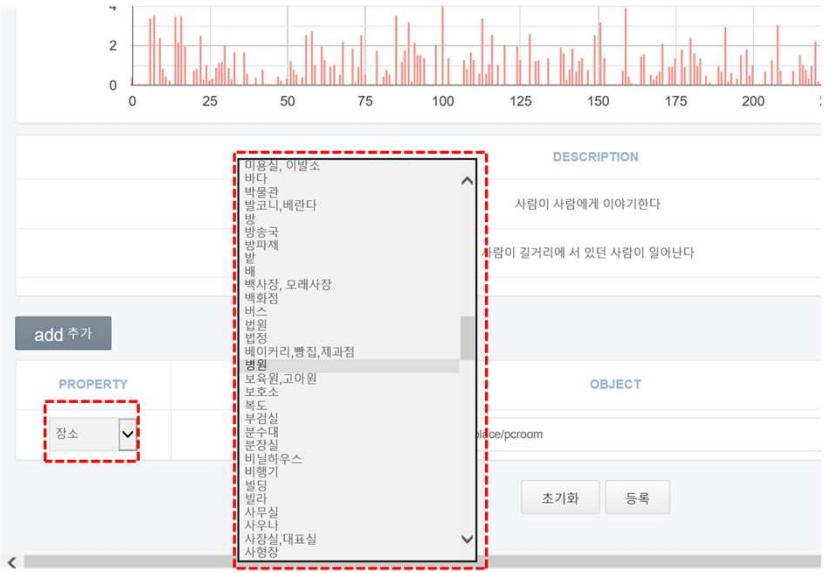


딥러닝 기반 영상 메타데이터 생성 기술 구조도

2. 기술이전 내용 및 범위 (1/3)

기술이전 내용

- 딥러닝 기반 영상 메타데이터 생성 모델 학습 기능
- 딥러닝 기반 영상 메타데이터 생성 기능
- 도메인 온톨로지 기반 영상 메타데이터 확장 기능
- 도메인 지식 구축 기능
- 도메인 지식 제공 기능



2. 기술이전 내용 및 범위 (2/3)

기술이전 범위

- 딥러닝 기반 영상 메타데이터 생성 기술 시스템 요구사항 정의서 1종
- 딥러닝 기반 영상 메타데이터 생성 기술 상세설계서 1종
- 딥러닝 기반 영상 메타데이터 생성 기술 시험절차 및 결과서 1종
- 딥러닝 기반 영상 메타데이터 생성 기술 프로그램 3종
  - 딥러닝 기반 영상 메타데이터 학습기 1.0
  - 딥러닝 기반 영상 메타데이터 생성기 1.0
  - 도메인 지식 기반 장면 지식 생성 1.0

2. 기술이전 내용 및 범위 (3/3)

기술 개발 현황

기술성숙도(TRL : Technology Readiness Level) :6단계

구 분	단계	정 의	세 부 설 명
기초 연구 단계	1	기초 이론/실험	◦ 기초이론 정립 단계
	2	실용목적의 아이디어, 특허 등 개념정립	◦ 기술개발개념정립및아이디어에대한특허출원단계
실험 단계	3	실험실 규모의 기본성능 검증	◦ 실험실 환경에서 실험 또는 전산 시뮬레이션을 통해 기본성능이 검증될 수 있는 단계 ◦ 개발하려는부품/시스템의기본설계도면을확보하는단계
	4	실험실 규모의 소재/부품/시스템 핵심성능 평가	◦ 시험샘플을제작하여핵심성능에대한평가가완료된단계 ◦ 3단계에서 도출된 다양한 결과 중에서 최적의 결과를 선택하려는 단계 ◦ 컴퓨터 모사가 가능한 경우 최적화를 완료하는 단계
시작품 단계	5	확정된 소재/부품/ 시스템시작품제작 및 성능 평가	◦ 확정된 소재/부품/시스템의 실험실 시작품 제작 및 성능 평가가 완료된 단계 ◦ 개발 대상의 생산을 고려하여 설계하나 실제 제작한 시작품 샘플은 1~수개 미만인 단계 ◦ 경제성을 고려하지 않고 기술의 핵심성능으로만 볼 때, 실제로 판매가 될 수 있는 정도로 목표 성능을 달성한 단계
	6	파일롯 규모 시작품 제작 및 성능 평가	◦ 파일롯 규모(복수 개~양산규모의 1/10정도)의 시작품 제작 및 평가가 완료된 단계 ◦ 파일롯규모생산품에대하생산량생산용량불량률등제시 ◦ 파일롯 생산을 위한 대규모 투자가 동반되는 단계 ◦ 생산기업이 수요기업 적용환경에 유사하게 자체 현장테스트를 실시하여 목표 성능을 만족시킨 단계 ◦ 성능평가결과에대하가능하면공인인증기관의성적서확보
실용화 단계	7	신뢰성평가 및 수요기업 평가	◦ 실제 환경에서 성능 검증이 이루어지는 단계 ◦ 부품 및 소재개발의 경우 수요업체에서 직접 파일롯 시작품을 현장 평가(성능 및 신뢰성 평가) ◦ 가능하면 인증기관의 신뢰성 평가 결과 제출
	8	시제품 인증 및 표준화	◦ 표준화 및 인허가 취득 단계
사업화	9	사업화	◦ 본격적인 양산 및 사업화 단계 ◦ 6-시그마 등 품질관리가 중요한 단계

3. 기술 동향

기술 현황

국외 기술현황

- 미국의 Clarifai는 Deep learning 기반의 이미지 태깅 기술을 기반으로, 영상 콘텐츠에서 추출된 각 프레임에 대한 태깅 툴 및 API를 제공하고 있음.

  - Clarifai의 태깅 기술은 Convolution Neural Network를 기반으로 구현되었으며, 개, 의자, 책상, 다리 등의 객체부터 도시, 아침, 밤, 즐거움 등 추상적인 개념에 이르기까지 수 천개의 컨셉에 대한 학습 결과를 바탕으로 이미지에 대한 태깅을 제공함.

국내 기술현황

- KBS는 영상에 대해 세그멘테이션을 수행한 후, 수동 태깅/인덱싱 관리 저작도구를 개발하였음.
- 한밭대 김수경 교수팀은 동영상 콘텐츠의 장면 검색을 위한 장면에 내포된 개체(인물, 사건, 사물 등)의 메타체계를 온톨로지로 구성하는 연구를 수행한바 있음.
- 영화 메타데이터의 경우, 국내에서는 현재 메타데이터 정보 체계를 확립하는 수준이며 영상 검색 등 상용화 서비스에는 적용된 바 없음.

4. 기술의 사업성 (1/2)

예상 응용 제품 및 서비스

예상 제품/서비스	예상 수요자(층)
VOD 클립 서비스	VOD 콘텐츠 사업자, 포탈 사업자 등
자동 캡션 배포 서비스	방송 사업자
e-Learning 서비스	인터넷 교육 서비스 사업자

사업성

예상 제품 /서비스	예상단가 (천원)	이전기술의 비중(%)	잠재적/현재적 경쟁자와 가격,시장 등에서 경쟁상 유리한 점	판매 가능 시기
딥러닝 기반 영상 메타데이터 생성 서비스	-	50%	a. 가격경쟁력면: 신규 서비스로 가격형성 b. 시장환경면: 2021년 이후 연간 10억 예상	2021년

4. 기술의 사업성 (2/2)

상용화까지 단계별 주요 일정

- 1단계: 기술 이해 및 서비스를 위한 서버 구축 (약 6개월 소요 추정)
- 2단계: 콘텐츠 확보 및 메타데이터 생성 (초기 6개월 소요 및 지속적 유지/보수
- 3단계: 상용화 (약 1년 후 상용화 가능 추정)

상용화를 위한 추가비용

- 영상 콘텐츠에 대한 영상 메타데이터 학습 및 생성 시 콘텐츠 확보 및 DB 관리에 따른 추가비용이 필요함.
- 자체 콘텐츠 적용에서의 성능 향상을 위해서는 학습 데이터의 추가적인 확보가 필요함.
- 영상 메타데이터 학습 및 생성 작업을 지원하기 위한 서버 구축 비용이 발생함.
- 도메인 지식 관리 및 저장을 위한 온톨로지 저장소 구입 및 수동 메타데이터 생성을 위한 인력이 필요할 수 있음.

5. 국내외 시장 동향

관련 제품/서비스 국내외 시장 규모 (추정)

관련 제품 /서비스	시장	2021	2022	2023	2024	2025
딥러닝 기반 영상 메타데이터 생성 서비스	해외 (백만불)	20	30	50	75	110
	국내 (억원)	10	15	25	40	60

예상 제품/서비스의 예상 매출액 (추정)

관련 제품 /서비스	시장	2021	2022	2023	2024	2025
딥러닝 기반 영상 메타데이터 생성 서비스	해외 (백만불)	사업준비	0.6	1	2.25	4.4
	국내 (억원)	1	3	5	12	21

감사합니다

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0122496
(43) 공개일자 2021년10월12일

(51) 국제특허분류(Int. Cl.)
H04N 21/435 (2011.01) *G06N 20/00* (2019.01)
H04N 21/43 (2011.01) *H04N 21/854* (2011.01)
(52) CPC특허분류
H04N 21/435 (2013.01)
G06N 20/00 (2021.08)
(21) 출원번호 10-2020-0039705
(22) 출원일자 2020년04월01일
심사청구일자 없음

(71) 출원인
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
곽창욱
대전광역시 유성구 가정로 270, B동 907호
김선중
세종특별자치시 남세종로 357, 103동 1001호
(뒷면에 계속)
(74) 대리인
한양특허법인

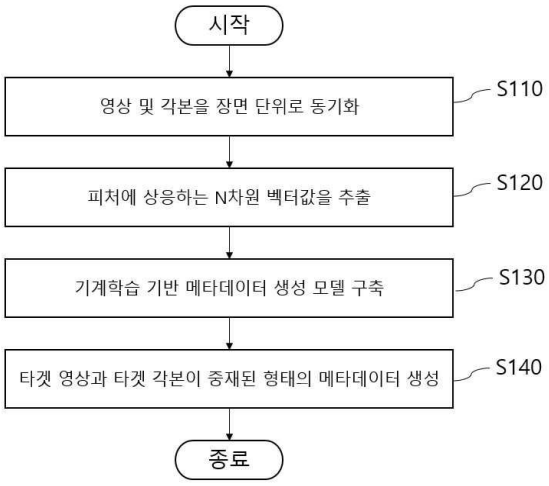
전체 청구항 수 : 총 1 항

(54) 발명의 명칭 **벡터를 이용한 장면 묘사 기반의 메타데이터 생성 방법 및 이를 위한 장치**

(57) 요 약

벡터를 이용한 장면 묘사 기반의 메타데이터 생성 방법 및 이를 위한 장치가 개시된다. 본 발명의 일실시예에 따른 메타데이터 생성 방법은 입력 영상에서 추출된 적어도 하나의 특징을 기반으로 상기 입력 영상을 입력 각본과 장면 단위로 동기화하고, 동기화된 장면 단위의 영상 및 각본 각각에 대해 상기 적어도 하나의 특징에 상응하는 N차원의 벡터값을 추출하고, 상기 N차원의 벡터값을 기반으로 영상 잠재 벡터 모델 및 각본 잠재 벡터 모델에 대한 기계학습을 수행하여 메타데이터 생성 모델을 구축하고, 타겟 영상 및 타겟 각본 중 어느 하나만 입력된 경우, 상기 메타데이터 생성 모델을 기반으로 상기 타겟 영상과 상기 타겟 각본이 중재된 형태의 메타데이터를 생성한다.

대 표 도 - 도1



(52) CPC특허분류
H04N 21/4302 (2020.08)
H04N 21/854 (2013.01)
(72) 발명자
손정우
대전광역시 유성구 봉명로 93, 608동 105호
이호재
대전광역시 유성구 지족로 362, 307동 402호
한민호
대전광역시 유성구 엑스포로123번길 46-15, 202동 303호
함경준
대전광역시 유성구 어은로 57, 120동 105호
김상권
대전광역시 유성구 어은로 57, 133동 505호
박종현
대전광역시 유성구 대덕대로 617, 101동 501호

이 발명을 지원한 국가연구개발사업

과제고유번호	1711101947
과제번호	20ZH1200
부처명	과학기술정보통신부
과제관리(전문)기관명	한국전자통신연구원
연구사업명	한국전자통신연구원연구운영비지원(R&D)(주요사업비)
연구과제명	오픈시나리오 기반 프로그래머블 인터랙티브 미디어 창작 서비스 플랫폼 개발
기 여 율	1/1
과제수행기관명	한국전자통신연구원
연구기간	2020.01.01 ~ 2020.12.31

명 세 서

청구범위

청구항 1

입력 영상에서 추출된 적어도 하나의 특징을 기반으로 상기 입력 영상을 입력 각본과 장면 단위로 동기화하는 단계;

동기화된 장면 단위의 영상 및 각본 각각에 대해 상기 적어도 하나의 특징에 상응하는 N차원의 벡터값을 추출하는 단계;

상기 N차원의 벡터값을 기반으로 영상 잠재 벡터 모델 및 각본 잠재 벡터 모델에 대한 기계학습을 수행하여 메타데이터 생성 모델을 구축하는 단계; 및

타겟 영상 및 타겟 각본 중 어느 하나만 입력된 경우, 상기 메타데이터 생성 모델을 기반으로 상기 타겟 영상과 상기 타겟 각본이 중재된 형태의 메타데이터를 생성하는 단계

를 포함하는 것을 특징으로 하는 메타데이터 생성 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 메타데이터를 자동으로 생성하는 기술에 관한 것으로, 특히 영상을 기반으로 장면을 묘사하는 정보를 추출하여 의미적 메타데이터를 생성하는 기술에 관한 것이다.

배 경 기 술

[0002] 방송국이 주도했던 영상 콘텐츠 시장은 온라인의 영상 플랫폼으로 중심이 이동되고 있다. 과거 텔레비전을 통해 수동적인 소비가 이루어졌다면, 최근에는 기술의 진보와 변화에 따라 생산, 유통, 소비의 방법이 다양해졌다.

[0003] 초기 영상 콘텐츠 스트리밍 서비스는 1시간 ~ 2시간 길이의 콘텐츠 전체를 활용했다. 하지만 최근 주요 OTT 업체에서는 3~4분 내외의 장면 클립으로 영상을 분할하여 온라인 플랫폼에서 서비스를 하고 있다. 이러한 클립 단위의 서비스는 빠르고 짧은 것을 선호하는 젊은 층의 소비 성향과 일치하기도 하지만, 주요 수익수단이 미디어 콘텐츠 재생 사이에 노출되는 광고라는 점을 볼 때, 온라인 콘텐츠 서비스에도 최적화된 단위이다. 영상의 단위가 축소되면서 짜깁기, 재편집과 같은 기존 영상을 재활용하는 산업적 시도도 발생하고 있다.

[0004] 이처럼 서비스에 활용되는 영상의 단위가 장면으로 축소되고, 기존 영상의 재활용이 확대됨에 따라 정확한 정보가 포함된 영상을 검색하는 것이 중요해졌다. 기존에는 콘텐츠 제목, 배우와 같은 큰 규모의 영상을 대상으로 사용자의 검색이 이루어졌다면, 장면 단위에서는 영상의 장소, 시간, 등장인물, 행동, 객체 등과 같은 다양한 정보들의 검색 요구가 발생하고 있다.

[0005] 현재, 영상 서비스 업체에서는 해시태그를 기반으로 주요한 정보들을 5개 내외의 키워드로 메타데이터를 태깅하고 이를 활용한 영상 검색이 이루어진다. 이러한 정보들은 사람들이 직접 태깅하고 있다는 점에서 시간적, 물리적 비용이 크기 때문에, 자동화된 영상 분석 및 태깅 시스템이 필요한 실정이다. 뿐만 아니라, 생성되는 키워드의 기준이 불명확하기 때문에 누락되는 정보들이 대부분이다. 따라서, 영상에 포함된 다양한 정보들을 메타데이터로 생성할 수 있는 시스템이 필요하다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국 공개 특허 제10-2018-0087969호, 2018년 8월 3일 공개(명칭: 동영상 장면과 메타데이터 저작 방법)

발명의 내용

해결하려는 과제

[0007] 본 발명의 목적은 영상의 의미적 정보들을 효과적으로 벡터 형식으로 압축 표현함으로써 영상에 포함된 내용들을 효과적으로 나타낼 수 있는 의미적 메타데이터를 생성하여 제공하는 것이다.

[0008] 또한, 본 발명의 목적은 영상이나 각본 중 어느 하나의 리소스의 부재에도 두 리소스 전체를 반영한 메타데이터를 생성하는 것이다.

[0009] 또한, 본 발명의 목적은 벡터 기반의 메타데이터를 제공함으로써 벡터 사이의 유사도 계산을 통해 보다 효율적인 영상 검색 서비스를 제공하는 것이다.

과제의 해결 수단

[0010] 상기한 목적을 달성하기 위한 본 발명에 따른 메타데이터 생성 방법은 입력 영상에서 추출된 적어도 하나의 특징을 기반으로 상기 입력 영상을 입력 각본과 장면 단위로 동기화하는 단계; 동기화된 장면 단위의 영상 및 각본 각각에 대해 상기 적어도 하나의 특징에 상응하는 N차원의 벡터값을 추출하는 단계; 상기 N차원의 벡터값을 기반으로 영상 잠재 벡터 모델 및 각본 잠재 벡터 모델에 대한 기계학습을 수행하여 메타데이터 생성 모델을 구축하는 단계; 및 타겟 영상 및 타겟 각본 중 어느 하나만 입력된 경우, 상기 메타데이터 생성 모델을 기반으로 상기 타겟 영상과 상기 타겟 각본이 중재된 형태의 메타데이터를 생성하는 단계를 포함한다.

[0011] 이 때, 영상 잠재 벡터 모델은 상기 장면 단위의 영상에 대한 메타데이터를 생성하고, 상기 각본 잠재 벡터 모델은 상기 장면 단위의 각본에 대한 메타데이터를 생성할 수 있다.

[0012] 이 때, 구축하는 단계는 손실 함수를 기반으로 상기 영상 잠재 벡터 모델 및 상기 각본 잠재 벡터 모델 각각에 포함된 인코더 및 디코더를 상기 영상 잠재 벡터 모델과 상기 각본 잠재 벡터 모델 사이의 차이가 감소하도록 학습시키는 단계를 포함할 수 있다.

[0013] 이 때, 학습시키는 단계는 상기 영상 잠재 벡터 모델 및 상기 각본 잠재 벡터 모델 각각에 상응하는 적어도 하나의 입력 벡터 및 출력 벡터를 획득하는 단계; 상기 영상 잠재 벡터 모델과 상기 각본 잠재 벡터 모델 간 상기 적어도 하나의 입력 벡터 및 출력 벡터의 오차가 감소하도록 상기 인코더 및 디코더의 변수를 조정하는 단계를 포함할 수 있다.

[0014] 이 때, 메타데이터는 실수형의 벡터 형식에 상응할 수 있다.

[0015] 이 때, 동기화하는 단계는 상기 적어도 하나의 특징을 기반으로 상기 입력 영상을 샷 단위로 분할하는 단계; 및 상기 샷 단위로 분할된 복수개의 샷 영상들을 의미적 연속성을 갖는 장면 단위로 병합하는 단계를 포함할 수 있다.

[0016] 또한, 본 발명의 일실시예에 따른 메타데이터 생성 장치는, 입력 영상에서 추출된 적어도 하나의 특징을 기반으로 상기 입력 영상을 입력 각본과 장면 단위로 동기화하고, 동기화된 장면 단위의 영상 및 각본 각각에 대해 상기 적어도 하나의 특징에 상응하는 N차원의 벡터값을 추출하고, 상기 N차원의 벡터값을 기반으로 영상 잠재 벡터 모델 및 각본 잠재 벡터 모델에 대한 기계학습을 수행하여 메타데이터 생성 모델을 구축하고, 타겟 영상 및 타겟 각본 중 어느 하나만 입력된 경우, 상기 메타데이터 생성 모델을 기반으로 상기 타겟 영상과 상기 타겟 각본이 중재된 형태의 메타데이터를 생성하는 프로세서; 및 상기 메타데이터 생성 모델을 저장하는 메모리를 포함한다.

[0017] 이 때, 영상 잠재 벡터 모델은 상기 장면 단위의 영상에 대한 메타데이터를 생성하고, 상기 각본 잠재 벡터 모델은 상기 장면 단위의 각본에 대한 메타데이터를 생성할 수 있다.

[0018] 이 때, 프로세서는 손실 함수를 기반으로 상기 영상 잠재 벡터 모델 및 상기 각본 잠재 벡터 모델 각각에 포함된 인코더 및 디코더를 상기 영상 잠재 벡터 모델과 상기 각본 잠재 벡터 모델 사이의 차이가 감소하도록 학습시킬 수 있다.

[0019] 이 때, 프로세서는 상기 영상 잠재 벡터 모델 및 상기 각본 잠재 벡터 모델 각각에 상응하는 적어도 하나의 입력 벡터 및 출력 벡터를 획득하고, 상기 영상 잠재 벡터 모델과 상기 각본 잠재 벡터 모델 간 상기 적어도 하나의 입력 벡터 및 출력 벡터의 오차가 감소하도록 상기 인코더 및 디코더의 변수를 조정할 수 있다.

[0020] 이 때, 메타데이터는 실수형의 벡터 형식에 상응할 수 있다.

[0021] 이 때, 프로세서는 상기 적어도 하나의 특징을 기반으로 상기 입력 영상을 샷 단위로 분할하는 단계; 및 상기 샷 단위로 분할된 복수개의 샷 영상들을 의미적 연속성을 갖는 장면 단위로 병합할 수 있다.

발명의 효과

[0022] 본 발명에 따르면, 영상의 의미적 정보들을 효과적으로 벡터 형식으로 압축 표현함으로써 영상에 포함된 내용들을 효과적으로 나타낼 수 있는 의미적 메타데이터를 생성하여 제공할 수 있다.

[0023] 또한, 본 발명은 영상이나 각본 중 어느 하나의 리소스의 부재에도 두 리소스 전체를 반영한 메타데이터를 생성할 수 있다.

[0024] 또한, 본 발명은 벡터 기반의 메타데이터를 제공함으로써 벡터 사이의 유사도 계산을 통해 보다 효율적인 영상 검색 서비스를 제공할 수 있다.

도면의 간단한 설명

[0025] 도 1은 본 발명의 일실시예에 따른 벡터를 이용한 장면 묘사 기반의 메타데이터 생성 방법을 나타낸 동작 흐름도이다.

도 2는 본 발명에 따른 메타데이터 생성 과정을 나타낸 도면이다.

도 3은 본 발명에 따른 N차원 벡터값을 추출하는 과정을 나타낸 도면이다.

도 4는 본 발명에 따른 메타데이터 생성 모델을 구축하는 과정을 나타낸 도면이다.

도 5는 본 발명에 따른 메타데이터 생성 모델 구축 과정을 상세하게 나타낸 동작 흐름도이다.

도 6 내지 도 7은 본 발명에 따른 타겟 영상만으로 메타데이터를 생성하는 과정의 일 예를 나타낸 도면이다.

도 8 내지 도 9는 본 발명에 따른 타겟 각본만으로 메타데이터를 생성하는 과정의 일 예를 나타낸 도면이다.

도 10은 본 발명의 일실시예에 따른 벡터를 이용한 장면 묘사 기반의 메타데이터 장치를 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0026] 본 발명을 첨부된 도면을 참조하여 상세히 설명하면 다음과 같다. 여기서, 반복되는 설명, 본 발명의 요지를 불필요하게 흐릴 수 있는 공지 기능, 및 구성에 대한 상세한 설명은 생략한다. 본 발명의 실시형태는 당 업계에서 평균적인 지식을 가진 자에게 본 발명을 보다 완전하게 설명하기 위해서 제공되는 것이다. 따라서, 도면에서의 요소들의 형상 및 크기 등은 보다 명확한 설명을 위해 과장될 수 있다.

[0027] 이하, 본 발명에 따른 바람직한 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0029] 도 1은 본 발명의 일실시예에 따른 벡터를 이용한 장면 묘사 기반의 메타데이터 생성 방법을 나타낸 동작 흐름도이다.

[0030] 본 발명은 기존의 키워드 기반의 메타데이터 생성 기술에서 진보하여, 영상에서 추출 가능한 여러 특징 정보들이 통합된 벡터 기반의 메타데이터를 생성하는 것이다. 기존의 일반적인 영상 메타데이터 생성 방법들이 영상에 등장하는 키워드들을 태깅하는 것이었다면, 본 발명에서는 영상의 의미적 요소들을 포함한 벡터 기반의 메타데이터를 생성하는 방법을 제공한다. 이를 위해, 기계학습 기반의 영상 분석을 통해 장소, 시간, 행위, 지문 등의 특징 정보들을 추출할 수 있다. 각각의 특징 정보들은 N차원 벡터 형식의 메타데이터로 표현할 수 있으며, 벡터로 변환하는 인코더 및 디코더의 학습을 통해 벡터들은 다시 텍스트로 재현할 수 있다.

[0031] 이러한 개념을 확장하여, 학습을 통해 영상에서 추출한 특징 정보들을 통합해 하나의 벡터로 표현하고, 통합된 벡터를 기반으로 다시 정보들을 재현할 수 있다. 즉, 통합된 벡터는 추출된 특징 정보들이 축약된 것으로써 영상의 정보들을 나타내는 메타데이터로써 사용될 수 있다는 것을 의미한다.

[0032] 본 발명은 크게 메타데이터를 생성하는 메타데이터 생성 모델을 학습하는 단계와 학습된 메타데이터 생성 모델을 기반으로 메타데이터를 생성하는 단계로 구분할 수 있는데, 이하에서는 먼저 메타데이터 생성 모델을 학습하

는 단계를 설명하도록 한다.

[0033] 도 1을 참조하면, 본 발명의 일실시예에 따른 벡터를 이용한 장면 묘사 기반의 메타데이터 생성 방법은 입력 영상에서 추출된 적어도 하나의 특징을 기반으로 입력 영상을 입력 각본과 장면 단위로 동기화한다(S110).

[0034] 예를 들어, 도 2를 참조하면, 입력된 각본(210)에 대해 각본 분할(S210)을 수행하고, 입력된 영상(220)에 대해 특징 분석(S220)후 샷 분할(S230) 및 장면 분할(S240)을 수행한 뒤 각각 분할된 영상과 각본에 대해 동기화를 수행할 수 있다(S250).

[0035] 이 때, 입력 각본은 입력 영상이 제작되기에 앞서 기록된 텍스트 형식의 문서에 상응하는 것으로, 장소, 시간, 장면 등을 설명하는 지문이 기록될 수 있다.

[0036] 이 때, 입력된 영상(220)에 대한 특징 분석(S220)은 메타데이터 생성 장치에 포함된 별도의 영상 특징 분석 모듈을 통해 수행될 수 있다.

[0037] 이 때, 적어도 하나의 특징을 기반으로 입력 영상을 샷 단위로 분할할 수 있다.

[0038] 이 때, 샷 단위로 분할된 복수개의 샷 영상들을 의미적 연속성을 갖는 장면 단위로 병합할 수 있다.

[0039] 예를 들어, 영상에 대한 특징 분석(S220)을 통해 추출된 특징 정보들을 별도의 샷 분할 모듈을 이용하여 1~3초 내외의 샷으로 분할할 수 있다. 이렇게 분할된 샷은 장면 분할 모듈을 통해 의미적으로 연속성을 갖는 장면으로 병합되어 장면 단위의 영상으로 생성될 수 있다.

[0040] 이와 같이 장면 단위의 영상은 샷 분할, 장면 분할 과정을 통해 생성될 수 있으며, 영상을 장면 단위로 분할하는 방법은 특정한 방법에 한정되지 않는다.

[0041] 이 때, 입력 각본은 별도의 각본 분할 모듈을 통해서 영상과 동일한 장면 단위로 분할되어 장소, 시간, 지문 형식의 인스턴스로 변화될 수 있다.

[0042] 따라서, 이와 같이 분할된 각본은 별도의 동기화 모듈을 기반으로 장면 단위로 분할된 영상과 동기화될 수 있다.

[0043] 또한, 본 발명의 일실시예에 따른 벡터를 이용한 장면 묘사 기반의 메타데이터 생성 방법은 동기화된 장면 단위의 영상 및 각본 각각에 대해 적어도 하나의 특징에 상응하는 N차원의 벡터값을 추출한다(S120).

[0044] 예를 들어, 도 2를 참조하면, 장면 분할(S240) 단계를 통해 장면 단위로 분할된 영상을 기반으로 적어도 하나의 특징에 상응하는 N차원 벡터값을 추출할 수 있다(S260).

[0045] 이 때, 장면 단위의 영상에서는 영상의 행위 정보, 장소 정보, 시간 정보, 객체 정보, 지문 정보에 상응하는 특징 정보들이 추출될 수 있다.

[0046] 예를 들어, 행위 정보는 '달린다', '때린다', '이야기한다' 등과 같이 인물의 행동을 기준으로 생성되는 정보에 상응할 수 있다. 장소 정보는 영상의 배경이 되는 장소에 상응하는 것으로 '공원', '학교', '도로' 등이 될 수 있다. 시간 정보는 '밤', '낮', '아침' 같이 영상의 시간적 배경이 되는 정보에 상응할 수 있다. 객체 정보는 '전화기', '컵', '벤치' 와 같이 영상에 나타난 객체들에 대한 정보에 해당할 수 있다. 지문 정보는 영상 캡셔닝의 결과로써 '남자와 여자가 이야기하고 있다' 또는 '도시의 야경이 펼쳐지고 있다' 등 영상을 설명하는 정보에 상응할 수 있다.

[0047] 이 때, 도 3을 참조하면, 본 발명에서는 메타데이터 생성 장치에 포함된 정보 추출 모듈(310)로 장면 영상을 입력함으로써 적어도 하나의 특징에 상응하는 N차원 벡터값(320)을 추출할 수 있다.

[0048] 이 때, 각각의 벡터값들은 정보 추출 모듈(310)에 포함된 각 인식 모델을 통해 추출된 벡터에 상응하는 것으로써 분류 태스크에서 Softmax 함수를 통해 클래스 정보를 텍스트로 표현할 수 있다. 이 때, 본 발명에서는 Softmax 함수를 사용하기 전의 잠재 벡터를 장면에 대한 특징 정보로 사용할 수 있다.

[0049] 또한, 본 발명의 일실시예에 따른 벡터를 이용한 장면 묘사 기반의 메타데이터 생성 방법은 N차원의 벡터값으로 기반으로 영상 잠재 벡터 모델 및 각본 잠재 벡터 모델에 대한 기계학습을 수행하여 메타데이터 생성 모델을 구축한다(S130).

[0050] 이 때, 영상 잠재 벡터 모델은 장면 단위의 영상에 대한 메타데이터를 생성하고, 각본 잠재 벡터 모델은 장면 단위의 각본에 대한 메타데이터를 생성할 수 있다.

[0051] 이 때, 손실 함수를 기반으로 영상 잠재 벡터 모델 및 각본 잠재 벡터 모델 각각에 포함된 인코더 및 디코더를 영상 잠재 벡터 모델과 각본 잠재 벡터 모델 사이의 차이가 감소하도록 학습시킬 수 있다.

[0052] 이 때, 영상 잠재 벡터 모델 및 각본 잠재 벡터 모델 각각에 상응하는 적어도 하나의 입력 벡터 및 출력 벡터를 획득할 수 있다.

[0053] 예를 들어, 도 4를 참조하면, 영상 잠재 벡터 모델(410)은 정보 추출 모듈을 통해 입력되는 장면 단위의 영상에 대한 N차원 벡터값을 입력값으로 학습될 수 있고, 각본 잠재 벡터 모델(420)은 각본 분할 모듈을 통해 입력되는 장면 단위의 각본에 대한 N차원 벡터값을 입력값으로 학습될 수 있다. 즉, 영상과 각본에서 각각 추출된 인스턴스 정보를 이용하여 잠재 벡터로 표현할 수 있는 인코더 및 디코더를 학습시킬 수 있다.

[0054] 이 때, 영상 잠재 벡터 모델(410)과 각본 잠재 벡터 모델(420)에 각각 포함된 잠재 벡터는 입력된 벡터값들을 인코더를 통해 압축 표현한 것으로써 다시 디코더를 통해 각각의 벡터로 재현할 수 있다.

[0055] 예를 들어, 장면 단위의 영상에서 추출된 장소 정보, 시간 정보, 행위 정보, 객체 정보 및 지문 정보들은 영상 잠재 벡터 모델(410)의 인코더를 통해 N차원의 영상 잠재 벡터로 표현될 수 있으며, 영상 잠재 벡터는 디코더를 통해 다시 장소 정보, 시간 정보, 행위 정보, 객체 정보 및 지문 정보로 재현될 수 있다.

[0056] 이와 마찬가지로 장면 단위의 각본에서 추출된 장소, 시간, 지문 인스턴스는 워드 임베딩 모델을 통해 N차원 벡터값으로 변환될 수 있고, 벡터값으로 변환된 각각의 인스턴스 정보들은 각본 잠재 벡터 모델(420)을 통해 N차원의 각본 잠재 벡터로 표현될 수 있다. 또한, 디코더를 통해 각본 잠재 벡터를 다시 벡터값으로 변환된 각각의 인스턴스 정보들로 재현할 수도 있다.

[0057] 이 때, 영상 잠재 벡터 모델과 각본 잠재 벡터 모델 간 적어도 하나의 입력 벡터 및 출력 벡터의 오차가 감소하도록 인코더 및 디코더의 변수를 조정할 수 있다.

[0058] 예를 들어, 손실함수는 [수학식 1]과 같이 영상 잠재 벡터 모델과 각본 잠재 벡터 모델 사이의 차이가 작아지도록 학습함으로써, 메타데이터 생성 모델이 입력된 영상 정보들과 각본 정보가 중재된 벡터를 가지도록 할 수 있다. 이는 추후에 영상이나 각본 중 하나의 리소스의 부재에도 두 리소스 전체를 반영한 메타데이터를 생성할 수 있는 효과가 있다.

[0059] [수학식 1]

$$\text{Loss} = \sum_{i=1}^{m} (\text{영상잠재벡터}_i - \text{각본잠재벡터}_i)^2$$

[0060]

[0061] 상기와 같이 메타데이터 생성 모델을 학습시키는 과정을 상세하게 나타내면 도 5와 같이 나타낼 수 있다.

[0062] 도 5를 참조하면, 영상과 각본이 입력되면(S510), 입력된 데이터를 구분하여(S515) 영상과 각본을 구별할 수 있다.

[0063] 이 후, 영상의 특징을 분석하여 추출하고(S520), 추출된 특징을 기반으로 샷 분할과 장면 분할을 수행하여(S530) 입력된 영상에 대한 장면 단위의 영상을 생성할 수 있다. 각본 또한 영상에서 추출된 특징을 기반으로 분할하고(S540), 장면 단위의 각본에서 정보를 추출할 수 있다(S550).

[0064] 이 후, 장면 단위의 영상과 각본을 동기화할 수 있다(S560).

[0065] 이 후, 또 다시 데이터를 영상과 각본으로 구분하여(S565), 장면 단위의 영상에 대해 추출된 특징에 상응하는 N차원 벡터값을 추출한 뒤 영상 잠재 벡터 모델을 학습시키고(S570), 각본에 대해 추출된 특징에 상응하는 N차원 벡터값을 추출한 뒤 각본 잠재 벡터 모델을 학습시킬 수 있다(S580).

[0066] 이 때, 손실 함수를 이용하여 영상 잠재 벡터 모델 및 각본 잠재 벡터 모델 각각에 포함된 인코더 및 디코더를 영상 잠재 벡터 모델과 각본 잠재 벡터 모델 사이의 차이가 감소하도록 학습시킬 수 있다.

[0067] 이와 같은 학습을 통해 메타데이터 생성 모델을 구축할 수 있다(S590).

[0068] 이하에서는 학습된 메타데이터 생성 모델을 기반으로 메타데이터를 생성하는 단계를 설명하도록 한다.

[0069] 또한, 본 발명의 일실시예에 따른 벡터를 이용한 장면 묘사 기반의 메타데이터 생성 방법은 타겟 영상 및 타겟

각본 중 어느 하나만 입력된 경우, 메타데이터 생성 모델을 기반으로 타겟 영상과 타겟 각본이 중재된 형태의 메타데이터를 생성한다(S140).

[0070] 예를 들어, 도 6 내지 도 7을 참조하면, 타겟 영상(610)만 입력되었다고 가정하면(S710), 타겟 영상에 대한 특징을 분석 및 추출할 수 있다(S720). 이 후, 추출된 특징을 기반으로 타겟 영상에 대한 샷 분할 및 장면 분할을 수행하고(S730), N차원 벡터값을 추출하여 학습된 메타데이터 생성 모델(620)로 입력할 수 있다(S740). 이 후, 메타데이터 생성 모델(620)이 영상 및 각본이 중재된 메타데이터(630)를 생성하여 제공하면(S750), 별도의 저장소에 장면 단위의 영상에 대해 생성된 메타데이터(630)를 저장할 수 있다(S760).

[0071] 다른 예를 들어, 도 8 내지 도 9를 참조하면, 타겟 각본(810)만 입력되었다고 가정하면(S910), 타겟 각본에 대한 정보를 추출할 수 있다(S920). 이 후, 추출된 정보를 기반으로 N차원 벡터값을 추출하여 학습된 메타데이터 생성 모델(820)로 입력할 수 있다(S930). 이 후, 메타데이터 생성 모델(820)이 영상 및 각본이 중재된 메타데이터(830)를 생성하여 제공하면(S940), 별도의 저장소에 장면 단위의 각본에 대해 생성된 메타데이터(830)를 저장할 수 있다(9560).

[0072] 이 때, 메타데이터는 실수형의 벡터 형식에 상응할 수 있다.

[0073] 예를 들어, 메타데이터는 아래와 같이 실수형의 벡터 형식으로 표현될 수 있다. '-0.3188 -0.456 0.1647 -0.2495 -0.9385 -0.3563 -0.0732 0.4643 0.2975 0.5199 -0.393 -0.0246 0.3141 -0.2959 0.4304 -0.471 0.3575 -0.0843 -0.213 -0.1231 0.0305 -0.1432 0.1642 -0.3709 -0.3492 0.3128 -0.3125 0.197 -0.3428 0.0803 0.7957 0.2245 -0.1008 -0.2649 -0.2901 0.1465 -0.2685 0.0587 0.3844 -0.0293 -0.0189 0.352 0.058 0.9351 0.0297 -0.5063 -0.256 0.6041 -0.1265 0.1885 -0.0964 -0.3505 -0.2208 -0.6876 -0.4202 0.0777 -0.0117 0.6559 0.3044 0.1288 -0.0751 -0.8191 -0.2499 -0.3286 -0.0498 -0.0646 0.2347 -0.0313 -0.1417 0.0131 -0.0557 0.4898 0.2188 -0.4096 -0.0245 -0.4827 0.0847 0.2517 0.203 0.2854 -0.0851 -0.2978 -0.0002 -0.4072 0.4154 0.4654 -0.1107 0.7675 0.2345 - 쪋'

[0074] 또한, 도 1에는 도시하지 아니하였으나, 본 발명의 일실시예에 따른 벡터를 이용한 장면 묘사 기반의 메타데이터 생성 방법은 상술한 메타데이터 생성 과정에서 발생하는 다양한 정보를 별도의 저장 모듈에 저장할 수 있다.

[0075] 이와 같은 메타데이터 생성 방법을 이용함으로써 영상을 기반으로 장면을 묘사하는 정보를 추출하고, 영상에 대한 의미적 메타데이터를 생성 및 제공할 수 있다.

[0076] 또한, 기계학습 기반의 영상 분석을 통해 영상을 묘사하는 장소, 시간, 행동 등과 같은 의미적인 정보들을 추출할 수 있으며, 이러한 정보들에 기반한 학습을 통해 통합된 하나의 벡터 형식의 의미적 메타데이터를 생성할 수도 있다.

[0077] 또한, 각본의 수급이 비교적 어려운 환경에서 각본이 없이도 영상만으로 각본을 충분히 고려한 메타데이터를 제공하는 효과를 얻을 수도 있다.

[0079] 도 10은 본 발명의 일실시예에 따른 벡터를 이용한 장면 묘사 기반의 메타데이터 장치를 나타낸 도면이다.

[0080] 본 발명은 기존의 키워드 기반의 메타데이터 생성 기술에서 진보하여, 영상에서 추출 가능한 여러 특징 정보들이 통합된 벡터 기반의 메타데이터를 생성하는 것이다. 기존의 일반적인 영상 메타데이터 생성 방법들이 영상에 등장하는 키워드들을 태깅하는 것이었다면, 본 발명에서는 영상의 의미적 요소들을 포함한 벡터 기반의 메타데이터를 생성하는 방법을 제공한다. 이를 위해, 기계학습 기반의 영상 분석을 통해 장소, 시간, 행위, 지문 등의 특징 정보들을 추출할 수 있다. 각각의 특징 정보들은 N차원 벡터 형식의 메타데이터로 표현할 수 있으며, 벡터로 변환하는 인코더 및 디코더의 학습을 통해 벡터들은 다시 텍스트로 재현할 수 있다.

[0081] 이러한 개념을 확장하여, 학습을 통해 영상에서 추출한 특징 정보들을 통합해 하나의 벡터로 표현하고, 통합된 벡터를 기반으로 다시 정보들을 재현할 수 있다. 즉, 통합된 벡터는 추출된 특징 정보들이 축약된 것으로써 영상의 정보들을 나타내는 메타데이터로써 사용될 수 있다는 것을 의미한다.

[0082] 본 발명은 크게 메타데이터를 생성하는 메타데이터 생성 모델을 학습하는 단계와 학습된 메타데이터 생성 모델을 기반으로 메타데이터를 생성하는 단계로 구분할 수 있는데, 이하에서는 먼저 메타데이터 생성 모델을 학습하는 단계를 설명하도록 한다.

[0083] 도 10을 참조하면, 본 발명의 일실시예에 따른 벡터를 이용한 장면 묘사 기반의 메타데이터 생성 장치는 통신부

(1010), 프로세서(1020) 및 메모리(1030)를 포함한다.

[0084] 통신부(1010)는 네트워크와 같은 통신망을 통해 정보 공유를 위해 필요한 정보를 송수신하는 역할을 한다.

[0085] 프로세서(1020)는 입력 영상에서 추출된 적어도 하나의 특징을 기반으로 입력 영상을 입력 각본과 장면 단위로 동기화한다.

[0086] 예를 들어, 도 2를 참조하면, 입력된 각본(210)에 대해 각본 분할(S210)을 수행하고, 입력된 영상(220)에 대해 특징 분석(S220)후 샷 분할(S230) 및 장면 분할(S240)을 수행한 뒤 각각 분할된 영상과 각본에 대해 동기화를 수행할 수 있다(S250).

[0087] 이 때, 입력 각본은 입력 영상이 제작되기에 앞서 기록된 텍스트 형식의 문서에 상응하는 것으로, 장소, 시간, 장면 등을 설명하는 지문이 기록될 수 있다.

[0088] 이 때, 입력된 영상(220)에 대한 특징 분석(S220)은 메타데이터 생성 장치에 포함된 별도의 영상 특징 분석 모듈을 통해 수행될 수 있다.

[0089] 이 때, 적어도 하나의 특징을 기반으로 입력 영상을 샷 단위로 분할할 수 있다.

[0090] 이 때, 샷 단위로 분할된 복수개의 샷 영상들을 의미적 연속성을 갖는 장면 단위로 병합할 수 있다.

[0091] 예를 들어, 영상에 대한 특징 분석(S220)을 통해 추출된 특징 정보들을 별도의 샷 분할 모듈을 이용하여 1~3초 내외의 샷으로 분할할 수 있다. 이렇게 분할된 샷은 장면 분할 모듈을 통해 의미적으로 연속성을 갖는 장면으로 병합되어 장면 단위의 영상으로 생성될 수 있다.

[0092] 이와 같이 장면 단위의 영상은 샷 분할, 장면 분할 과정을 통해 생성될 수 있으며, 영상을 장면 단위로 분할하는 방법은 특정한 방법에 한정되지 않는다.

[0093] 이 때, 입력 각본은 별도의 각본 분할 모듈을 통해서 영상과 동일한 장면 단위로 분할되어 장소, 시간, 지문 형식의 인스턴스로 변화될 수 있다.

[0094] 따라서, 이와 같이 분할된 각본은 별도의 동기화 모듈을 기반으로 장면 단위로 분할된 영상과 동기화될 수 있다.

[0095] 또한, 프로세서(1020)는 동기화된 장면 단위의 영상 및 각본 각각에 대해 적어도 하나의 특징에 상응하는 N차원의 벡터값을 추출한다.

[0096] 예를 들어, 도 2를 참조하면, 장면 분할(S240) 단계를 통해 장면 단위로 분할된 영상을 기반으로 적어도 하나의 특징에 상응하는 N차원 벡터값을 추출할 수 있다(S260).

[0097] 이 때, 장면 단위의 영상에서는 영상의 행위 정보, 장소 정보, 시간 정보, 객체 정보, 지문 정보에 상응하는 특징 정보들이 추출될 수 있다.

[0098] 예를 들어, 행위 정보는 '달린다', '때린다', '이야기한다' 등과 같이 인물의 행동을 기준으로 생성되는 정보에 상응할 수 있다. 장소 정보는 영상의 배경이 되는 장소에 상응하는 것으로 '공원', '학교', '도로' 등이 될 수 있다. 시간 정보는 '밤', '낮', '아침' 같이 영상의 시간적 배경이 되는 정보에 상응할 수 있다. 객체 정보는 '전화기', '컵', '벤치' 와 같이 영상에 나타난 객체들에 대한 정보에 해당할 수 있다. 지문 정보는 영상 캡셔닝의 결과로써 '남자와 여자가 이야기하고 있다' 또는 '도시의 야경이 펼쳐지고 있다' 등 영상을 설명하는 정보에 상응할 수 있다.

[0099] 이 때, 도 3을 참조하면, 본 발명에서는 메타데이터 생성 장치에 포함된 정보 추출 모듈(310)로 장면 영상을 입력함으로써 적어도 하나의 특징에 상응하는 N차원 벡터값(320)을 추출할 수 있다.

[0100] 이 때, 각각의 벡터값들은 정보 추출 모듈(310)에 포함된 각 인식 모델을 통해 추출된 벡터에 상응하는 것으로써 분류 태스크에서 Softmax 함수를 통해 클래스 정보를 텍스트로 표현할 수 있다. 이 때, 본 발명에서는 Softmax 함수를 사용하기 전의 잠재 벡터를 장면에 대한 특징 정보로 사용할 수 있다.

[0101] 또한, 프로세서(1020)는 N차원의 벡터값으ㄹ 기반으로 영상 잠재 벡터 모델 및 각본 잠재 벡터 모델에 대한 기계학습을 수행하여 메타데이터 생성 모델을 구축한다.

[0102] 이 때, 영상 잠재 벡터 모델은 장면 단위의 영상에 대한 메타데이터를 생성하고, 각본 잠재 벡터 모델은 장면 단위의 각본에 대한 메타데이터를 생성할 수 있다.

[0103] 이 때, 손실 함수를 기반으로 영상 잠재 벡터 모델 및 각본 잠재 벡터 모델 각각에 포함된 인코더 및 디코더를 영상 잠재 벡터 모델과 각본 잠재 벡터 모델 사이의 차이가 감소하도록 학습시킬 수 있다.

[0104] 이 때, 영상 잠재 벡터 모델 및 각본 잠재 벡터 모델 각각에 상응하는 적어도 하나의 입력 벡터 및 출력 벡터를 획득할 수 있다.

[0105] 예를 들어, 도 4를 참조하면, 영상 잠재 벡터 모델(410)은 정보 추출 모듈을 통해 입력되는 장면 단위의 영상에 대한 N차원 벡터값을 입력값으로 학습될 수 있고, 각본 잠재 벡터 모델(420)은 각본 분할 모듈을 통해 입력되는 장면 단위의 각본에 대한 N차원 벡터값을 입력값으로 학습될 수 있다. 즉, 영상과 각본에서 각각 추출된 인스턴스 정보를 이용하여 잠재 벡터로 표현할 수 있는 인코더 및 디코더를 학습시킬 수 있다.

[0106] 이 때, 영상 잠재 벡터 모델(410)과 각본 잠재 벡터 모델(420)에 각각 포함된 잠재 벡터는 입력된 벡터값들을 인코더를 통해 압축 표현한 것으로써 다시 디코더를 통해 각각의 벡터로 재현할 수 있다.

[0107] 예를 들어, 장면 단위의 영상에서 추출된 장소 정보, 시간 정보, 행위 정보, 객체 정보 및 지문 정보들은 영상 잠재 벡터 모델(410)의 인코더를 통해 N차원의 영상 잠재 벡터로 표현될 수 있으며, 영상 잠재 벡터는 디코더를 통해 다시 장소 정보, 시간 정보, 행위 정보, 객체 정보 및 지문 정보로 재현될 수 있다.

[0108] 이와 마찬가지로 장면 단위의 각본에서 추출된 장소, 시간, 지문 인스턴스는 워드 임베딩 모델을 통해 N차원 벡터값으로 변환될 수 있고, 벡터값으로 변환된 각각의 인스턴스 정보들은 각본 잠재 벡터 모델(420)을 통해 N차원의 각본 잠재 벡터로 표현될 수 있다. 또한, 디코더를 통해 각본 잠재 벡터를 다시 벡터값으로 변환된 각각의 인스턴스 정보들로 재현할 수도 있다.

[0109] 이 때, 영상 잠재 벡터 모델과 각본 잠재 벡터 모델 간 적어도 하나의 입력 벡터 및 출력 벡터의 오차가 감소하도록 인코더 및 디코더의 변수를 조정할 수 있다.

[0110] 예를 들어, 손실함수는 [수학식 1]과 같이 영상 잠재 벡터 모델과 각본 잠재 벡터 모델 사이의 차이가 작아지도록 학습함으로써, 메타데이터 생성 모델이 입력된 영상 정보들과 각본 정보가 중재된 벡터를 가지도록 할 수 있다. 이는 추후에 영상이나 각본 중 하나의 리소스의 부재에도 두 리소스 전체를 반영한 메타데이터를 생성할 수 있는 효과가 있다.

[0111] [수학식 1]

$$\text{Loss} = \sum_{i=1}^{m} (\text{영상잠재벡터}_i - \text{각본잠재벡터}_i)^2$$

[0112]

[0113] 상기와 같이 메타데이터 생성 모델을 학습시키는 과정을 상세하게 나타내면 도 5와 같이 나타낼 수 있다.

[0114] 도 5를 참조하면, 영상과 각본이 입력되면(S510), 입력된 데이터를 구분하여(S515) 영상과 각본을 구별할 수 있다.

[0115] 이 후, 영상의 특징을 분석하여 추출하고(S520), 추출된 특징을 기반으로 샷 분할과 장면 분할을 수행하여(S530) 입력된 영상에 대한 장면 단위의 영상을 생성할 수 있다. 각본 또한 영상에서 추출된 특징을 기반으로 분할하고(S540), 장면 단위의 각본에서 정보를 추출할 수 있다(S550).

[0116] 이 후, 장면 단위의 영상과 각본을 동기화할 수 있다(S560).

[0117] 이 후, 또 다시 데이터를 영상과 각본으로 구분하여(S565), 장면 단위의 영상에 대해 추출된 특징에 상응하는 N차원 벡터값을 추출한 뒤 영상 잠재 벡터 모델을 학습시키고(S570), 각본에 대해 추출된 특징에 상응하는 N차원 벡터값을 추출한 뒤 각본 잠재 벡터 모델을 학습시킬 수 있다(S580).

[0118] 이 때, 손실 함수를 이용하여 영상 잠재 벡터 모델 및 각본 잠재 벡터 모델 각각에 포함된 인코더 및 디코더를 영상 잠재 벡터 모델과 각본 잠재 벡터 모델 사이의 차이가 감소하도록 학습시킬 수 있다.

[0119] 이와 같은 학습을 통해 메타데이터 생성 모델을 구축할 수 있다(S590).

[0120] 이하에서는 학습된 메타데이터 생성 모델을 기반으로 메타데이터를 생성하는 단계를 설명하도록 한다.

[0121] 또한, 프로세서(1020)는 타겟 영상 및 타겟 각본 중 어느 하나만 입력된 경우, 메타데이터 생성 모델을 기반으

로 타겟 영상과 타겟 각본이 중재된 형태의 메타데이터를 생성한다.

[0122] 예를 들어, 도 6 내지 도 7을 참조하면, 타겟 영상(610)만 입력되었다고 가정하면(S710), 타겟 영상에 대한 특징을 분석 및 추출할 수 있다(S720). 이 후, 추출된 특징을 기반으로 타겟 영상에 대한 샷 분할 및 장면 분할을 수행하고(S730), N차원 벡터값을 추출하여 학습된 메타데이터 생성 모델(620)로 입력할 수 있다(S740). 이 후, 메타데이터 생성 모델(620)이 영상 및 각본이 중재된 메타데이터(630)를 생성하여 제공하면(S750), 별도의 저장소에 장면 단위의 영상에 대해 생성된 메타데이터(630)를 저장할 수 있다(S760).

[0123] 다른 예를 들어, 도 8 내지 도 9를 참조하면, 타겟 각본(810)만 입력되었다고 가정하면(S910), 타겟 각본에 대한 정보를 추출할 수 있다(S920). 이 후, 추출된 정보를 기반으로 N차원 벡터값을 추출하여 학습된 메타데이터 생성 모델(820)로 입력할 수 있다(S930). 이 후, 메타데이터 생성 모델(820)이 영상 및 각본이 중재된 메타데이터(830)를 생성하여 제공하면(S940), 별도의 저장소에 장면 단위의 각본에 대해 생성된 메타데이터(830)를 저장할 수 있다(9560).

[0124] 이 때, 메타데이터는 실수형의 벡터 형식에 상응할 수 있다.

[0125] 예를 들어, 메타데이터는 아래와 같이 실수형의 벡터 형식으로 표현될 수 있다. '-0.3188 -0.456 0.1647 -0.2495 -0.9385 -0.3563 -0.0732 0.4643 0.2975 0.5199 -0.393 -0.0246 0.3141 -0.2959 0.4304 -0.471 0.3575 -0.0843 -0.213 -0.1231 0.0305 -0.1432 0.1642 -0.3709 -0.3492 0.3128 -0.3125 0.197 -0.3428 0.0803 0.7957 0.2245 -0.1008 -0.2649 -0.2901 0.1465 -0.2685 0.0587 0.3844 -0.0293 -0.0189 0.352 0.058 0.9351 0.0297 -0.5063 -0.256 0.6041 -0.1265 0.1885 -0.0964 -0.3505 -0.2208 -0.6876 -0.4202 0.0777 -0.0117 0.6559 0.3044 0.1288 -0.0751 -0.8191 -0.2499 -0.3286 -0.0498 -0.0646 0.2347 -0.0313 -0.1417 0.0131 -0.0557 0.4898 0.2188 -0.4096 -0.0245 -0.4827 0.0847 0.2517 0.203 0.2854 -0.0851 -0.2978 -0.0002 -0.4072 0.4154 0.4654 -0.1107 0.7675 0.2345 - 쩐'

[0126] 메모리(1030)는 메타데이터 생성 모델을 저장할 수 있다.

[0127] 또한, 메모리(1030)는 메타데이터 생성 과정에서 발생하는 다양한 정보를 저장할 수 있다.

[0128] 이와 같은 메타데이터 생성 장치를 통해 영상을 기반으로 장면을 묘사하는 정보를 추출하고, 영상에 대한 의미적 메타데이터를 생성 및 제공할 수 있다.

[0129] 또한, 기계학습 기반의 영상 분석을 통해 영상을 묘사하는 장소, 시간, 행동 등과 같은 의미적인 정보들을 추출할 수 있으며, 이러한 정보들에 기반한 학습을 통해 통합된 하나의 벡터 형식의 의미적 메타데이터를 생성할 수도 있다.

[0130] 또한, 각본의 수급이 비교적 어려운 환경에서 각본이 없이도 영상만으로 각본을 충분히 고려한 메타데이터를 제공하는 효과를 얻을 수도 있다.

[0132] 이상에서와 같이 본 발명에 따른 벡터를 이용한 장면 묘사 기반의 메타데이터 생성 방법 및 이를 위한 장치는 상기한 바와 같이 설명된 실시예들의 구성과 방법이 한정되게 적용될 수 있는 것이 아니라, 상기 실시예들은 다양한 변형이 이루어질 수 있도록 각 실시예들의 전부 또는 일부가 선택적으로 조합되어 구성될 수도 있다.

부호의 설명

[0133] 210: 각본 220: 영상

230: 장면 영상 데이터베이스 240: 메타데이터 데이터베이스

310: 정보 추출 모듈 320: N차원 벡터값

410: 영상 잠재 벡터 모델 420: 각본 잠재 벡터 모델

610: 타겟 영상 620, 820: 메타데이터 생성 모델

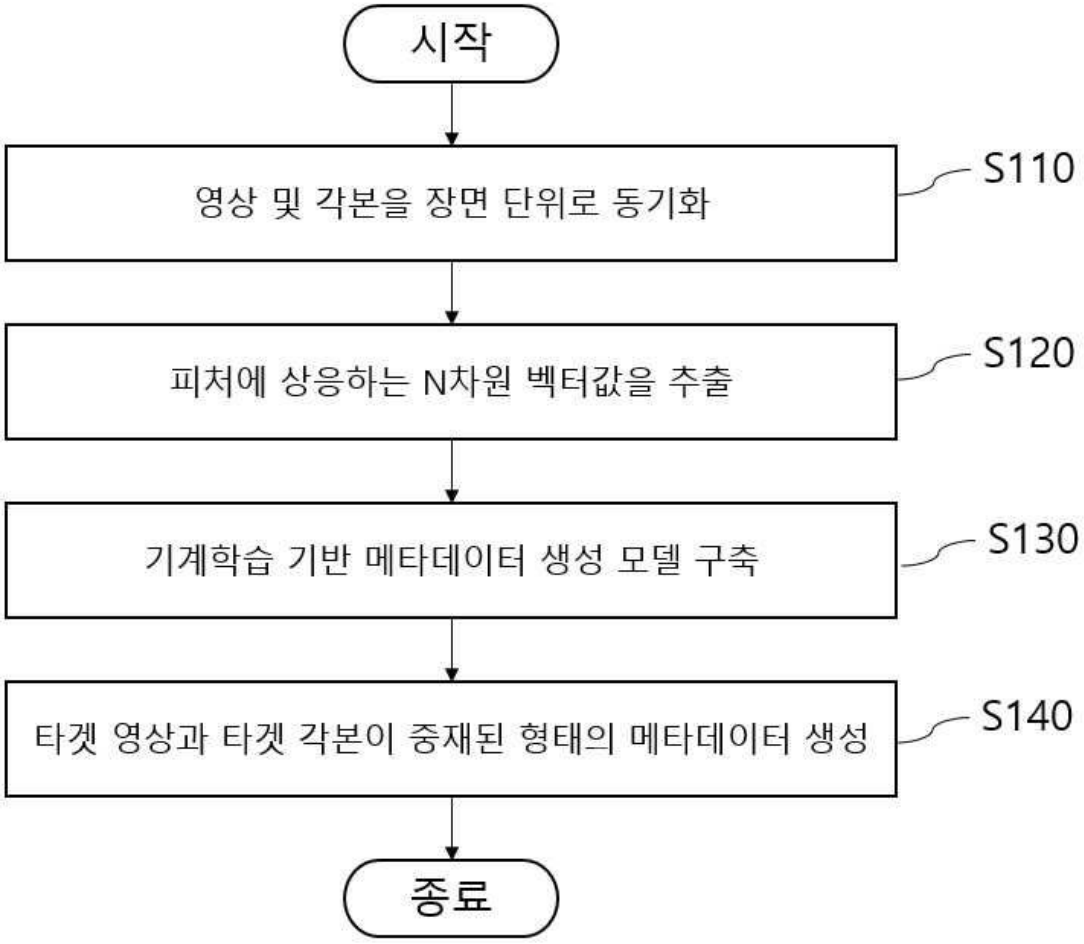
630, 830: 메타데이터 810: 타겟 각본

1010: 통신부 1020: 프로세서

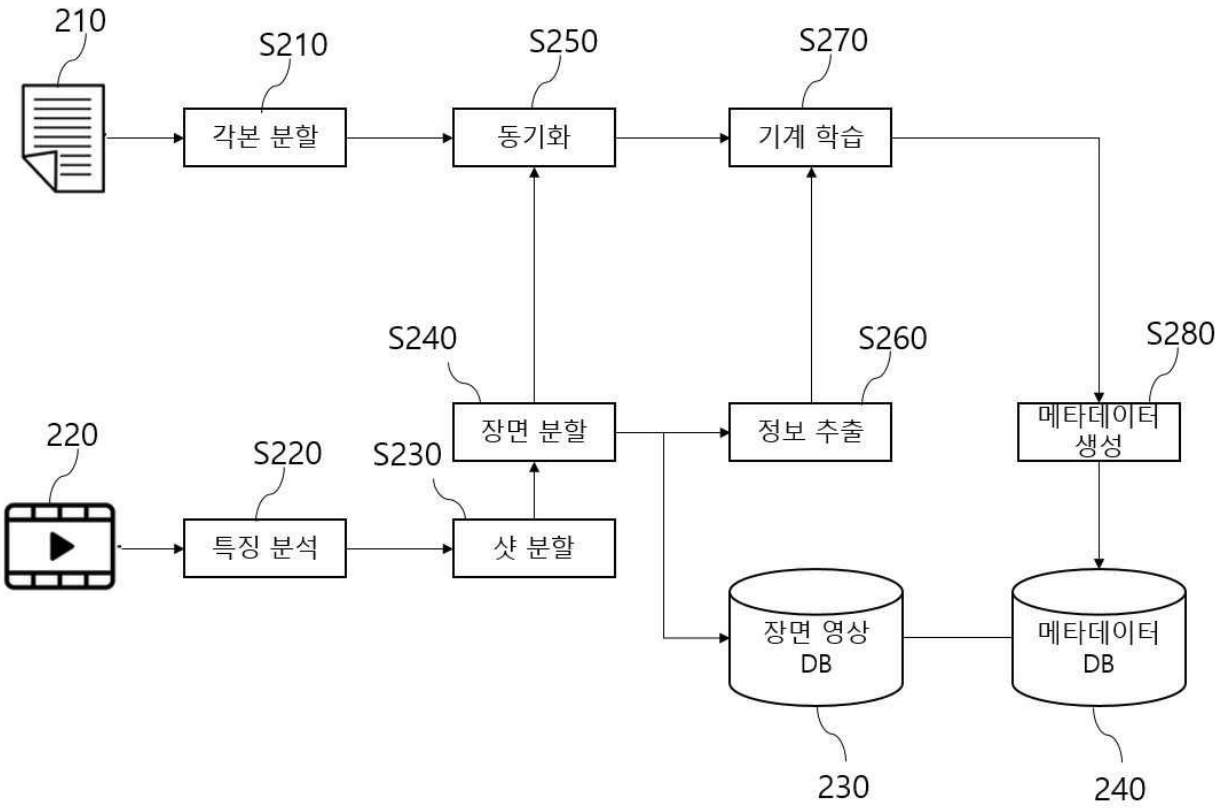
1030: 메모리

도면

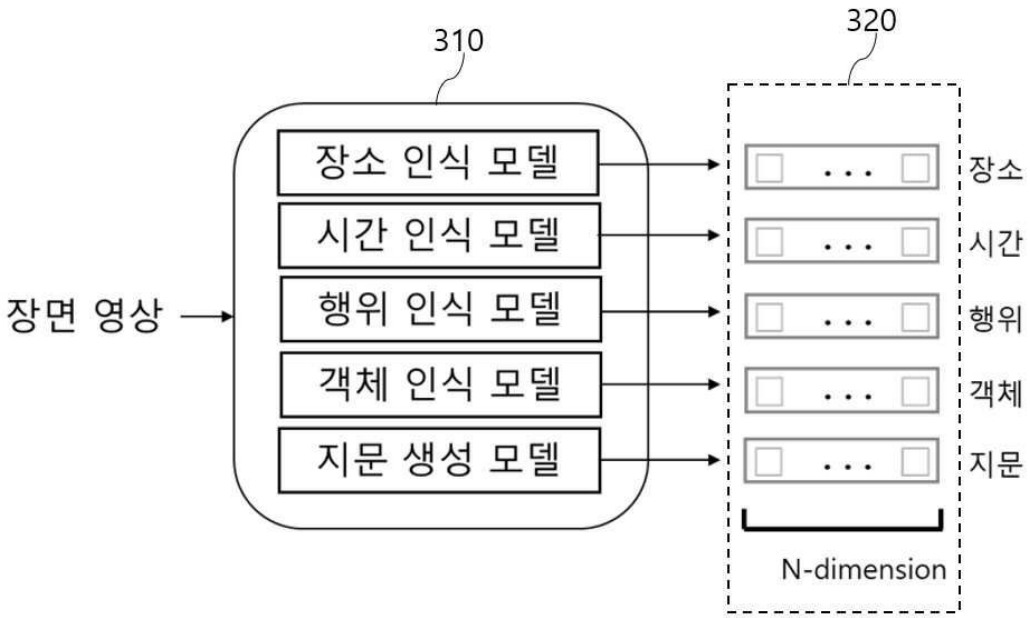
도면1


시작
영상 및 각본을 장면 단위로 동기화
S110
피처에 상응하는 N차원 벡터값을 추출
S120
기계학습 기반 메타데이터 생성 모델 구축
S130
타겟 영상과 타겟 각본이 중재된 형태의 메타데이터 생성
S140
종료

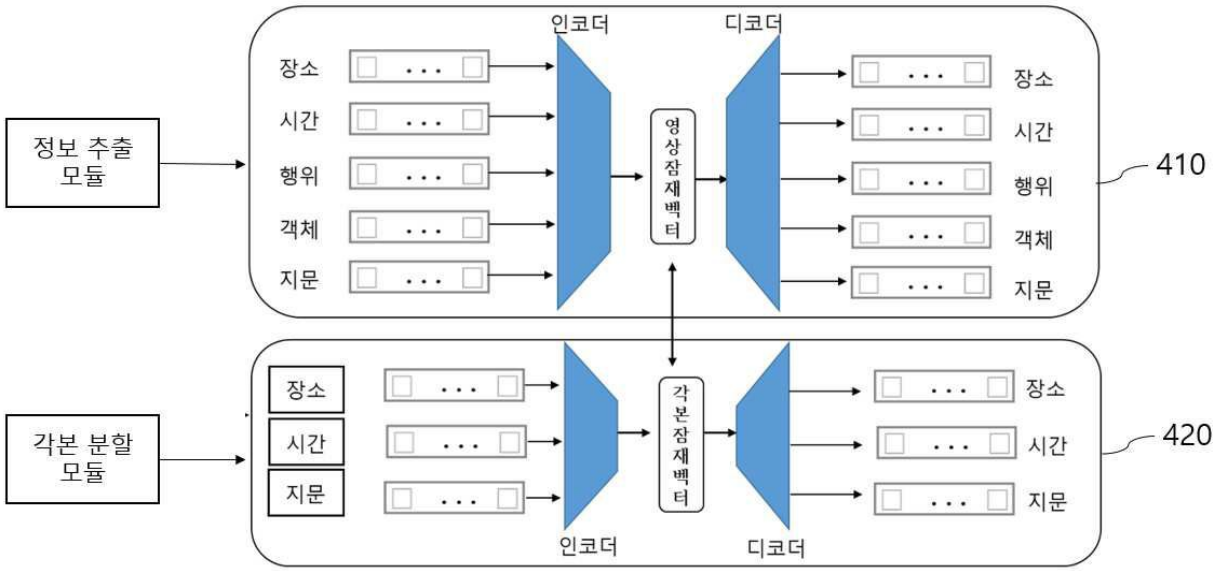
도면2


210
S210
각본 분할
S250
동기화
S270
기계 학습
S240
장면 분할
S260
정보 추출
S280
메타데이터 생성
220
S220
특징 분석
S230
샷 분할
장면 영상 DB
메타데이터 DB
230
240

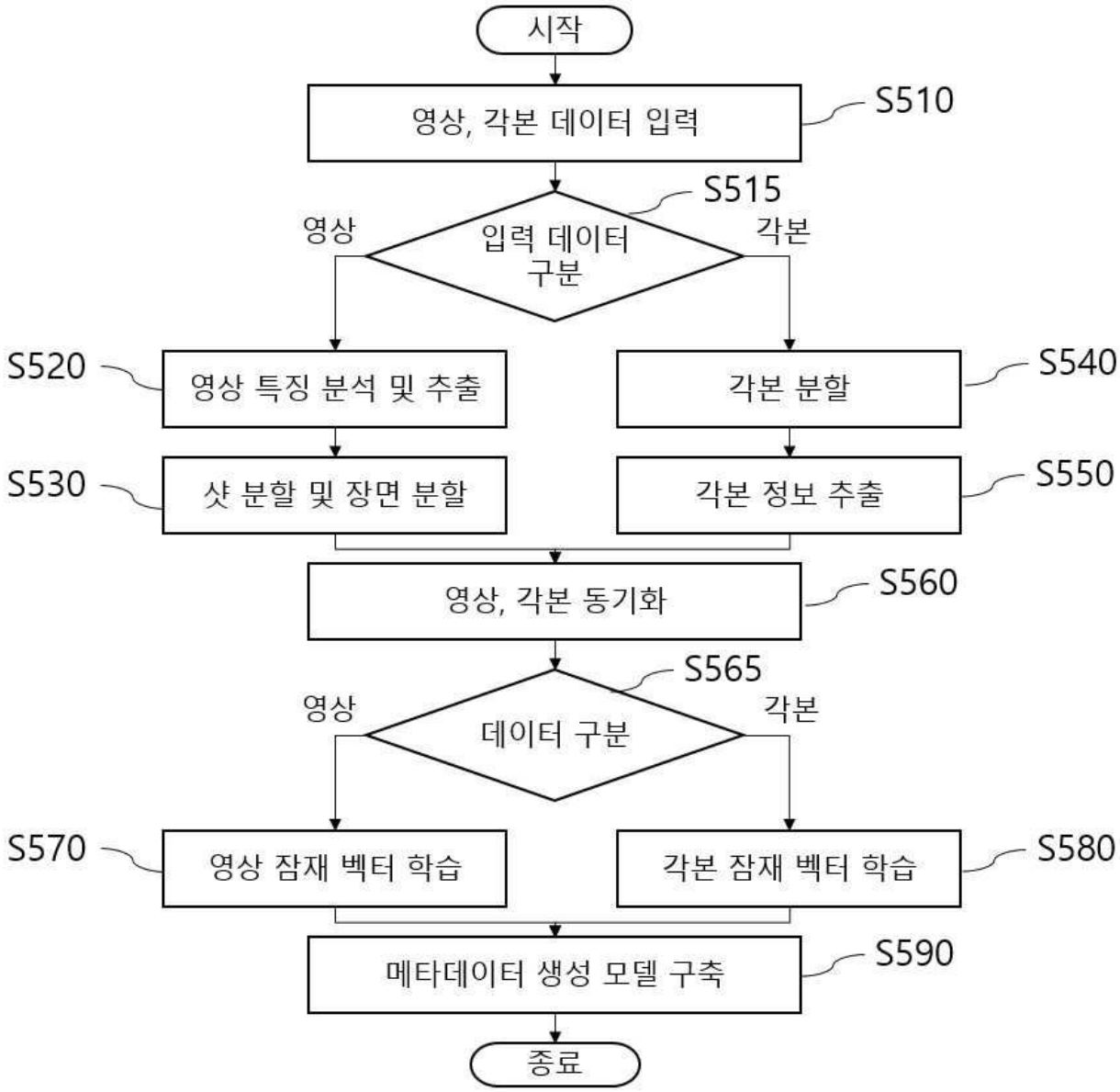
도면3


310
320
장면 영상
장소 인식 모델
시간 인식 모델
행위 인식 모델
객체 인식 모델
지문 생성 모델
장소
시간
행위
객체
지문
N-dimension

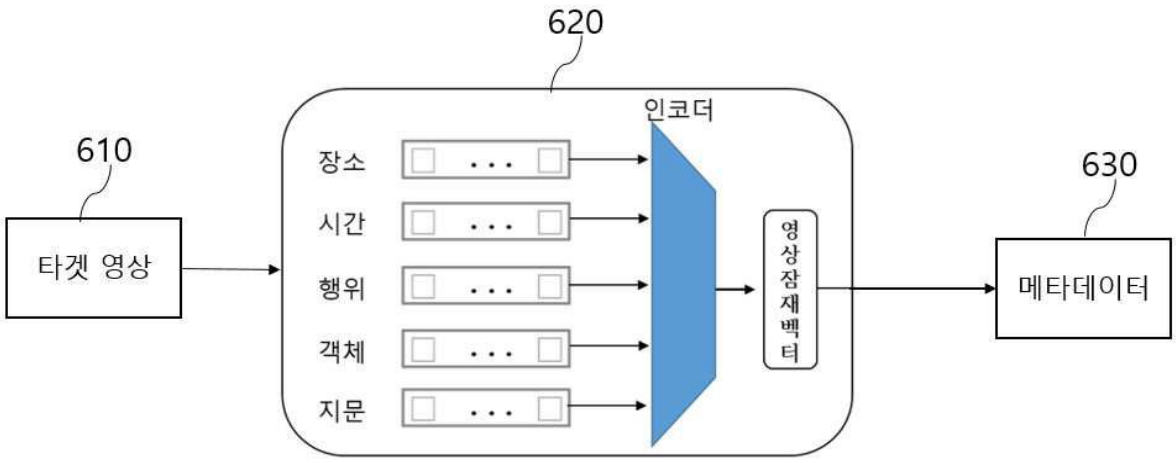
도면4


정보 추출 모듈
장소
시간
행위
객체
지문
인코더
영상잠재벡터
디코더
장소
시간
행위
객체
지문
410
각본 분할 모듈
장소
시간
지문
인코더
각본잠재벡터
디코더
장소
시간
지문
420

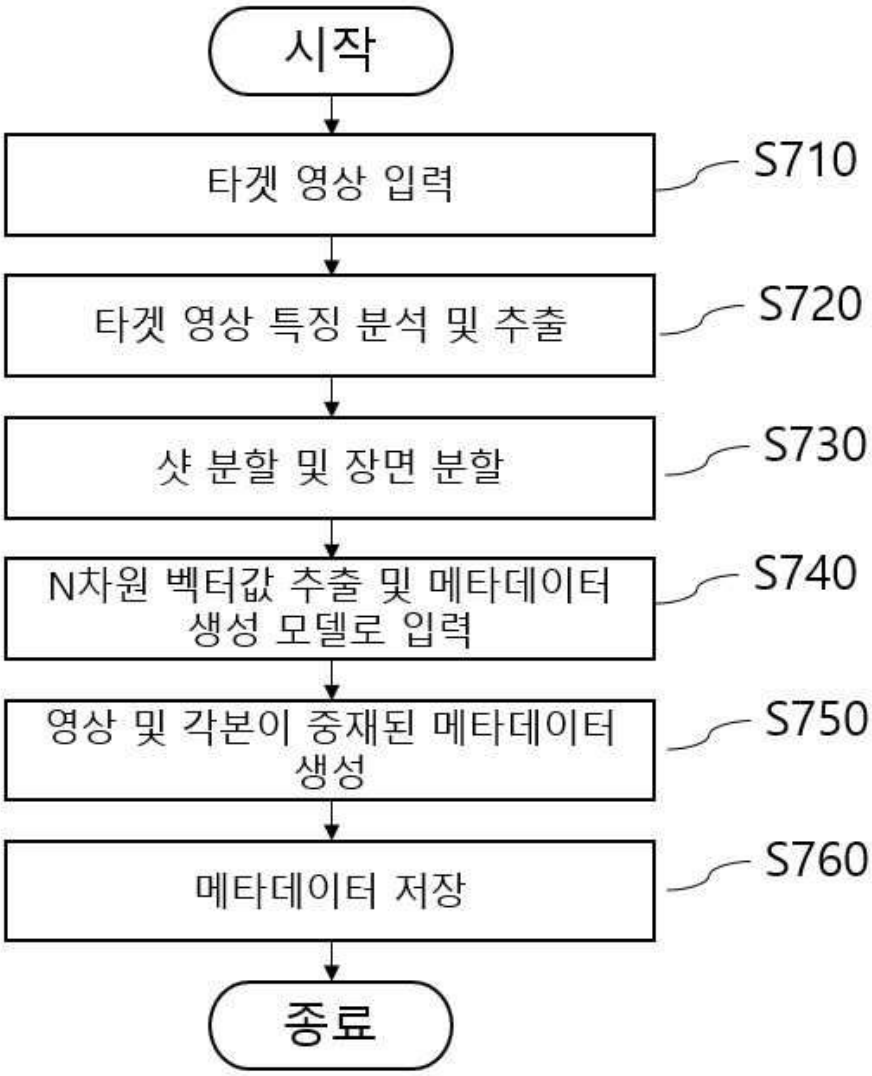
도면5


시작
영상, 각본 데이터 입력
S510
S515
영상
입력 데이터 구분
각본
S520
영상 특징 분석 및 추출
각본 분할
S540
S530
샷 분할 및 장면 분할
각본 정보 추출
S550
영상, 각본 동기화
S560
S565
영상
데이터 구분
각본
S570
영상 잠재 벡터 학습
각본 잠재 벡터 학습
S580
메타데이터 생성 모델 구축
S590
종료

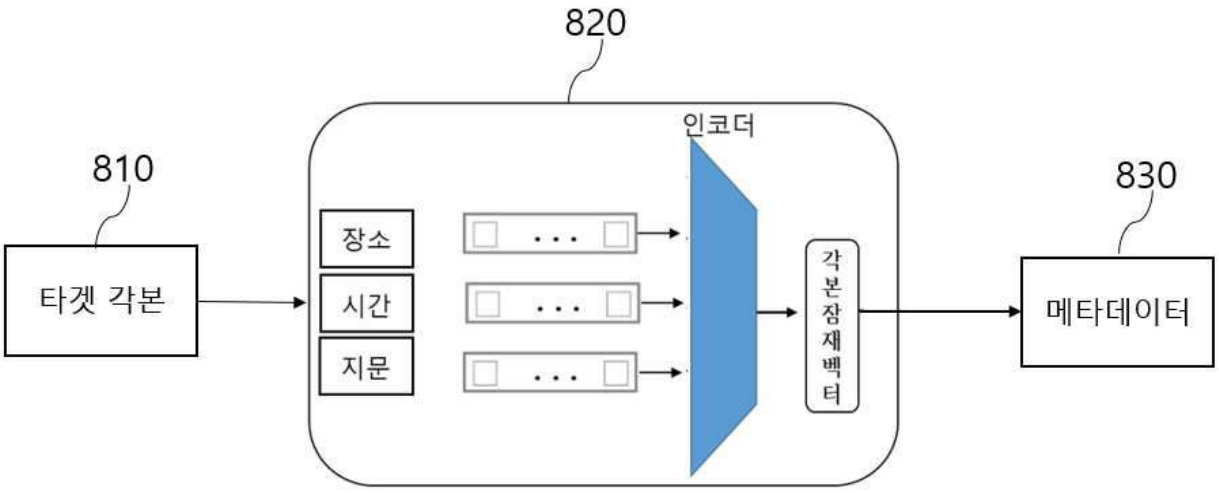
도면6


620
인코더
610
장소
시간
행위
객체
지문
영상잠재벡터
630
타겟 영상
메타데이터

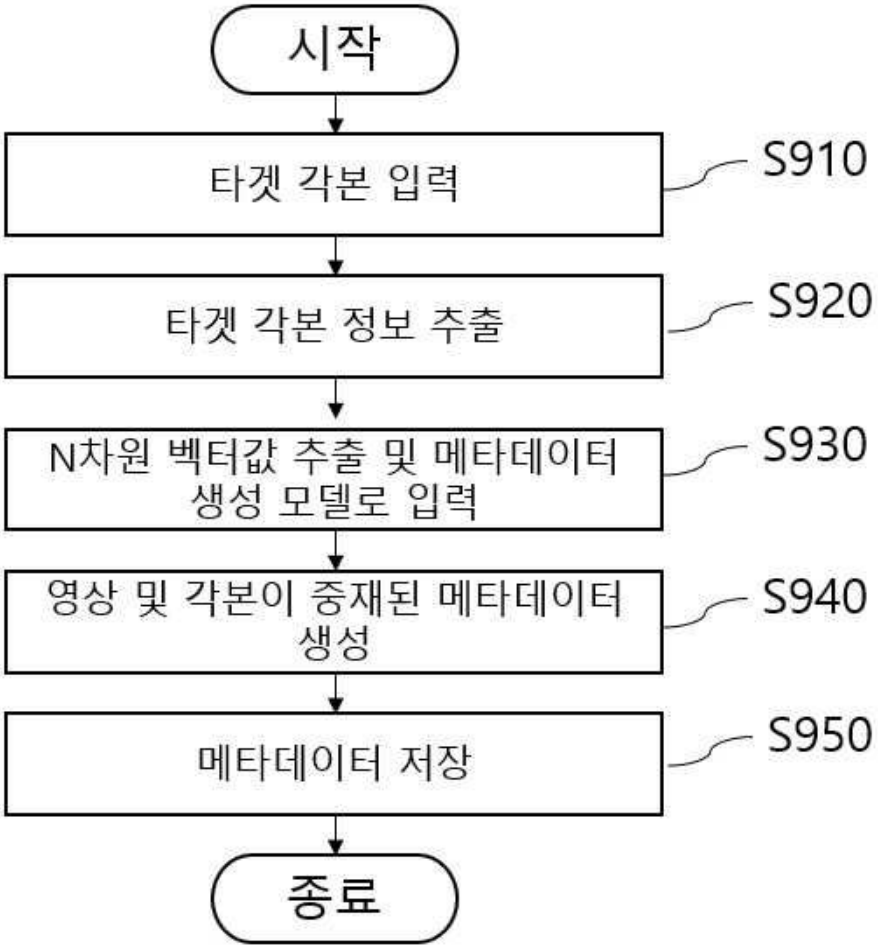
도면7


시작
타겟 영상 입력
S710
타겟 영상 특징 분석 및 추출
S720
샷 분할 및 장면 분할
S730
N차원 벡터값 추출 및 메타데이터 생성 모델로 입력
S740
영상 및 각본이 중재된 메타데이터 생성
S750
메타데이터 저장
S760
종료

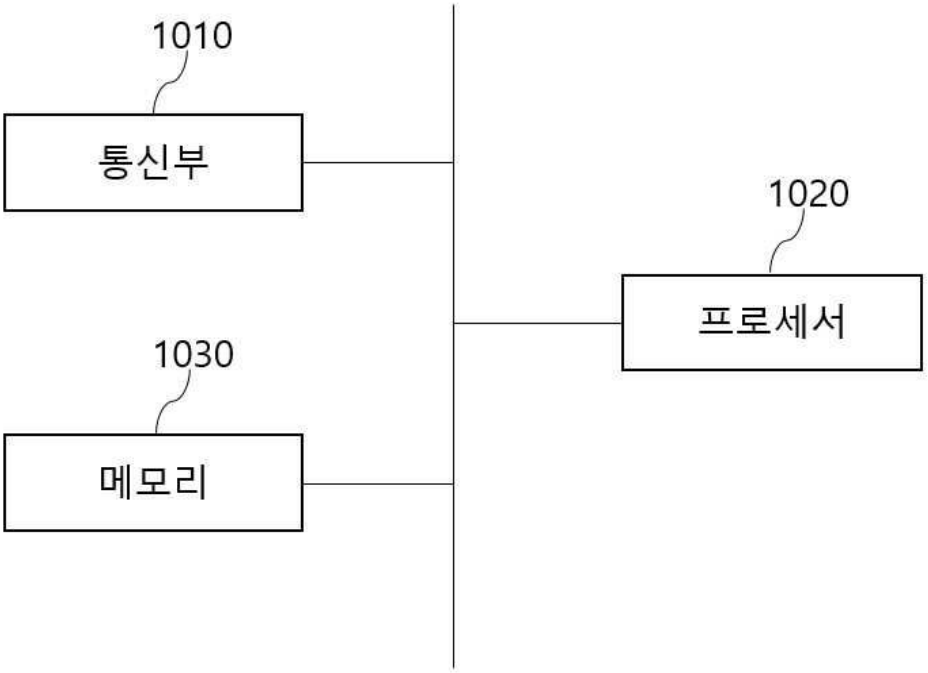
도면8


820
인코더
810
타겟 각본
장소
시간
지문
각본장재벡터
830
메타데이터

도면9


시작
타겟 각본 입력
S910
타겟 각본 정보 추출
S920
N차원 벡터값 추출 및 메타데이터 생성 모델로 입력
S930
영상 및 각본이 중재된 메타데이터 생성
S940
메타데이터 저장
S950
종료

도면10


1010
통신부
1020
프로세서
1030
메모리

IT R&D Global Leader

스마트기기 비전인식용 온디바이스 딥러닝 SW 플랫폼 기술

ETRI 한국전자통신연구원
www.etri.re.kr

목 차

1. 기술의 개요

2. 기술이전 내용 및 범위

3. 경쟁기술과 비교

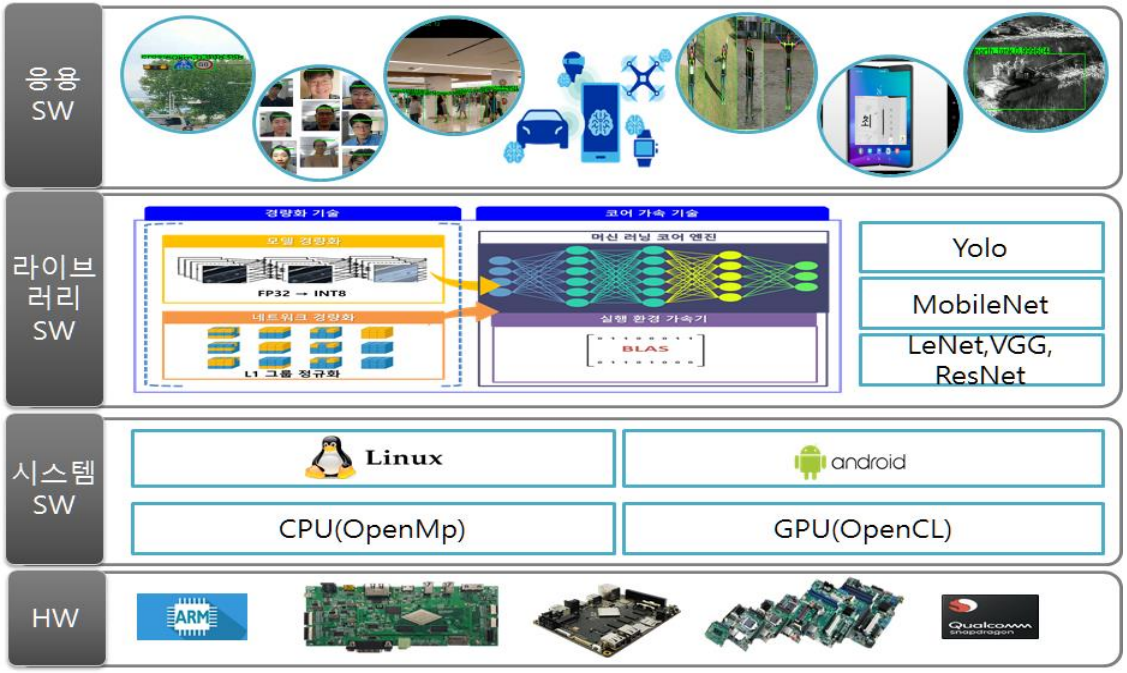
4. 기술의 사업성

5. 국내외 시장 동향

1. 기술의 개요

▣ 기술의 개념

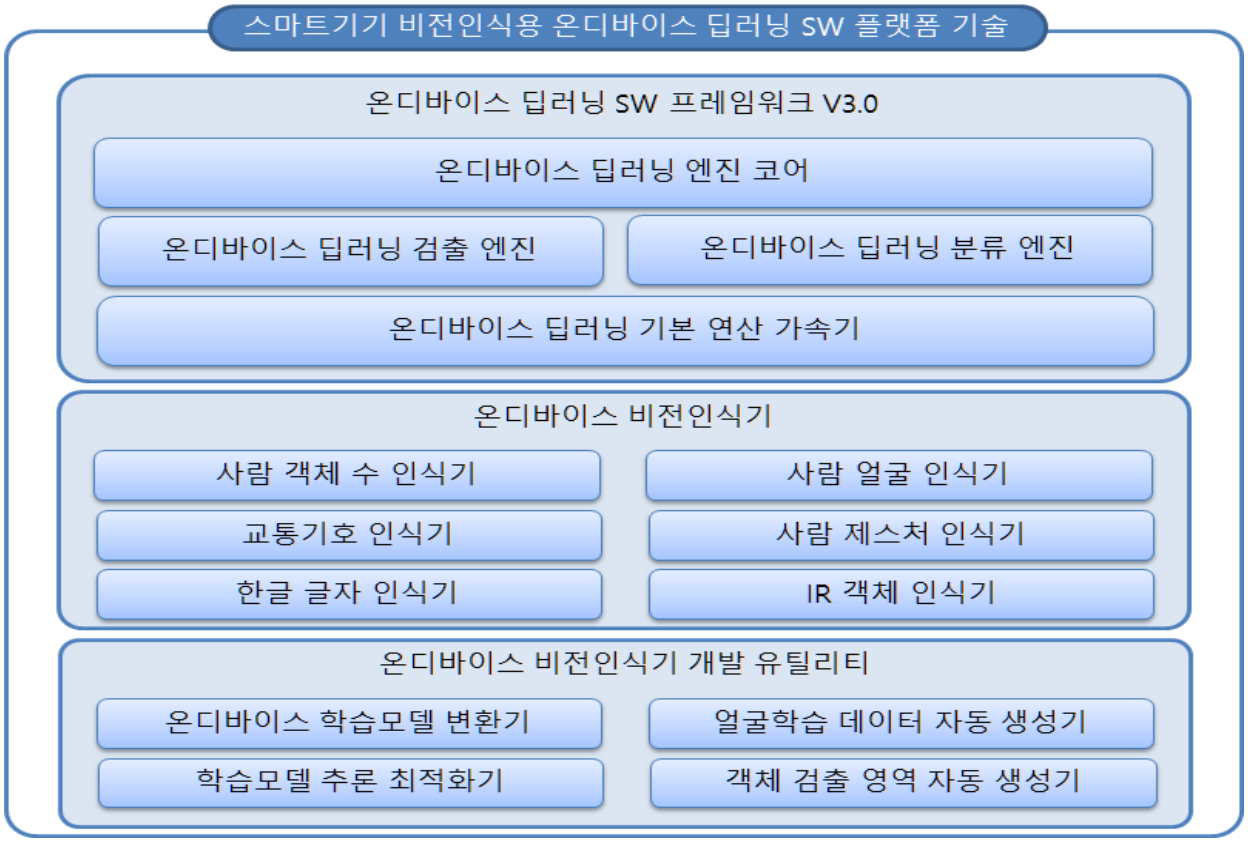
- ❖ 스마트기기에서 사용할 수 있는 온디바이스 인공지능 SW 기술로서 특히, 클라우드 연결없이 스마트기기 안에서 실시간 비전인식 서비스를 가능하게 하는 임베디드 시스템용 딥러닝 SW 기술
- ❖ 본 기술은 임베디드 시스템 HW 상에서 딥러닝 기반 비전인식을 가능하게 하는 기술로, 임베디드 리눅스 및 안드로이드 OS 상에서 실행 가능한 온디바이스 딥러닝 SW 프레임워크와 온디바이스 비전인식기 그리고 온디바이스 비전인식기 개발 유틸리티 등을 제공함



[기술의 개념도]

2. 기술이전 내용 및 범위[1/7]

▣ 기술이전 내용 및 범위



2. 기술이전 내용 및 범위[2/7]

▣ 기술이전 내용 및 범위

1. 스마트기기 비전인식용 온디바이스 딥러닝 SW 플랫폼 기술

1.1 온디바이스 딥러닝 SW 프레임워크 V3.0(바이너리)

: 온디바이스 딥러닝 엔진 코어
: 온디바이스 딥러닝 검출 엔진
: 온디바이스 딥러닝 분류 엔진
: 온디바이스 딥러닝 기본 연산 가속기

1.2 온디바이스 비전인식기(인식기 3종 선택, 소스)

: 사람 객체 수 인식기
: 사람 얼굴 인식기
: 교통기호 인식기
: 사람 제스처 인식기
: 한글 글자 인식기
: IR 객체 인식기

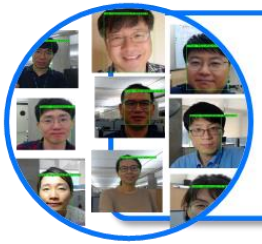
1.3 온디바이스 비전인식기 개발 유틸리티(바이너리)

: 온디바이스 학습모델 변환기
: 학습모델 추론 최적화기
: 얼굴 학습 데이터 자동 생성기
: 객체 검출 영역 자동 생성기

2. 기술이전 내용 및 범위[3/7]

▣ 기술 적용 예시

교통기호 탐지



얼굴 인식

제스쳐 인식

피플 카운터

필기글자 인식

IR 표적 탐지

2. 기술이전 내용 및 범위[4/7]

▣기술개발현황

❖ 기술성숙도(Technology Readiness Level): (5)단계

구 분	단계	정 의	세 부 설 명
기초연구 단계	1	기초 이론/실험	◦기초이론 정립 단계
	2	실용목적의아이디어, 특허 등 개념정립	◦기술개발개념정립및아이디어에대한특허출원단계
실험 단계	3	실험실 규모의 기본성능 검증	◦실험실 환경에서 실험 또는 전산 시뮬레이션을 통해 기본성능이 검증될 수 있는 단계 ◦개발하려는부품/시스템의기본설계도면을확보하는단계
	4	실험실 규모의 소재/부품/시스템 핵심성능 평가	◦시험샘플을제작하여핵심성능에대한평가가완료된단계 ◦3단계에서 도출된 다양한 결과 중에서 최적의 결과를 선택하려는 단계 ◦컴퓨터 모사가 가능한 경우 최적화를 완료하는 단계
시작품 단계	5	확정된 소재/부품/시스템시작품제작 및 성능 평가	◦확정된 소재/부품/시스템의 실험실 시작품 제작 및 성능 평가가 완료된 단계 ◦개발 대상의 생산을 고려하여 설계하나 실제 제작한 시작품 샘플은 1~수개 미만인 단계 ◦경제성을 고려하지 않고 기술의 핵심성능으로만 볼 때, 실제 판매 가능한 정도로 목표 성능을 달성한 단계
	6	파일롯 규모 시작품 제작 및 성능 평가	◦파일롯 규모(복수 개~양산규모의 1/10정도)의 시작품 제작 및 평가가 완료된 단계 ◦파일롯규모생산품에대해생산량,생산용량,불량률등제시 ◦파일롯 생산을 위한 대규모 투자가 동반되는 단계 ◦생산기업이 수요기업 적용환경에 유사하게 자체 현장테스트를 실시하여 목표 성능을 만족시킨 단계 ◦성능평가결과에대해가능하면공인인증기관의성적서확보
실용화 단계	7	신뢰성평가 및 수요기업 평가	◦실제 환경에서 성능 검증이 이루어지는 단계 ◦부품 및 소재개발의 경우 수요업체에서 직접 파일롯 시작품을 현장 평가(성능 및 신뢰성 평가) ◦가능하면 인증기관의 신뢰성 평가 결과 제출
	8	시제품 인증 및 표준화	◦표준화 및 인허가 취득 단계
사업화	9	사업화	◦본격적인 양산 및 사업화 단계 ◦6-시그마 등 품질관리가 중요한 단계

▣ 기술별 제공 기술 목록[1] : 기술문서

문서관리번호	기술자료 명칭	관련기술		
		1.1	1.2	1.3
1240-2020-00660	온디바이스 머신러닝 프레임워크 V3.0 설치방법	O		
1240-2020-00661	온디바이스 머신러닝 프레임워크 V3.0 기반 비전 인식기 솔루션 구성		O	
1240-2020-00662	온디바이스 머신러닝 프레임워크 V3.0 기반 비전 인식기 솔루션 사용법		O	
1240-2019-02836	온디바이스 추론 가속 라이브러리 기반 객체 인식기 구조 및 설치 방법	O		
1240-2019-02837	온디바이스 추론 가속 라이브러리 사용자 API	O		
1240-2020-00658	온디바이스 머신러닝 프레임워크 활용을 위한 머신러닝 기본 개요	O		
1240-2020-00659	온디바이스 머신러닝 프레임워크 비전 인식 솔루션 사용자 API		O	
1240-2020-00663	온디바이스 머신러닝 프레임워크 Yolo 앵커 계산기 사용법			O
1240-2020-00664	온디바이스 머신러닝 프레임워크 모델 변환기 사용법			O
1240-2020-00739	얼굴 학습 데이터 자동 생성기 설명서			O

2. 기술이전 내용 및 범위[6/7]

▣ 기술별 제공 기술 목록[2] : 특허

특허 출원번호	특허 명칭	관련 기술		
		1.1	1.2	1.3
2020-0055313	콘볼루션 신경망 양자화 추론 장치 및 방법	O		O
2020-0072486	임베디드 기기에서의 딥러닝 기반 객체 인식 장치	O	O	
2019-0004805	이기종 온디바이스 시스템에서의 그래프 기반 영상 처리 모델 실행 최적화 장치 및 그 방법		O	

▣ 기술별 제공 기술 목록[3] : 프로그램

문서관리번호 (등록번호)	프로그램 명칭	관련기술		
		1.1	1.2	1.3
1240-2019-02185	정적 관심영역 기반 얼굴 인식기 라이브러리	O		
1240-2019-02186	정적 관심영역 기반 얼굴 인식기	O	O	
1240-2019-02197	동적 관심영역 기반 얼굴 인식기 라이브러리	O		
1240-2019-02198	동적 관심영역 기반 얼굴 인식기	O	O	
1240-2020-00670	비디오 입력 기반 피플 카운터		O	
1240-2020-00671	비디오 파일 입력 기반 교통 신호 인식기		O	
1240-2020-00706	딥러닝 얼굴 인식을 위한 얼굴 학습 데이터 자동 생성기			O
1240-2020-00730	모델 추론 최적화 및 객체 검출 보조 영역 자동 생성기			O

3. 경쟁기술과 비교[1/2]

▣ 기술의 특징

기술명	특징
스마트기기 비전인식용 온디바이스 딥러닝 SW 플랫폼 기술	– 본 기술은 스마트기기의 제한된 컴퓨팅 자원으로도 딥러닝 기술을 실행 가능하게 함으로, 저사양 스마트기기에서도 인공지능 SW 솔루션을 신속히 적용 가능하게 하여 지능형 스마트기기 개발 기간 및 비용을 절감시킬 수 있으므로 다양한 산업 분야에서 지능형 제품 개발을 가능하게 하는 핵심 기술을 제공함 – 향후 부가가치가 높은 지능형 스마트기기 산업에서 국내 기업들의 인공지능 기술 도입에 소요되는 기간과 인력을 단축시키도록 지원함으로써 적시성 (Time-to-Market) 높은 제품 출시를 가능하게 함 – 따라서, 성장절벽에 처한 스마트기기 시장에서 관련 국내 기업의 매출을 회복할 수 있는 4차 산업혁명시대에 필요한 지능형 스마트기기 신제품 개발을 가능하게 하는 핵심 기술 확보로 미래 시장 선점을 지원함 – 주요 기능 ➢ 딥러닝 처리 병렬화와 경량 화를 기반으로 임베디드 시스템 상에서 비전인식용 실시간 추론 처리를 지원하는 온디바이스 딥러닝 프레임워크 기술 제공 ➢ 사람 객체 수/사람 얼굴/사람제스처/한글 글자/교통기호/IR 객체 인식 솔루션 제공 ➢ 온디바이스 딥러닝 프레임워크기반의 비전인식 솔루션 개발 편의를 지원하는 유틸리티 제공

3. 경쟁기술과 비교[2/2]

▣ 기존 경쟁기술 대비 개량된 부분

❖ 기술적 측면

- 본 기술은 이러한 스마트기기에서 인공지능 SW 도입의 한계를 극복할 수 있도록 저사양 임베디드 시스템에 최적화된 딥러닝 SW 프레임워크와 현장의 학습 데이터로 개발한 다양한 비전 인식 솔루션을 제공함
- 특히, 본 기술은 ARM 및 Qualcomm SoC기반의 HW와 Embedded Linux 및 Android 기반의 OS 지원 가능하며, 관련 기능과 성능을 공인시험기관을 통해 검증 완료함(KTL/KOLAS 인증)

❖ 사업적 측면

- 기존 스마트기기에 추가적인 HW 없이 딥러닝 기반 비전인식 솔루션을 탑재 가능하게 함으로써 시스템 개발 기간 및 비용을 절감시켜 다양한 산업 분야에서 지능형 스마트기기 제품 경쟁력을 확보 가능하게 하는 핵심 기술을 제공함
- 스마트 자동차, 드론, 로봇 등에서 실시간 주변 객체 인식 솔루션을 제공할 뿐 아니라, HMD, 스마트 폰, 투명 디바이스 등 차세대 소형 디바이스에 탑재되어 지능형 증강/가상 현실(AR/VR), 혼합현실(MR), 홀로그램 서비스 분야에서 경쟁력 있게 활용할 수 있음
- 민감한 개인 및 기업 정보를 활용한 온디바이스 학습을 통해 개인 및 기업정보 보호와 동시에 고품질의 개인 맞춤형 지능형 서비스를 제공할 수 있음

▣예상 제품(서비스) 및 사업 조건

❖ 예상 응용 제품 및 서비스

- (예상 수요자) 임베디드 디바이스 상에서의 인공 지능 서비스 제공 업체
  - ➢ 인지형 모바일기기, 자율이동체, 지능형 로봇, 지능형 영상감시기기, 지능형 IoT 디바이스
- (제품) 지능형 산업기기, 지능형 모바일기기, 지능형 영상감시기기, 지능형 국방 내장형 무기체계, 지능형 스마트 가전, 지능형 IoT 디바이스
- (활용분야) 다양한 산업분야에서 활용되는 임베디드 기기에 탑재되어 다양한 인공지능 서비스를 제공하거나 무인화 및 정밀화를 가능하게 하는 신제품을 개발할 수 있음

❖ 사업성

- 인공지능기반 비전인식 기술을 임베디드기기에 탑재 가능하게 하는 기술로 스마트기기 산업에 적용하여 연평균 성장률 30% 이상으로 급속히 증가하는 지능형 스마트기기 시장에서 사업성이 매우 높음

❖ 기술이전 업체 조건

- 임베디드 SW 및 HW 제품 개발 또는 딥러닝기반 비전인식 기술 경험이 있는 업체
- 2인 이상의 소프트웨어 엔지니어를 보유하고 있는 업체

❖ 사업화시 제약 조건

- 해당사항 없음

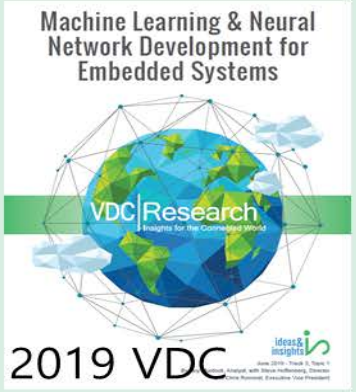
5. 국내외 시장 동향[1/4]

▣관련 제품/서비스 시장 동향

- 개발 기술이 적용될 디바이스 지능형 비전처리 세계 목표시장 규모는 '24년 428억 달러에서 '29년 688억 달러 규모로 연평균 10%의 성장이 전망되며, 국내 목표시장은 '24년 6,823억원에서 '29년 8,315억원 규모로 연평균 4% 성장 전망(출처: BCC Research('14), Frost & Sullivan('16), Gartner('16), IDC('15), Gartner('15), Strategy Analysis('14), 국방기술품질원('14))
- 인공지능 스마트 기기 시장은 2024년 기준 각 분야별로 자동 로봇(139억$), 디지털(지능형) 보조(80억$), 임베디드(20억$) 순으로 전망됨

5. 국내외 시장 동향[2/4]

최신 글로벌 시장동향 보고서에 따르면, 본 과제 핵심기술인 온디바이스 인공지능 SW 기술 시장은 고성장



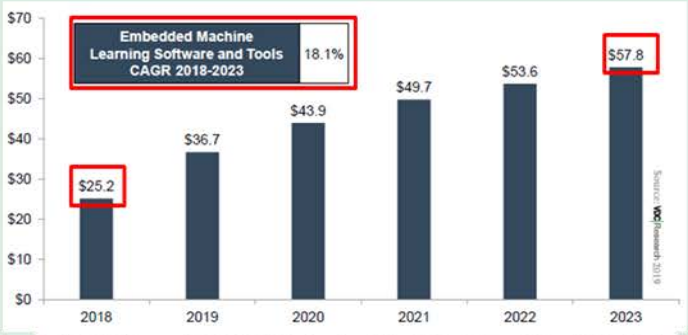


Exhibit 1: Global Revenue of Embedded Machine Learning and Neural Network Software, Tools, and Related Services (Millions of Dollars)

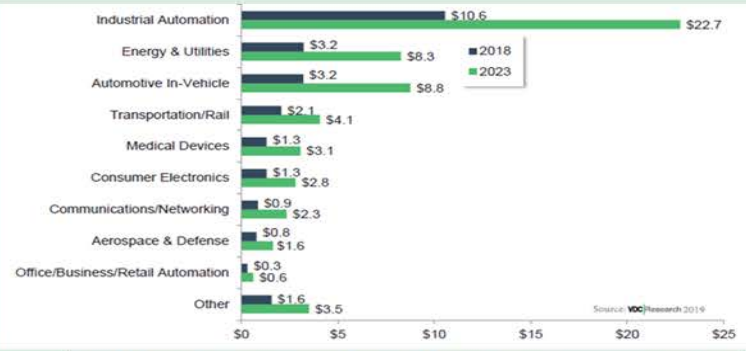
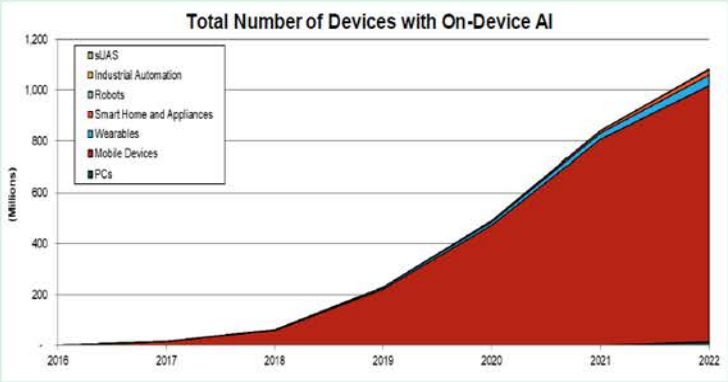


Exhibit 3A: Global Revenue of Embedded Machine Learning Software and Tools by Vertical Market (Millions of Dollars)



- By 2022, 26% of active AI devices will have embedded AI, totaling 1 billion devices, at a CAGR of 135%. (In 2017, 15 million devices)
- The on-device AI-capable segment will mainly be driven by mobile devices; the wave of AI enabling SoCs, frameworks, and software engines important role

5. 국내외 시장 동향[3/4]

▣관련 제품/서비스의 민수분야 관련 세계시장

구분		2014	2015	2016	2017	2018	2019	2020	CAGR ('14-'20)
머신비전(imaging SW)		1,980	2,160	2,367	2,595	2,844	3,117	3,410	
스마트 기기	스마트폰	6,176	6,786	7,679	8,599	9,485	10,407	10,973	
	웨어러블	10	17	120	582	1,117	1,890	2,808	
자율주행(ADAS)		44	57	69	83	105	123	152	
합계		8,210	9,021	10,235	11,858	13,550	15,537	17,342	13.3%

출처: BCC Research(2014), Frost & Sullivan(2016), Gartner(2016), IDC(2015), Giantt(2014)

▣관련 제품/서비스의 군수분야 관련 세계시장

구분		2014	2015	2016	2017	2018	2019	2020	CAGR ('14-'20)
감시 정찰	감시정찰-전자광학/적외선	2,456	2,400	2,408	2,363	2,425	2,247	2,050	
	전자전	2,587	2,471	3,583	3,332	3,289	3,865	3,149	
기동	무인체계(로봇무인체계)	85	75	76	75	92	112	129	
항공	무인기	775	1,110	1,241	1,273	1,446	1,487	1,682	
화력	정밀유도무기	6,099	6,542	6,961	7,010	6,937	7,228	7,286	
	합계	12,002	12,598	14,269	14,053	14,189	14,939	14,296	17.5%

출처 : 국방기술품질원(2014)

5. 국내외 시장 동향[4/4]

▣관련 제품/서비스의 국내외 예상 매출액

- ❖ 본 기술을 통해 '21년까지 지능형 스마트기기 분야에서 기술 경쟁력을 향상시켜 관련 민수 및 군수 국내시장 562억원 달성 및 국외 1249.1억원을 선점할 수 있을 것으로 예상됨
- ❖ 산출근거
  - 전체 전세계 지능형스마트 기기의 20% 규모를 국내 시장으로 가정하고, 2020년부터 상용화를 시작하여 매년 국외는 1%, 3%, 8% 그리고 국내시장은 5%, 8%,12% 점유한다고 예상함

감사합니다.

(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년05월25일
(11) 등록번호 10-2255365
(24) 등록일자 2021년05월17일

(51) 국제특허분류(Int. Cl.)
G06T 1/20 (2018.01) *G06F 15/78* (2006.01)
(52) CPC특허분류
G06T 1/20 (2013.01)
G06F 15/7807 (2013.01)
(21) 출원번호 10-2019-0004805
(22) 출원일자 2019년01월14일
심사청구일자 2019년05월17일
(65) 공개번호 10-2020-0088155
(43) 공개일자 2020년07월22일
(56) 선행기술조사문헌
KR101600231 B1*
KR1020160090919 A*
KR1020170046784 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
이문수
세종특별자치시 달빛1로 206, 907동 602호
김정시
대전광역시 유성구 도안동로 523, 202동 701호
(뒷면에 계속)
(74) 대리인
특허법인지명

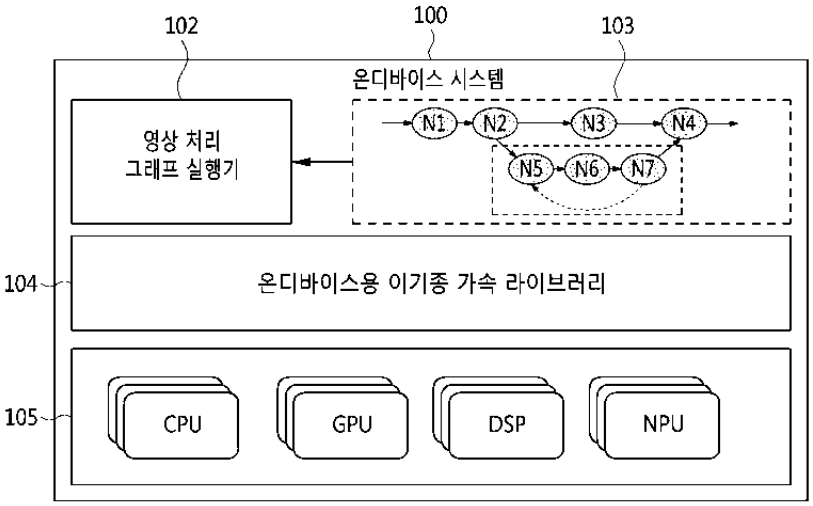
전체 청구항 수 : 총 8 항 심사관 : 김병성

(54) 발명의 명칭 **이기종 온디바이스 시스템에서의 그래프 기반 영상 처리 모델 실행 최적화 장치 및 그 방법**

(57) 요 약

본 발명의 일 실시예에 따른 이기종 온디바이스 시스템에서의 그래프 기반 영상 처리 모델 실행 최적화 장치는 사용자가 제공하는 영상 처리 그래프 모델에 대해 그래프 토폴러지(topology)를 이용하여 경로를 탐색하고, 실행 순서에 따라 노드들을 생성 및 순차 배치하여 영상 처리 그래프 모델의 실행을 위한 실행 경로를 생성하는 영상 처리 그래프 경로 생성부; 및 영상 처리 그래프 경로 생성부로부터 수신된 영상 처리 그래프 모델, 노드들에 대응되는 노드 목록 및 가상 슈퍼 노드 목록, 실행 경로에 기초하여 영상 처리 그래프 모델을 실행하는 영상 처리 그래프 경로 실행부;를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06T 2200/28 (2013.01)
(72) 발명자
배수영
대전광역시 유성구 반석동로 33, 501동 2202호
석종수
대전광역시 유성구 온천로 26, 1424호
정영준
세종특별자치시 새롬북로 13, 410동 2606호

이 발명을 지원한 국가연구개발사업

과제고유번호	2017-0-00142
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기술진흥센터
연구사업명	SW컴퓨팅산업원천기술개발
연구과제명	스마트기기를 위한 온디바이스 지능형 정보처리 가속화 SW플랫폼 기술 개발
기 여 율	1/1
과제수행기관명	한국전자통신연구원
연구기간	2018.01.01 ~ 2018.12.31

명 세 서

청구범위

청구항 1

삭제

청구항 2

사용자가 제공하는 영상 처리 그래프 모델에 대해 그래프 토폴러지(topology)를 이용하여 경로를 탐색하고, 상기 영상 처리 그래프 모델에 포함된 영상 처리 기능 단위에 대응되는 노드들을 생성하고, 상기 노드들을 실행 순서에 따라 순차 배치하여 상기 영상 처리 그래프 모델의 실행을 위한 실행 경로를 생성하는 영상 처리 그래프 경로 생성부; 및

상기 영상 처리 그래프 경로 생성부로부터 수신된 상기 영상 처리 그래프 모델, 상기 노드들에 대응되는 노드 목록, 상기 실행 경로를 이용하여 상기 영상 처리 그래프 모델을 실행하는 영상 처리 그래프 경로 실행부

를 포함하고,

상기 영상 처리 그래프 경로 생성부는

상기 노드들을 생성한 뒤, 상기 영상 처리 그래프 모델 내 순환 그래프가 존재하는 경우, 상기 순환 그래프에 대응되는 순환 노드들을 가상 슈퍼 노드로 변환하고, 상기 노드들 및 상기 가상 슈퍼 노드를 실행 순서에 따라 순차 배치하여 상기 실행 경로를 생성하고,

상기 영상 처리 그래프 경로 실행부는

상기 가상 슈퍼 노드에 대응되는 가상 슈퍼 노드 목록을 더 이용하여 상기 영상 처리 그래프 모델을 실행하는 것인, 이기종 온디바이스 시스템에서의 그래프 기반 영상 처리 모델 실행 최적화 장치.

청구항 3

청구항 2에 있어서,

상기 영상 처리 그래프 경로 생성부는

상기 영상 처리 그래프 모델 내 복수의 순환 그래프들이 존재하는 경우, 상기 순환 그래프들 각각에 대응되는 순환 노드들을 상기 가상 슈퍼 노드로 변환하는 것인, 이기종 온디바이스 시스템에서의 그래프 기반 영상 처리 모델 실행 최적화 장치.

청구항 4

청구항 3에 있어서,

상기 영상 처리 그래프 경로 생성부는

상기 순환 노드들 중 시작 노드 및 종료 노드를 결정하고, 상기 순환 노드들 및 상기 노드들 중 상기 순환 노드들을 제외한 나머지 노드들 간의 입출력 데이터를 상기 가상 슈퍼 노드의 입출력 파라미터로 결정하고, 상기 가상 슈퍼 노드의 순환 조건을 결정하고, 상기 영상 처리 그래프 모델에 포함된 상기 순환 노드들을 상기 가상 슈퍼 노드로 대치하는 것인, 이기종 온디바이스 시스템에서의 그래프 기반 영상 처리 모델 실행 최적화 장치.

청구항 5

청구항 4에 있어서,

상기 이기종 온디바이스 시스템은

복수의 온디바이스용 이기종 라이브러리들 및 복수의 이기종 컴퓨팅 자원들을 포함하고,

상기 복수의 온디바이스용 이기종 라이브러리들 각각은

복수의 이기종 컴퓨팅 자원들 중 적어도 하나를 이용하여 상기 영상 처리 기능 단위를 제공하는 소프트웨어 모듈로서 상기 노드들 각각과 일대일 대응되는 것인, 이기종 온디바이스 시스템에서의 그래프 기반 영상 처리 모델 실행 최적화 장치.

청구항 6

삭제

청구항 7

사용자가 제공하는 영상 처리 그래프 모델에 대해 그래프 토폴러지(topology)를 이용하여 경로를 탐색하는 단계;

상기 영상 처리 그래프 모델에 포함된 영상 처리 기능 단위에 대응되는 노드들을 생성하는 단계;

상기 노드들을 실행 순서에 따라 순차 배치하여 상기 영상 처리 그래프 모델의 실행을 위한 실행 경로를 생성하는 단계; 및

상기 영상 처리 그래프 모델, 상기 노드들에 대응되는 노드 목록, 상기 실행 경로를 이용하여 상기 영상 처리 그래프 모델을 실행하는 단계

를 포함하고,

상기 노드들을 생성하는 단계 이후,

상기 영상 처리 그래프 모델 내 순환 그래프가 존재하는 경우, 상기 순환 그래프에 대응되는 순환 노드들을 가상 슈퍼 노드로 변환하는 단계;

를 더 포함하고,

상기 실행 경로를 생성하는 단계는

상기 노드들 및 상기 가상 슈퍼 노드를 실행 순서에 따라 순차 배치하여 상기 실행 경로를 생성하고,

상기 영상 처리 그래프 모델을 실행하는 단계는

상기 가상 슈퍼 노드에 대응되는 가상 슈퍼 노드 목록을 더 이용하여 상기 영상 처리 그래프 모델을 실행하는 것인, 이기종 온디바이스 시스템에서의 그래프 기반 영상 처리 모델 실행 최적화 방법.

청구항 8

청구항 7에 있어서,

상기 순환 노드들을 가상 슈퍼 노드로 변환하는 단계는

상기 영상 처리 그래프 모델 내 복수의 순환 그래프들이 존재하는 경우, 상기 순환 그래프들 각각에 대응되는 순환 노드들을 상기 가상 슈퍼 노드로 변환하는 것인, 이기종 온디바이스 시스템에서의 그래프 기반 영상 처리 모델 실행 최적화 방법.

청구항 9

청구항 8에 있어서,

상기 순환 노드들을 가상 슈퍼 노드로 변환하는 단계는

상기 순환 노드들 중 시작 노드 및 종료 노드를 결정하는 단계;

상기 순환 노드들 및 상기 노드들 중 상기 순환 노드들을 제외한 나머지 노드들 간의 입출력 데이터를 상기 가상 슈퍼 노드의 입출력 파라미터로 결정하는 단계;

상기 가상 슈퍼 노드의 순환 조건을 결정하는 단계; 및

상기 영상 처리 그래프 모델에 포함된 상기 순환 노드들을 상기 가상 슈퍼 노드로 대치하는 단계;

를 포함하는 것인, 이기종 온디바이스 시스템에서의 그래프 기반 영상 처리 모델 실행 최적화 방법.

청구항 10

청구항 9에 있어서,

상기 이기종 온디바이스 시스템은

복수의 온디바이스용 이기종 라이브러리들 및 복수의 이기종 컴퓨팅 자원들을 포함하고,

상기 복수의 온디바이스용 이기종 라이브러리들 각각은

복수의 이기종 컴퓨팅 자원들 중 적어도 하나를 이용하여 상기 영상 처리 기능 단위를 제공하는 소프트웨어 모듈로서 상기 노드들 각각과 일대일 대응되는 것인, 이기종 온디바이스 시스템에서의 그래프 기반 영상 처리 모델 실행 최적화 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 이기종 온디바이스(on-device) 시스템에서 중앙 처리 장치(central processing unit; CPU)뿐만 아니라 그래픽 처리 장치(graphics processing unit; GPU), 디지털 신호 처리(digital signal processing) 장치, 신경망 처리 장치(neural processing unit; NPU) 등 다양한 컴퓨팅 자원 기반의 소프트웨어 라이브러리 조합으로 구성된 그래프 기반 영상 처리 모델을 효율적으로 실행하는 시스템에 관한 것이다.

배 경 기 술

[0002] 최근 머신 러닝 기술의 급속한 성장으로 인해 영상 인식 서비스가 보편화되면서 스마트폰과 같이 온디바이스 시스템에서도 컴퓨터 비전 처리 기능 탑재에 대한 수요가 급증하고 있다.

[0003] 이에 따라 칩 제조사들은 임베디드(Embedded) 시스템 특성상 소모 전력을 최소화하면서 동시 성능을 높이기 위해 GPU, DSP 장치, NPU(Neural Processing Unit), 필드 프로그래머블 게이트 어레이(Field Programmable Gate Array; FPGA) 등 다양한 전용 시스템 온 칩(System on Chip; SoC)들을 탑재하고 있다.

[0004] 이들 전용 칩들은 미디어 인코더/디코더와 같이 초기에 특정 응용 프로그램이나 서비스에 특화되어 최적의 성능을 얻기 위해 사용되었지만, 점차 다른 응용 프로그램이나 다른 서비스에서도 이러한 컴퓨팅 리소스를 활용하는 이기종 컴퓨팅(Heterogeneous Computing) 기술이 필요해 지고 있다.

[0005] 이를 위해 국제 표준 단체인 크로노스 그룹에서는 OpenVX 표준을 통해 영상 처리에 필요한 기본 영상 처리 기능들을 소프트웨어 라이브러리 형태로 제공될 수 있도록 세분화하여 정의하였으나, 일반적으로 세분화된 라이브러리와 그래픽 기반 응용 모델링 기법은 복잡한 영상 처리 응용을 개발할 경우 다음과 같은 문제점이 있다.

[0006] 첫 번째로 라이브러리의 세분화는 그래픽 기반의 개발 환경으로 응용을 쉽게 구현할 수 있는 장점이 있다. 하지만 이를 위해서는 각 처리 모듈에 대한 인터페이스가 단순하게 되어 복합적인 입력 데이터를 처리하기 위해서는 그래픽 기반 응용 모델러를 통해 다수개의 영상처리 노드들과 파이프라인 형태의 노드 별 실행 흐름을 통해 정의하게 된다. 따라서, 실행 모델은 영상 처리를 위한 각 단위 기능이 노드 별로 구성되고, 이 노드들은 실행 순서에 맞게 순차적으로 실행하게 된다. 이러한 응용 모델로부터 자동 생성된 코드들은 목표로 하는 기능을 수행하기 위해 간략한 구조로 표현된 코드로서 실행하는데 문제는 없으나 성능을 높이거나 최적화하기 위해서는 여전히 개발자가 직접 코드를 추가로 수정해야 하는 문제가 있다

[0007] 두 번째로 파이프라인 형태의 실행 모델에서 일부 구간에서 동일 영상 처리를 반복 처리가 필요할 경우 그래픽 기반 모델링에서는 이들을 분리하여 별도의 하위 그래프로 표현하게 된다. 각 하위 그래프들은 자신의 그래프 관리 영역 내에서 리소스를 생성하고 해제를 하게 되는데, 상위 그래프가 반복되는 경우 매번 리소스를 생성 및 해제해야 하는 문제가 발생한다. 또한, GPU와 같은 병렬 처리 컴퓨팅 리소스의 경우 실행 태스크마다 GPU가 사용하는 메모리들을 생성 초기화(initialize) 및 종료(deinitialize)가 필요한데 이로 인해 성능 저하가 발생할 수 있다. 예컨대, GPU 기반 오픈 컴퓨팅 랭귀지(open computing language; OpenCL)용 라이브러리의 경우 라이브러리 실행할 때 마다 OpenCL 커널(kernel) 소스 코드를 로딩하여 동적 컴파일을 수행하고, 해당 커널에 필요한 GPU를 위한 메모리 할당 및 초기화를 수행해야 한다.

[0008] 마지막으로 영상 추적이나 정확도를 높이기 위해서 이전 처리 결과에 대한 피드백을 가지는 순환(cyclic) 구조의 그래프가 많이 필요로 하게 되지만 그래픽 기반 모델의 경우 무한 루프 구조가 발생 가능성이 있어 순환 구

조의 그래프 사용이 제한되는 문제가 있다.

선행기술문헌

특허문헌

[0009] (특허문헌 0001) 한국 공개 특허 제10-2017-0115185호, 2017년 10월 17일 공개(명칭: 소프트웨어 빌드 모듈을 포함하는 임베디드 시스템)

발명의 내용

해결하려는 과제

[0010] 본 발명은 상술한 문제점을 해결하기 위하여, 다양한 이기종 컴퓨팅 자원이 탑재된 온디바이스 시스템 환경에서 독립적인 하위 그래프나 순환 그래프가 포함된 그래픽 기반 응용 모델에 대한 성능을 태스크 레벨에서 최적화하기 위한 장치 및 방법을 제공하는 것을 목적으로 한다.

과제의 해결 수단

[0011] 상기한 목적을 달성하기 위한 본 발명의 일 실시예에 따른 이기종 온디바이스 시스템에서의 그래프 기반 영상 처리 모델 실행 최적화 장치는 사용자가 제공하는 영상 처리 그래프 모델에 대해 그래프 토폴러지(topology)를 이용하여 경로를 탐색하고, 실행 순서에 따라 노드들을 생성 및 순차 배치하여 영상 처리 그래프 모델의 실행을 위한 실행 경로를 생성하는 영상 처리 그래프 경로 생성부; 및 영상 처리 그래프 경로 생성부로부터 수신된 영상 처리 그래프 모델, 노드들에 대응되는 노드 목록 및 가상 슈퍼 노드 목록, 실행 경로에 기초하여 영상 처리 그래프 모델을 실행하는 영상 처리 그래프 경로 실행부;를 포함한다.

[0012] 또한, 상기한 목적을 달성하기 위한 본 발명의 일 실시예에 따른 이기종 온디바이스 시스템에서의 그래프 기반 영상 처리 모델 실행 최적화 방법은, 영상처리 그래프 경로 생성부가 사용자가 작성한 영상 처리 그래프 모델에서 노드 별 입출력 데이터를 기반으로 위상정렬(토폴러지 정렬)하여 실행 가능한 경로 후보 목록들을 생성하고, 영상 처리 그래프 모델 내에 순환 하위 그래프가 존재하는지 여부를 확인하고, 영상 처리 그래프 모델 내에 순환 하위 그래프가 존재하는 경우, 순환 구조 내부의 시작 노드와 종료 노드를 식별하고 하나의 가상 슈퍼 노드에 대응되는 블록(노드들)을 결정하고, 가상 슈퍼 노드 내의 외부와 연결되는 입출력 데이터를 자동 탐색하고, 이들에 대해 입출력 데이터를 가상 슈퍼 노드의 입출력 파라미터로 등록하고, 가상 슈퍼 노드가 별도의 쓰레드를 통해 동작되도록 실행 반복 회수 및/또는 시작/종료 시점을 결정하고, 가상 슈퍼 노드의 입출력 연결 정보를 통해 영상 처리 그래프 모델(300) 내에 새로운 노드를 생성하여 추가하고, 영상처리 그래프 경로 실행부가 영상 처리 그래프 모델을 실행한다.

발명의 효과

[0013] 본 발명에 따르면, 다양한 이기종 컴퓨팅 자원을 가지고 있는 온디바이스 시스템 상에서 독립적인 하위 그래프나 순환 그래프가 포함된 파이프라인 형태의 영상 처리 모델을 가상 슈퍼 노드와 그 실행 매커니즘을 제공함으로써 피드백 형태의 영상 처리 응용 프로그램의 개발이 가능하도록 기능을 확장할 수 있고, 응용 프로그램 내의 쓰레드(thread)나 프로세스들을 자동 생성 및 노드 간의 데이터 동기화를 통해 라이브러리들 간 응답 지연 시간을 감소시킬 수 있다.

도면의 간단한 설명

[0014] 도 1은 본 발명의 일 실시예에 따른 이기종 온디바이스 시스템을 나타낸 도면이다.

도 2는 본 발명의 일 실시예에 따른 영상 처리 그래프 실행기의 구성을 나타낸 도면이다.

도 3a 내지 3c는 본 발명의 일 실시예에 따른 영상 처리 그래프 모델들을 나타낸 도면이다.

도 4는 도 2에 도시된 영상 처리 그래프 경로 생성부의 가상 슈퍼 노드를 생성하는 방법의 일 예를 나타낸 동작 흐름도이다.

도 5는 본 발명의 일 실시예에 따른 가상 슈퍼 노드를 나타낸 도면이다.

도 6은 본 발명의 일 실시예에 따른 컴퓨터 시스템을 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0015] 본 발명을 첨부된 도면을 참조하여 상세히 설명하면 다음과 같다. 여기서, 반복되는 설명, 본 발명의 요지를 불필요하게 흐릴 수 있는 공지 기능, 및 구성에 대한 상세한 설명은 생략한다. 본 발명의 실시형태는 당 업계에서 평균적인 지식을 가진 자에게 본 발명을 보다 완전하게 설명하기 위해서 제공되는 것이다. 따라서, 도면에서의 요소들의 형상 및 크기 등은 보다 명확한 설명을 위해 과장될 수 있다.

[0016] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성 요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0017] 이하, 본 발명에 따른 바람직한 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0018] 도 1은 본 발명의 일 실시예에 따른 이기종 온디바이스 시스템을 나타낸 도면이다.

[0019] 특히, 도 1은 그래프 기반 모델을 실행해주는 다양한 이기종 컴퓨팅 장치를 포함하고 있는 온디바이스 시스템의 일 예를 나타낸 것이다.

[0020] 도 1을 참조하면, 온디바이스 시스템(100)은 영상, 센서 데이터, 실시간 데이터 등 스트림 형태로 입력되는 데이터들을 그래프 기반 실행 모델(103)을 기반으로 시스템의 다양한 이기종 컴퓨팅 자원(105)을 활용할 수 있도록 한다.

[0021] 온디바이스 시스템(100)은 그래프 기반 실행 모델(103)에 포함된 각 단위 영상 처리 기능을 구현한 온디바이스용 이기종 라이브러리(104)와 이를 실제 수행하기 위한 이기종 컴퓨팅 자원(105), 사용자가 원하는 타겟 서비스를 기능 중심의 시나리오로 정의된 그래프 기반 실행 모델(103)과 이 실행 모델을 해석하고 온디바이스용 이기종 라이브러리(104)를 호출 실행해 주는 영상 처리 그래프 실행기(102)로 구성된다.

[0022] 이기종 컴퓨팅 자원(105)은 일반적으로 사용되는 컴퓨팅 시스템에 존재하는 멀티 코어 CPU와 GPU뿐만 아니라, 저전력 DSP 장치, 인공지능을 위한 NPU, 영상 처리에 사용되는 전용 코덱 등 특수 목적의 시스템 온 칩(SoC)을 포함할 수 있다.

[0023] 온디바이스용 이기종 라이브러리(104)는 온디바이스 시스템(100)에 탑재되어 있는 이기종 컴퓨팅 자원(105)을 최소 한 개 이상 이용하여 데이터를 처리하기 위한 하나의 소프트웨어 모듈이다.

[0024] 여기서 온디바이스용 이기종 라이브러리(104)는 단위 데이터 처리 기능을 제공하고, 그래프 기반 실행 모델(103)내의 노드와 1:1로 매칭된다. 또한, 이기종 라이브러리는 동일 기능을 가지더라도, 해당 기능을 수행하기 위해 이용하는 자원 종류에 따라 서로 다른 것으로 구별될 수 있다. 예컨대, CPU 자원만을 이용하는 라이브러리가 존재할 수 있고, GPU나 DSP 자원만을 이용하는 라이브러리가 존재할 수 있는데, 본 발명에서는 이들을 각각 다른 모듈로서 구별할 수 있다.

[0025] 그래프 기반 실행 모델(103)은 노드(node)와 에지(edge)로 구성된다. 노드는 데이터를 처리하기 위한 단위 기능에 해당되고, 에지는 다음 단계의 데이터 처리를 알려주는 연결 정보이다. 일반적인 그래프 기반 실행 모델은 무한 반복을 피하기 위해 순환(cyclic) 구조를 제한하지만 본 발명에서는 순환 구조를 포함한다.

[0026] 영상 처리 그래프 실행기(102)는 사용자가 정의한 그래프 기반 실행 모델(103)을 해석하여 온디바이스 시스템(100)이 가지고 있는 이기종 컴퓨팅 자원(105)으로 최적의 실행이 가능하도록 하는, 가장 효율적인 경로를 생성하고 실행 스케줄을 생성한다.

[0027] 도 2는 본 발명의 일 실시예에 따른 영상 처리 그래프 실행기의 구성을 나타낸 도면이다.

[0028] 도 2를 참조하면, 사용자는 자신이 원하는 영상 처리 순서를 하나의 시나리오 형태로 영상 처리 그래프 모델(201)을 생성한다.

[0029] 영상 처리 그래프 실행기(200)는 사용자가 제공하는 영상 처리 그래프 모델(201)을 해석하여, 이를 효과적으로 실행하기 위한 경로를 생성하고 이를 실행하게 해 준다.

[0030] 영상 처리 그래프 실행기(200)는 영상 처리 그래프 경로 생성부(202)와 생성된 경로를 기반으로 온디바이스용 이기종 라이브러리(도 1의 104 참조) 내 해당 라이브러리를 동적으로 호출해주는 영상 처리 그래프 경로 실행부

(203)로 구성된다.

[0031] 영상 처리 그래프 경로 생성부(202)는 그래프 토폴러지(topology)를 이용하여 경로를 탐색하고, 실행 순서에 따라 노드(204)들을 순차적으로 배치하여 실행 경로(207)를 생성한다.

[0032] 여기서, 영상 처리 그래프 경로 생성부(202)는 그래프 내에 순환(cyclic) 구조가 있는지 확인하고, 이에 대한 정보를 추출하여 가상 슈퍼 노드(205) 후보로 결정한다.

[0033] 노드 실행 추적 관리부(206)는 영상 처리 그래프 경로 생성부(202)에서 생성한 경로 순서에 따라 노드 별 해당되는 라이브러리를 호출해준다.

[0034] 도 3a 내지 3c는 본 발명의 일 실시예에 따른 영상 처리 그래프 모델들을 나타낸 도면이다.

[0035] 특히, 도 3a 내지 3c는 사용자가 영상 처리 응용 프로그램을 생성하기 위해 정의한 영상 처리 그래프 모델(201)의 예들을 나타낸 것이다.

[0036] 도 3a를 참조하면, 영상 처리 그래프 모델(300)은 2개의 입력 데이터 IN1(301)과 IN2(302), 영상 처리함수(N1 내지 N7) 및 최종 출력 데이터 OUT(303)으로 구성된다.

[0037] 여기서, 영상 처리 그래프 모델(300)은 N5, N6 및 N7 노드들로 구성된 하위 그래프(304)를 포함할 수 있다.

[0038] 도 3b를 참조하면, 영상 처리 그래프 모델(305)은 N5, N6 및 N7 노드들이 N5-N6-N7-N5 순서로 순환되는 상호 순환 구조(306)를 포함할 수 있다.

[0039] 도 3c를 참조하면, 영상 처리 그래프 모델(307)은 N8 노드와 같이 가상 슈퍼 노드(308)를 포함할 수 있다.

[0040] 여기서, 영상 처리 그래프 모델(307)은 도 3a 및 3b에 도시된 영상 처리 그래프 모델들(300 및 305)에서 N5, N6 및 N7 노드들을 하나의 가상 슈퍼 노드(308)로 대치하여 생성된 것일 수 있다.

[0041] 영상 처리 그래프 경로 생성부(도 2의 202 참조)는 상호 순환 구조가 없는 영상 처리 그래프 모델(307)과 노드(도 2의 204 참조) 목록, 가상 슈퍼 노드(도 2의 205 참조) 목록 및 실행 경로(도 2의 207 참조) 정보를 영상 처리 그래프 경로 실행부(도 2의 203 참조)에 전달할 수 있다.

[0042] 도 4는 도 2에 도시된 영상 처리 그래프 경로 생성부(202)의 가상 슈퍼 노드를 생성하는 방법의 일 예를 나타낸 동작 흐름도이다.

[0043] 도 4를 참조하면, 먼저 영상 처리 그래프 경로 생성부(202)가 사용자가 작성한 영상 처리 그래프 모델(도 3의 300 참조)에서 노드 별 입출력 데이터를 기반으로 위상정렬(토폴러지 정렬)하여 실행 가능한 경로 후보 목록들을 생성한다(S400).

[0044] 다음으로, 영상 처리 그래프 경로 생성부(202)가 영상 처리 그래프 모델(300)내에 순환 하위 그래프가 존재하는지 여부를 확인한다(S401).

[0045] 단계(S401)의 판단 결과, 영상 처리 그래프 모델(도 3의 300 참조) 내에 순환 하위 그래프가 존재하는 경우, 영상 처리 그래프 경로 생성부(202)가 순환 구조 내부의 시작 노드와 종료 노드를 식별하고 하나의 가상 슈퍼 노드에 대응되는 블록(노드들)을 결정한다(S402).

[0046] 다음으로, 영상 처리 그래프 경로 생성부(202)가 가상 슈퍼 노드 내의 외부와 연결되는 입출력 데이터를 자동 탐색하고, 이들에 대해 입출력 데이터를 가상 슈퍼 노드의 입출력 파라미터로 등록한다(S403).

[0047] 다음으로, 영상 처리 그래프 경로 생성부(202)가, 가상 슈퍼 노드가 별도의 쓰레드를 통해 동작되도록 실행 반복 회수 및/또는 시작/종료 시점을 결정한다(S404).

[0048] 여기서, 실행 반복 회수 또는 시작/종료 시점은 가상 슈퍼 노드 외부의 입력 데이터를 이용하여 결정될 수도 있고, 사용자가 지정한 파라미터를 이용하여 결정될 수도 있다.

[0049] 다음으로, 영상 처리 그래프 경로 생성부(202)가 가상 슈퍼 노드의 입출력 연결 정보를 통해 영상 처리 그래프 모델(도 3의 300 참조) 내에 새로운 노드를 생성하여 추가한다(S405).

[0050] 단계(S401) 내지 단계(S405)는 영상 처리 그래프 모델(도 3의 300 참조) 내에 순환 그래프가 존재하지 않을 때까지 반복 수행될 수 있다.

[0051] 단계(S401)의 판단 결과, 영상 처리 그래프 모델(도 3의 300 참조) 내에 순환 그래프가 존재하지 않는 경우, 영

상 처리 그래프 경로 생성부(도 2의 202 참조)가 단계(S401) 내지 단계(S405)를 통해 추가 생성된 가상 슈퍼 노드들이 메인 그래프와 독립적으로 실행될 수 있도록 내부적으로 별도의 쓰레드나 프로세스를 생성하고, 종료 노드 수행이 완료되면 해당 동기화 신호를 어플리케이션 실행기(도 2의 200 참조)에 전달한다(S406).

[0052] 도 5는 본 발명의 일 실시예에 따른 가상 슈퍼 노드를 나타낸 도면이다.

[0053] 도 5를 참조하면, 가상 슈퍼 노드(500)는 일반 노드와 동일한 입출력 인터페이스(501), 가상 슈퍼 노드(500) 내의 하위 그래프(506)를 실행하기 위한 순환 그래프 실행 엔진(508), 반복 실행 종료 조건 관리부(505) 및 반복 실행 종료 후 상위의 어플리케이션 실행기(도 2의 200 참조)에 노드 종료 동기화 신호를 제공하는 동기화 신호 생성부(507)로 구성된다.

[0054] 비록 도 5에는 도시되지 아니하였으나, 본 발명에 따른 가상 슈퍼 노드(500)의 입출력 인터페이스(501)는 노드 초기화 처리 인터페이스(502), 노드 실행 처리 인터페이스(503) 및 노드 종료 처리 인터페이스(504)로 제한되는 것은 아니며, 필요에 따라 다른 기능을 포함하는 입출력 인터페이스가 더 추가될 수 있다.

[0055] 노드 초기화 처리 인터페이스(502)는 노드를 실행하기에 앞서 변수 초기화, 메모리 초기화 등 노드 실행 처리 인터페이스(503)를 통한 기능 수행에 필요한 사전 조건들을 초기화 하는 기능을 포함하는 것일 수 있다. 또한, 온디바이스 시스템(도 1의 100 참조) 내의 이기종 컴퓨팅 자원(도 1의 105 참조)를 사용하기 위한 초기화 작업을 포함할 수 있다. 예컨대, 노드 실행 처리에 GPU가 사용되는 경우, 실시간 성능을 높이기 위해 노드 초기화 처리 인터페이스(502)에 OpenCL이나 쿠다(compute unified device architecture; CUDA) 라이브러리를 사용하기 위한 커널 소스를 컴파일하는 작업이 포함될 수 있다.

[0056] 노드 종료 처리 인터페이스(504)는 메인 그래프 실행이 완료되면 그에 따라 가상 슈퍼 노드(500)에 사용된 리소스를 반환/해제하는 기능을 포함할 수 있다.

[0057] 노드 실행 처리 인터페이스(503)는 노드들을 순서대로 실행하는 기능을 포함하며, 실행 과정에서 가상 슈퍼 노드의 실행이 호출되면, 내부의 순환 그래프 실행 엔진(508)이 분리된 하위 그래프(506)의 하위 노드들을 실행하도록 할 수 있다.

[0058] 반복 실행 종료 조건 관리부(505)는 순환 그래프 실행 엔진(508)이 하위 그래프 노드을 이용해 가며 실행하게 되는데, 이들을 실행 중지 할 수 있는 조건을 포함할 수 있다.

[0059] 동기화 신호 생성부(507)는 가상 슈퍼 노드의 반복 회수가 완료되면 그에 따라 실행 완료 메시지를 영상 처리 그래프 경로 실행부(도 2의 203 참조)에 보내어 다음 순서의 노드가 실행되도록 할 수 있다.

[0060] 도 6은 본 발명의 일 실시예에 따른 컴퓨터 시스템을 나타낸 도면이다.

[0061] 본 발명에 따른 온디바이스 시스템 및/또는 영상 처리 그래프 실행기는 컴퓨터 시스템(600)으로서 구현될 수 있다.

[0062] 도 6을 참조하면, 컴퓨터 시스템(600)은 버스(620)를 통하여 서로 통신하는 하나 이상의 프로세서(610), 메모리(630), 사용자 인터페이스 입력 장치(640), 사용자 인터페이스 출력 장치(650) 및 스토리지(660)를 포함할 수 있다. 또한, 컴퓨터 시스템(600)은 네트워크(680)에 연결되는 네트워크 인터페이스(670)를 더 포함할 수 있다. 프로세서(610)는 중앙 처리 장치 또는 메모리(630)나 스토리지(660)에 저장된 프로세싱 인스트럭션들을 실행하는 반도체 장치일 수 있다. 메모리(630) 및 스토리지(660)는 다양한 형태의 휘발성 또는 비휘발성 저장 매체일 수 있다. 예컨대, 메모리는 ROM(631)이나 RAM(632)을 포함할 수 있다.

[0063] 본 발명에서 설명하는 특정 실행들은 일 실시예들로서, 어떠한 방법으로도 본 발명의 범위를 한정하는 것은 아니다. 명세서의 간결함을 위하여, 종래 전자적인 구성들, 제어 시스템들, 소프트웨어, 상기 시스템들의 다른 기능적인 측면들의 기재는 생략될 수 있다. 또한, 도면에 도시된 구성 요소들 간의 선들의 연결 또는 연결 부재들은 기능적인 연결 및/또는 물리적 또는 회로적 연결들을 예시적으로 나타낸 것으로서, 실제 장치에서는 대체 가능하거나 추가의 다양한 기능적인 연결, 물리적인 연결, 또는 회로 연결들로서 나타내어질 수 있다. 또한, “필수적인”, “중요하게” 등과 같이 구체적인 언급이 없다면 본 발명의 적용을 위하여 반드시 필요한 구성 요소가 아닐 수 있다.

[0064] 따라서, 본 발명의 사상은 상기 설명된 실시예에 국한되어 정해져서는 아니 되며, 후술하는 특허청구범위뿐만 아니라 이 특허청구범위와 균등한 또는 이로부터 등가적으로 변경된 모든 범위는 본 발명의 사상의 범주에 속한다고 할 것이다.

부호의 설명

[0065] 100: 온디바이스 시스템 102: 영상 처리 그래프 실행기

103: 그래프 기반 실행 모델 104: 온디바이스용 이기종 라이브러리

105: 이기종 컴퓨팅 자원 200: 영상 처리 그래프 실행기

201, 300, 305 및 307: 영상 처리 그래프 모델

202: 영상 처리 그래프 경로 생성부

203: 영상 처리 그래프 경로 실행부

204: 노드 205, 308 및 500: 가상 슈퍼 노드

206: 노드 실행 추적 관리부 207: 실행 경로

304 및 506: 하위 그래프 306: 상호 순환 구조

501: 입출력 인터페이스 502: 노드 초기화 처리 인터페이스

503: 노드 실행 처리 인터페이스 504: 노드 종료 처리 인터페이스

505: 반복 실행 종료 조건 관리부 507: 동기화 신호 생성부

508: 순환 그래프 실행 엔진 600: 컴퓨터 시스템

610: 프로세서 620: 버스

630: 메모리 631: ROM

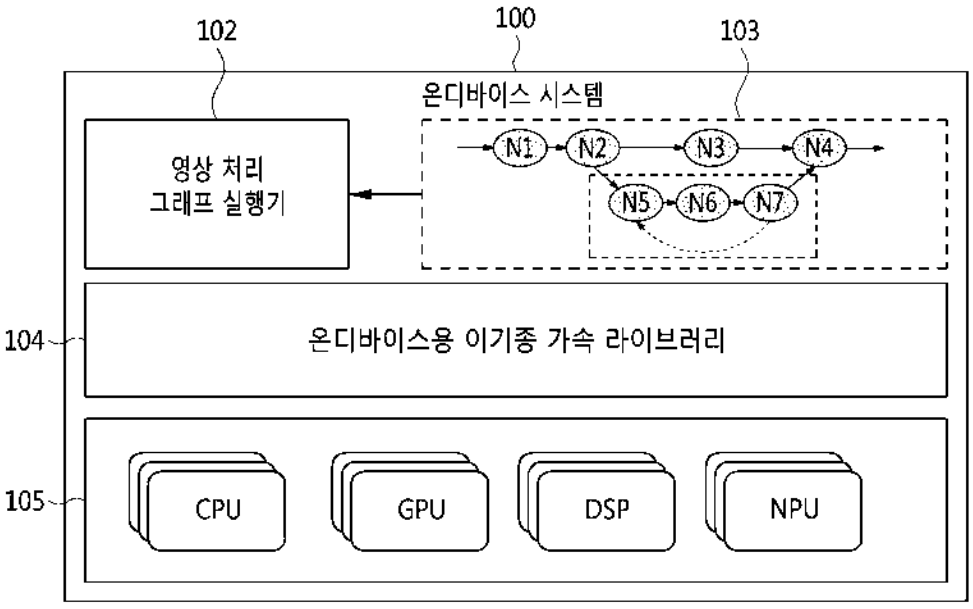
632: RAM 640: 사용자 인터페이스 입력 장치

650: 사용자 인터페이스 출력 장치 660: 스토리지

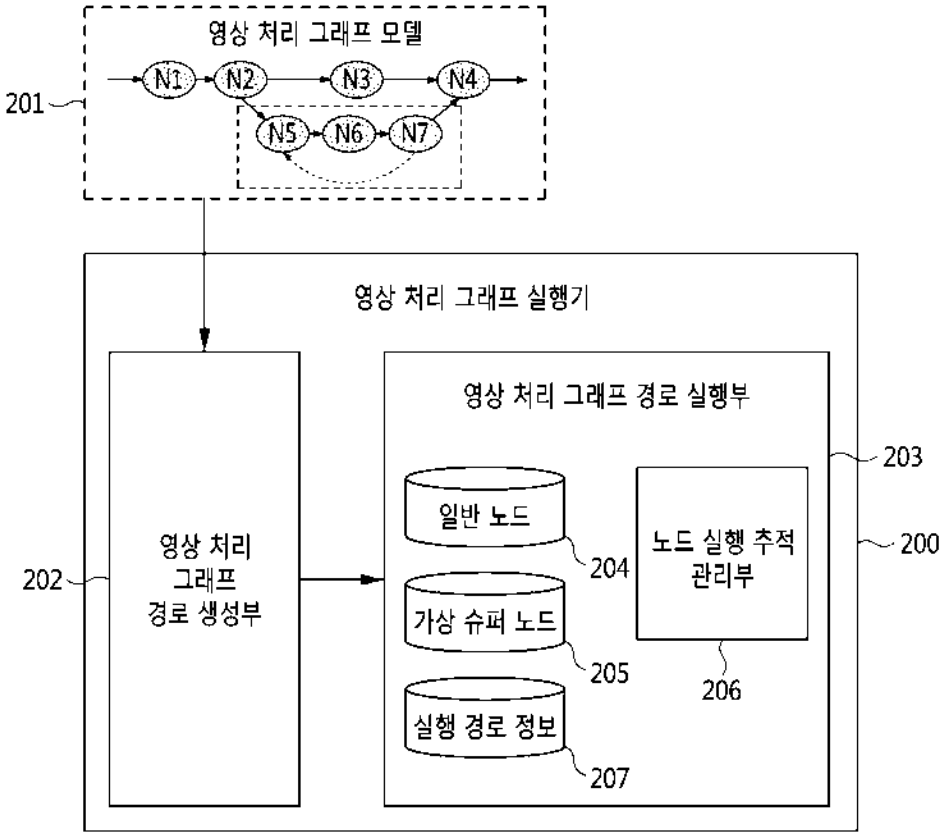
670: 네트워크 인터페이스 680: 네트워크

도면

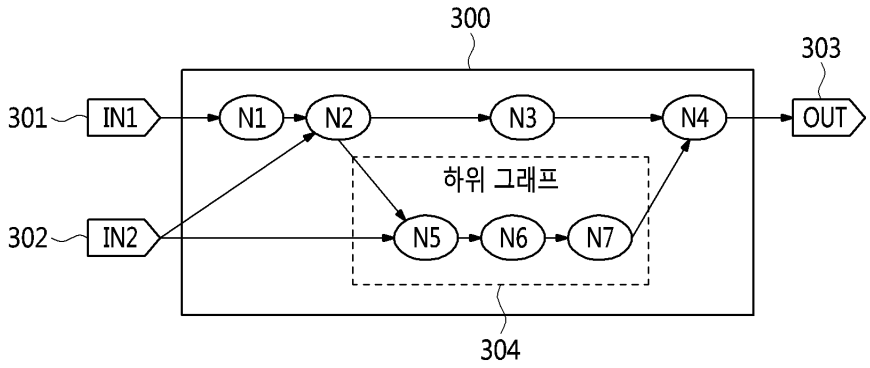
도면1



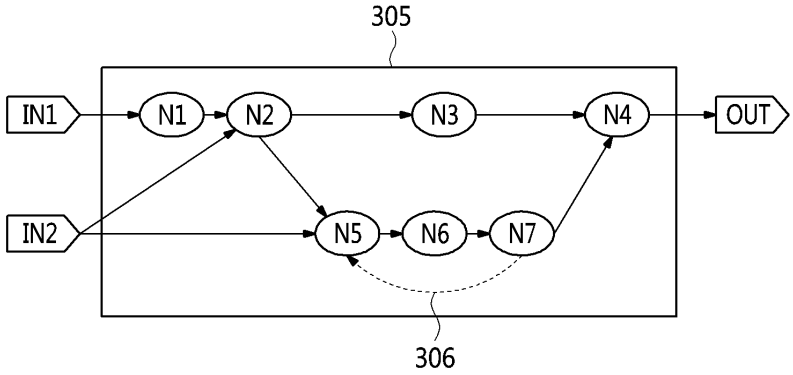
도면2


영상 처리 그래프 모델
N1
N2
N3
N4
N5
N6
N7
201
영상 처리 그래프 실행기
영상 처리
그래프
경로 생성부
202
영상 처리 그래프 경로 실행부
일반 노드
204
가상 슈퍼 노드
205
실행 경로 정보
207
노드 실행 추적
관리부
206
203
200

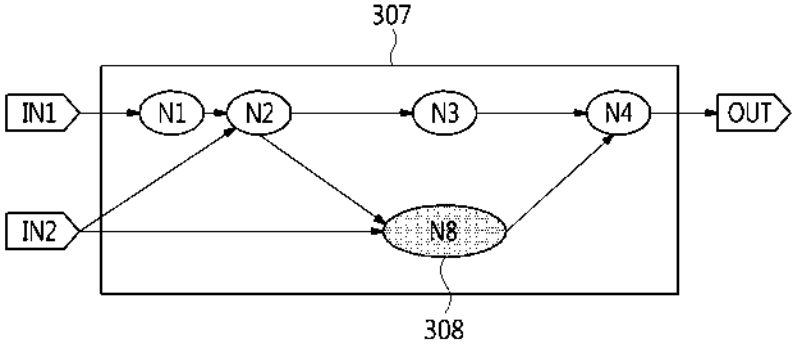
도면3a


300
303
301
IN1
N1
N2
N3
N4
OUT
하위 그래프
302
IN2
N5
N6
N7
304

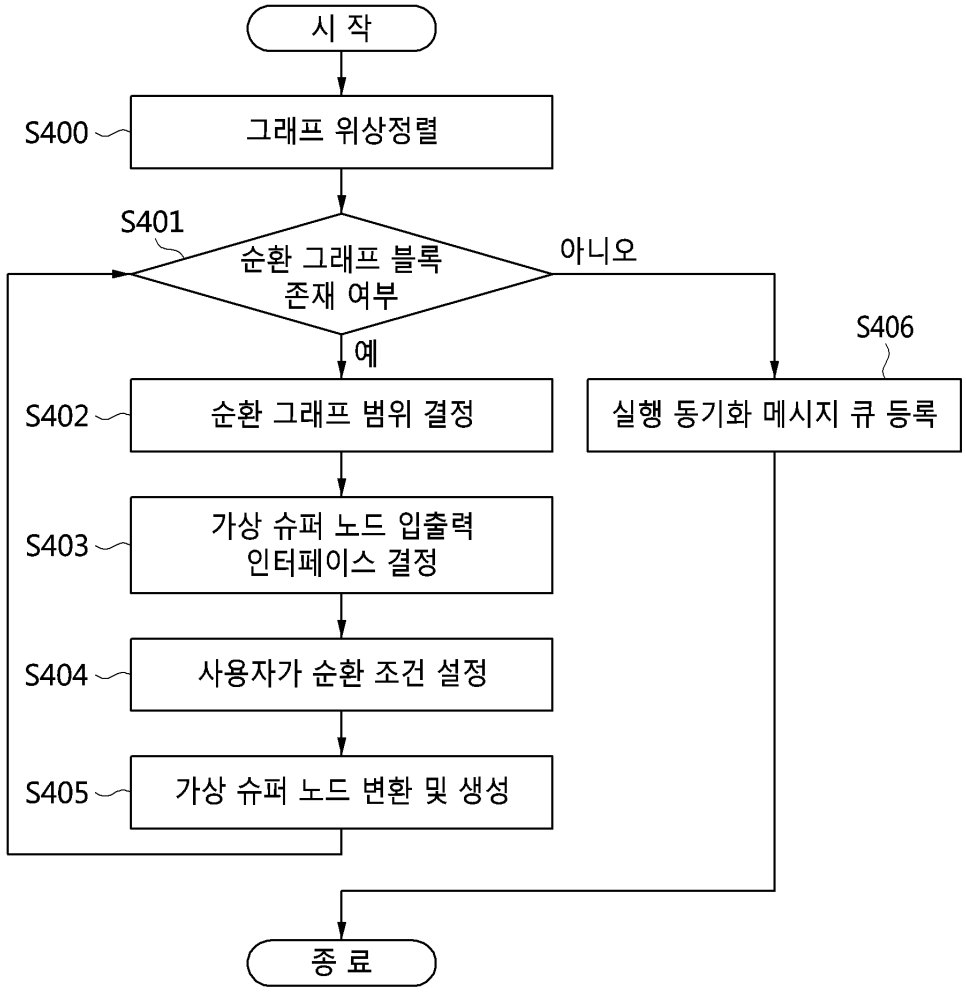
도면3b


305
IN1
N1
N2
N3
N4
OUT
IN2
N5
N6
N7
306

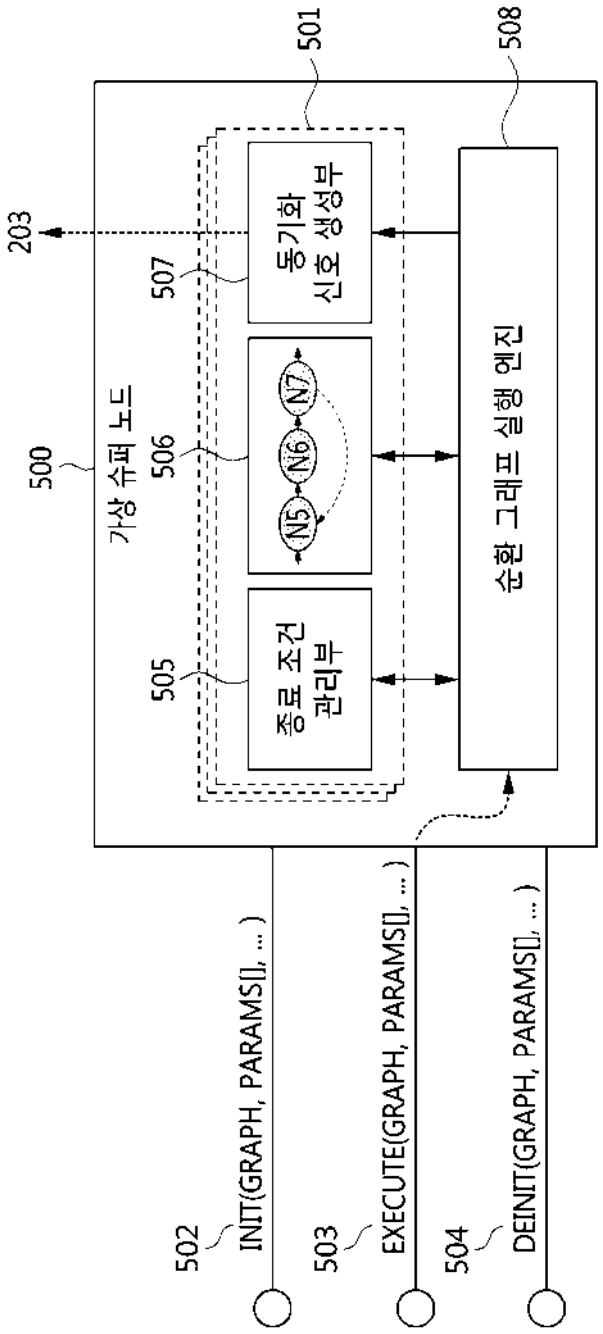
도면3c


307
IN1
N1
N2
N3
N4
OUT
IN2
N8
308

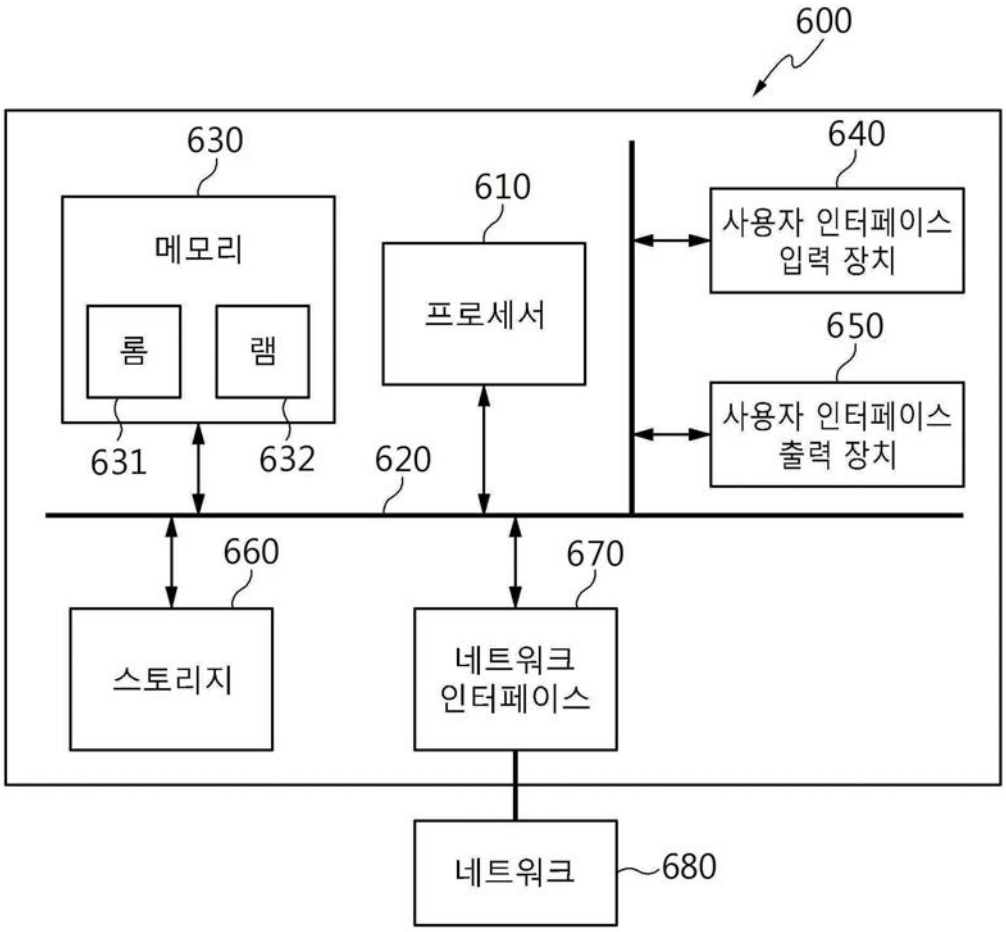
도면4


시 작
S400
그래프 위상정렬
S401
순환 그래프 블록
존재 여부
아니오
예
S406
실행 동기화 메시지 큐 등록
S402
순환 그래프 범위 결정
S403
가상 슈퍼 노드 입출력
인터페이스 결정
S404
사용자가 순환 조건 설정
S405
가상 슈퍼 노드 변환 및 생성
종 료

도면5


501
508
203
507
동기화 신호 생성부
500
가상 슈퍼 노드
506
N7
N6
N5
505
종료 조건 관리부
순환 그래프 실행 엔진
502
INIT(GRAPH, PARAMS[], ...)
503
EXECUTE(GRAPH, PARAMS[], ...)
504
DEINIT(GRAPH, PARAMS[], ...)

도면6


600
630
메모리
롬
램
631
632
610
프로세서
620
640
사용자 인터페이스
입력 장치
650
사용자 인터페이스
출력 장치
660
스토리지
670
네트워크
인터페이스
네트워크
680

딥러닝기반 의상정보(종류,속성) 인식기술

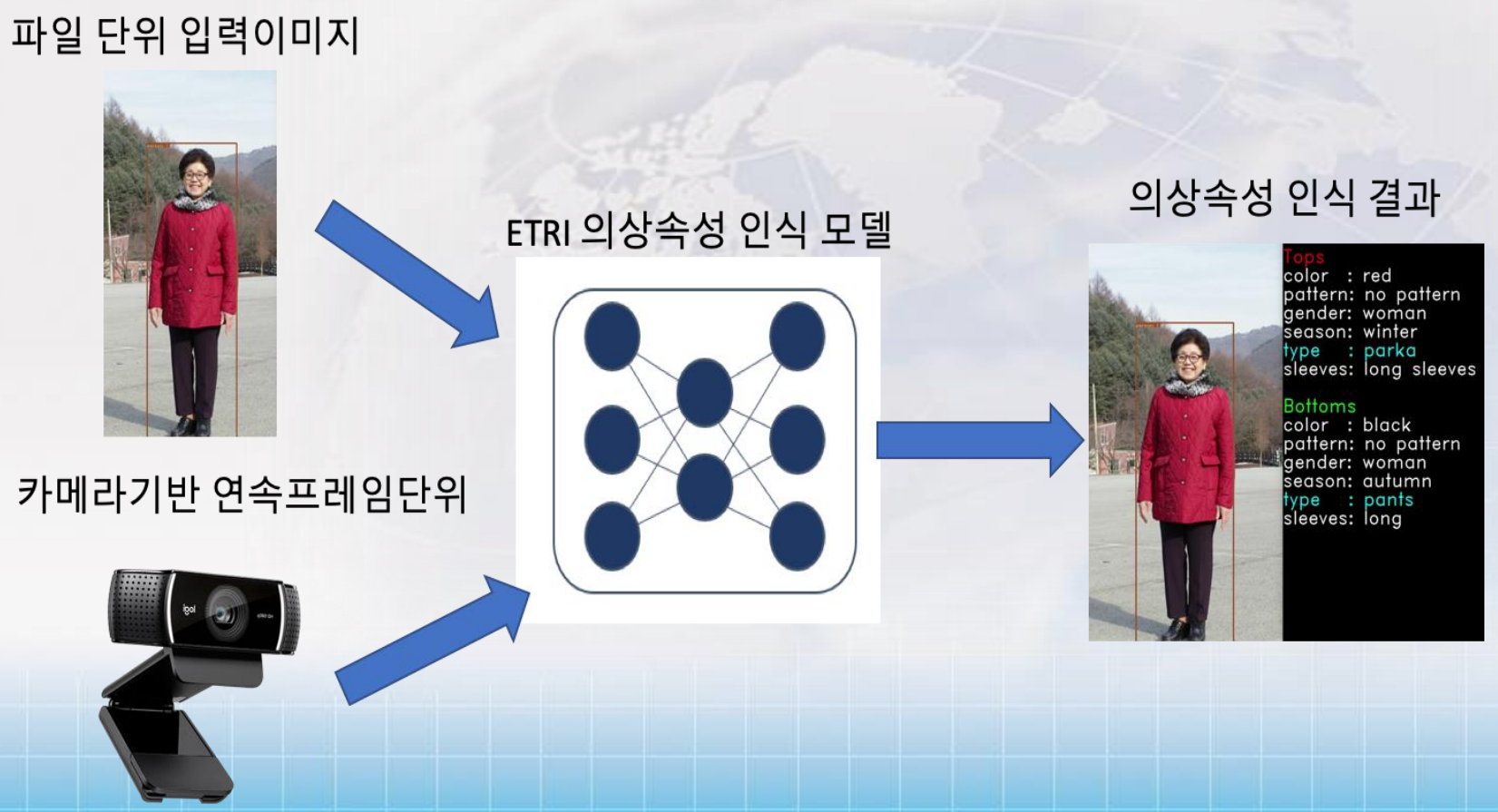
목 차

- 기술 개요
- 기술의 특성
- 기존 기술과의 차별성
- 기술 이전 범위
- 기술 응용 분야
- 시장성 예측
- 기술료 수준

기술 개요 (1): 소개

■ 딥러닝기반 의상정보(종류, 속성) 인식 기술

– 다양한 형태의(스마트폰, CCTV, 웹캠등) 카메라로부터 동영상 or 이미지들로부터 사람이 착용하고 있는 의상(상의, 하의)의 정보(종류, 12종 다중 속성)을 인식하여 추출하는 기술



기술 개요 (1): 소개

■ 딥러닝기반 의상정보(종류, 속성)인식 기술

의상의 표현 : 다중 속성(12종)정보 추출

색상, 무늬패턴, 성별, 계절, 옷의 종류, 소매길이, 바지길이

상의:
blue floral woman summer shirt long sleeves

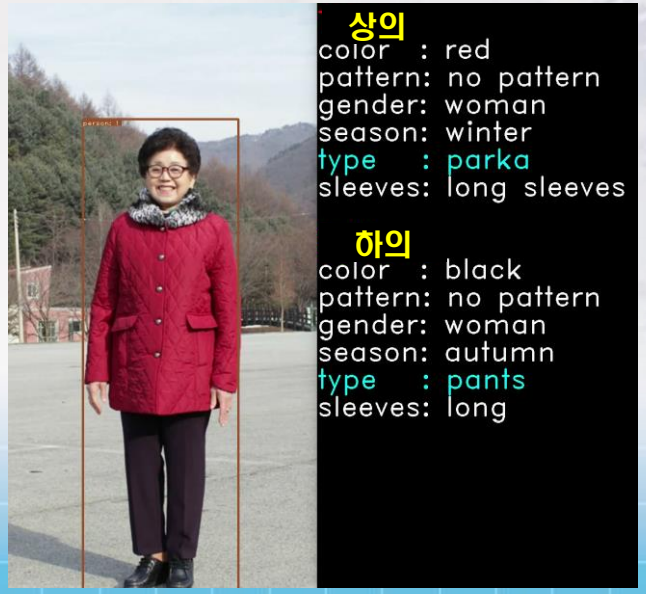
하의:
blue no pattern woman spring long pants

기술 개요 (1): 소개

■ 딥러닝기반 의상정보(종류, 속성)인식 기술

인식가능한 의상 종류 및 다양한 속성 정보 제공 : 12가지 속성그룹과 66가지 속성들

1	상의 종류(7)	shirt, jumper, jacket, vest, parka, coat, dress
2	상의 색상(14)	white, black, gray, pink, red, green, blue, brown, navy, beige, yellow, purple, orange, mixed
3	상의 계절(4)	spring, summer, autunm, winter
4	상의 패턴(6)	plain, checker, dotted, floral, striped, mixed
5	상의 소매(3)	short sleeves, long sleeves, no sleeves
6	상의 성별(2)	man, woman
7	하의 종류(2)	pants, skirt
8	하의 색상(14)	white, black, gray, pink, red, green, blue, brown, navy, beige, yellow, purple, orange, mixed
9	하의 계절(4)	spring, summer, autunm, winter
10	하의 패턴(6)	plain, checker, dotted, floral, striped, mixed
11	하의 길이(2)	short pants, long pants
12	하의 성별(2)	man, woman



기술의 특성

■ 딥러닝기반 의상정보(종류, 속성)인식 기술

- 다양한 카메라 환경(스마트폰, CCTV, webcam 등) or image 환경을 지원
- 사람 ROI기반 의상 정보 추출 기능: 상의, 하의 정보 구분 가능
- 이미지내의 복수명이 사람이 존재할 경우 복수명에 대한 의상(종류, 다중속성) 정보 추출
- 다양한 의상의 종류(9종)와 다양한 속성 12종(66가지)의 풍부한 의상 표현 정보 제공
- GPU 사용시 실시간 사용 가능 (15- 20fps처리)
- python pytorch기반 개발 환경

기술 이전 범위

■ 기술이전 내용

A. 기술명: 딥러닝기반 의상정보(종류, 속성)인식 기술

- 사람이 착용하고 있는 상의, 하의 의상 종류 및 다중 속성 12종 추출 기술
- 이미지내에 복수명의 사람이 있을 경우 복수명에 대한 의상정보 추출 기술

■ 기술이전 범위

A. 의상정보(종류, 속성)인식 엔진 소스 제공 (python, pytorch기반)

B. 의상정보(종류, 속성)인식 모델 제공 (python, pytorch기반)

C. 기술문서:

- 기술 설명서
- 성능평가 시험절차서 및 결과서

기술 응용 분야

■ 응용 분야

- 대형 IT 검색업체, 소셜네트워서비스업체, 패션관련 쇼핑몰, 의상정보 색인, 검색, 추천관련 서비스

예상 제품/서비스	예상 수요자(층)
의상정보(종류, 속성) 자동 태깅 서비스	IT검색업체, SNS, 쇼핑몰, 패션관련검색
다중속성지원 의상 정보 검색 서비스	쇼핑몰, 패션관련 업체, 일반 소비자
다중속성지원 의상 추천 서비스	쇼핑몰, 패션관련 업체, 일반 소비자

기술료 수준

구 분		공동연구 참여기업		일반 기업		
		중소기업	대기업	중소기업	중견기업	대기업
딥러닝기반 의상정보 (종류,속성) 인식기술	정액 기본료(원)	-	-	50,000,000	150,000,000	200,000,000

감사합니다

인공지능기반 2D->3D 얼굴 자동생성기술

목 차

- 기술 개요
- 기술의 특성
- 기존 기술과의 차별성
- 기술 이전 범위
- 기술 응용 분야
- 시장성 예측
- 기술료 수준

기술 개요 (1): 소개

2D->3D 얼굴 자동생성 기술

- 다수의 2D 얼굴 영상과 3D 얼굴 영상을 학습데이터로 입력 받아, 딥러닝 기법으로 학습하여, 2D 얼굴 영상이 입력으로 들어왔을 때 대응되는 3D 얼굴을 자동으로 생성하는 기술

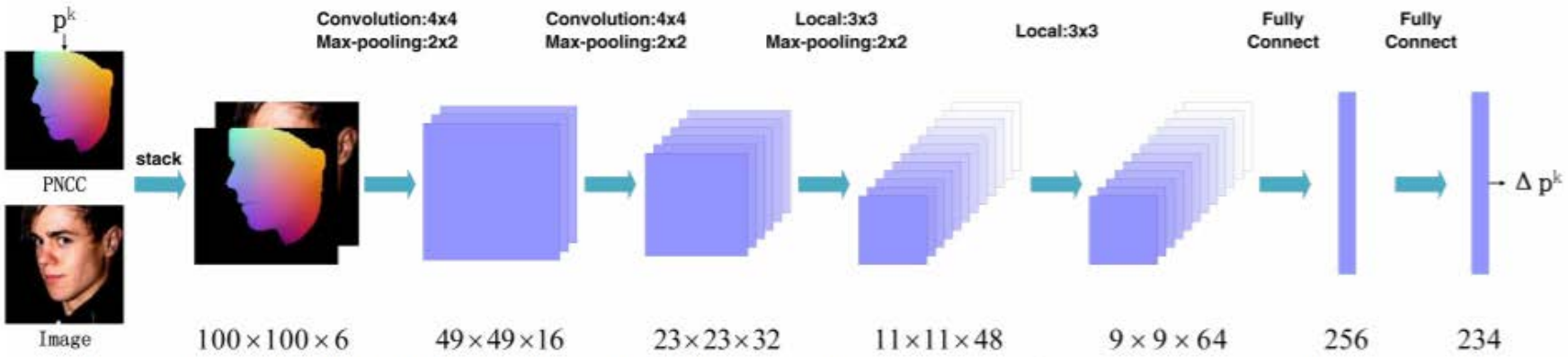
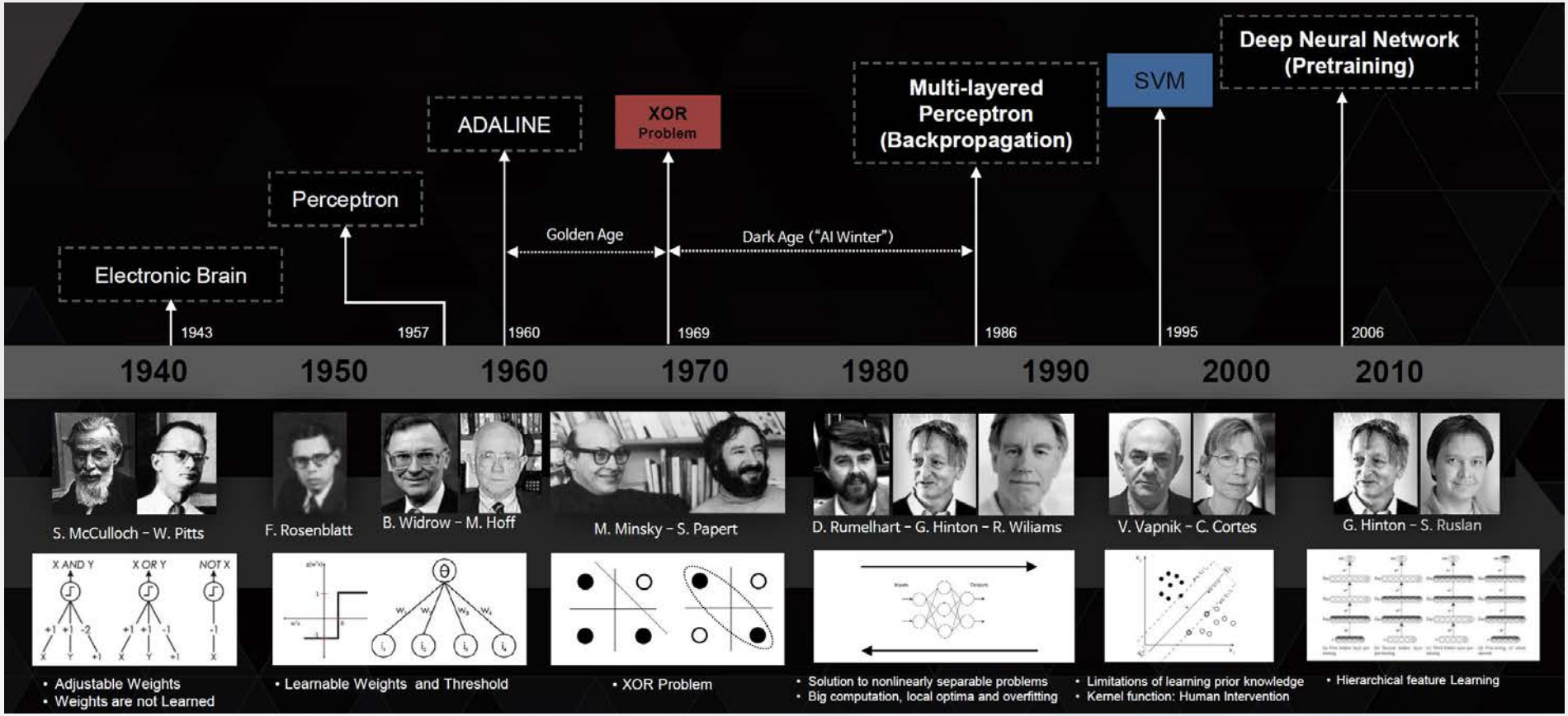


Figure 2. An overview of 3DDFA. At kth iteration, Net^k takes a medium parameter $\mathbf{p}^k$ as input, constructs the projected normalized coordinate code (PNCC), stacks it with the input image and sends it into CNN to predict the parameter update $\Delta\mathbf{p}^k$.

기술 개요 (1): 소개

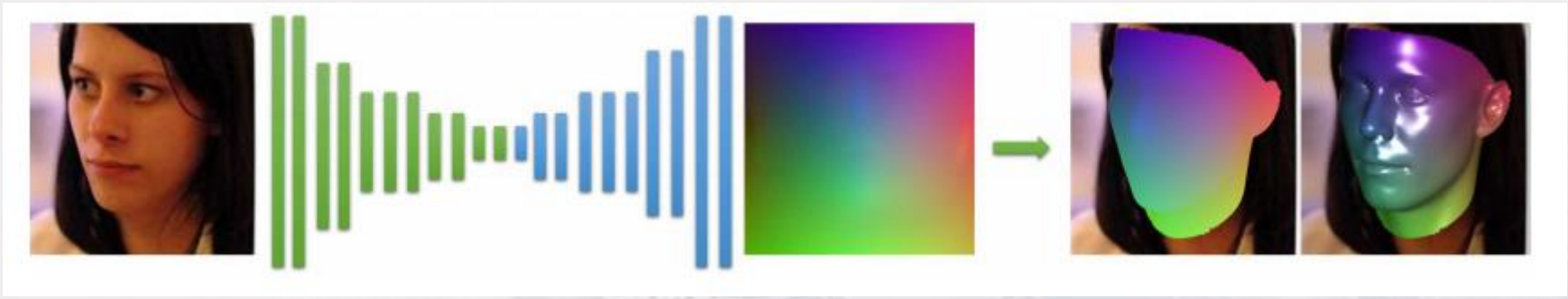
BRIEF HISTORY OF NEURAL NETWORK

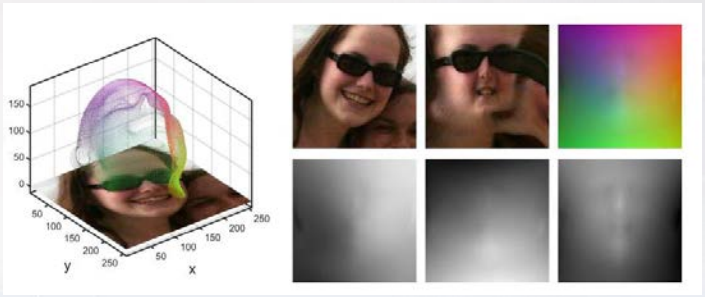


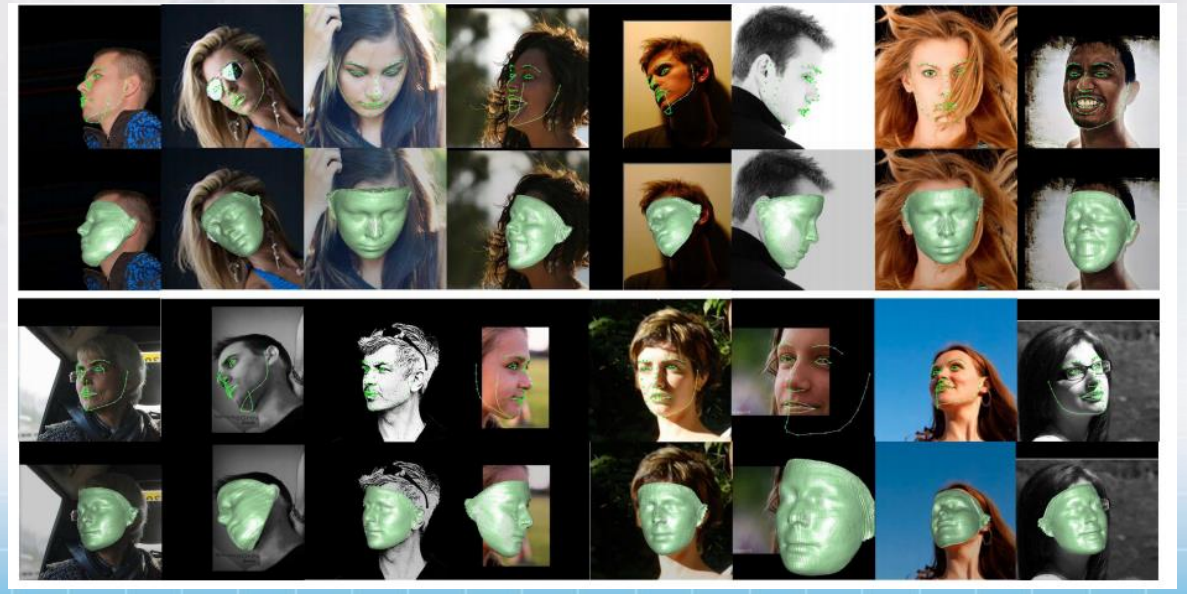
*출처: 2015 GTX KOREA VUNO(이예하)

기술 개요 (1): 소개

■ 딥러닝을 이용한 3D 얼굴 생성







기술의 특성

◆ **제안된 기술 장점 및 효과**

- 일반 CCTV 카메라나 저가의 USB 카메라 환경에서도 모두 적용 가능
- 정면, 왼측면, 오른 측면 세 번의 얼굴영상 획득으로 완벽한 3D 얼굴 모델 생성
- 다양한 조명, 거리, 헤어스타일을 가진 얼굴을 대상으로 검증 테스트시 겹침, 회전, 앞뒤 변화 변화에 모두 3D 얼굴 모델 생성 가능
- 640x480일반 사양의 컴퓨팅 환경에서 1~3초 이내의 실시간 3D 얼굴 모델링
- 얼굴 검출에 기반하여 3D 모델 생성시 3D 랜드마크 정보도 자동으로 추출
- 동시에 다수의 얼굴이 입력돼도 3D 얼굴 생성 가능
- Pytorch 3D를 기반으로 3D 얼굴 랜더링 가능
- 해외 및 국내 얼굴 데이터베이스를 이용하여 3D얼굴 모델링 성능 테스트

기술의 특성

- 시연 동영상


ETRI HRI
www.BANDICAM.com
Start
Image Capture
Create
Reset
Yaw Pitch Roll
0
10
0

기술의 특성

개발환경 및 주의사항

- **개발환경**
  - CPU : Intel i7-6700 @ 4.0GHz, 64G RAM
  - GPU : NVIDIA Geforce GTX 970
  - OS : Microsoft windows 10 Pro x64
  - Language : Pytorch

- **주의사항**
  - NVIDIA의 GPU기반 CUDA 연산을 수행하면 특징 추출의 속도가 빨라지기 때문에 GPU 사용 권장
  - NVIDIA Kepler / Maxwell architecture 에서 동작(GeForce 600 series 이상)
  - 대부분 GPU에서 처리되므로, CPU 성능이 좋을 필요 없음

기존 기술과의 차별성

기존(선행)기술과 비교하여 유리한 점

- 정면, 왼측면, 우측면 얼굴 영상 입력으로 모든 방향의 얼굴 생성 가능
- 비교적 저 사양의 컴퓨팅 환경에서 실시간 인식이 가능함
- 3D 모델 생성에 있어 딥러닝 기술을 이용함으로써 정확도를 84% 이상까지 확보함
- 얼굴인식 시도시, 다양한 포즈에서 입력되는 얼굴 영상을 3D 모델로 동일하게 생성함으로써, 인식 정확도 향상 가능

기존(선행)기술과 비교하여 불리한 점

- 기존 기술 대비 성능이나 속도에서 불리한 점은 없음.
- 실행 PC에 GPU기능이 포함된 그래픽카드의 사용을 추천함

기술 이전 범위

기술이전 범위

A. 기술명: 인공지능 기술을 이용한 2D->3D 얼굴 자동생성기술

- 내국인 혹은 외국인 얼굴 2D->3D 얼굴 자동생성 코드
- 내국인 혹은 외국인 얼굴 3D 랜드마크 자동 생성 코드
- 내국인 혹은 외국인 얼굴 시선 추적 코드
- Pytorch 코드로 기본 제공
- OS: Window and Linux 지원

특허 및 기술문서

A. 특허명:

- 해당사항 없음

B. 기술문서:

- 1420-2020-01084 : 소셜로봇 기술개발 요구사항 정의서
- 1420-2020-01085 : 소셜로봇 기술개발 시험결과 절차서

기술 응용 분야

■ 응용 분야

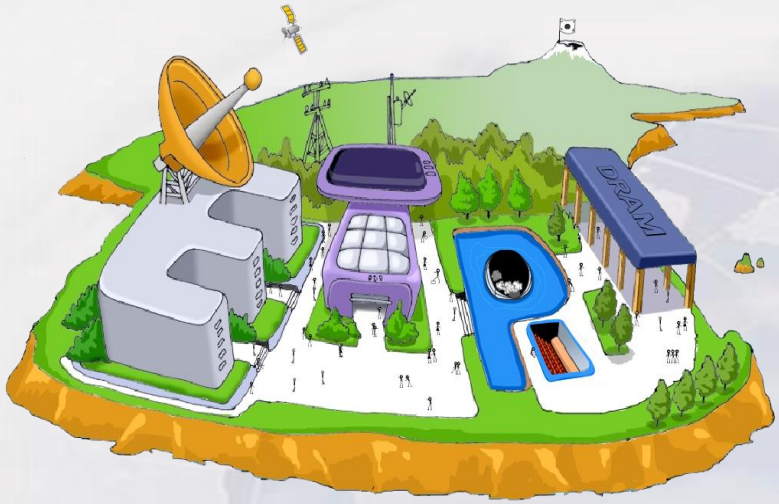
- 3D 얼굴 생성 기술은 활용도가 한정되나 높은 수요를 가진 기술로서 지능형 로봇, 지능형 CCTV 업체 들도 수요자가 될 수 있음

예상 제품/서비스	예상 수요자(층)
지능형로봇/사용자 맞춤형 서비스	헬스케어, 돌보미 로봇 업체(사용자 인식)
범죄자 검출 및 추적	지능형 CCTV 업체(범죄자 인식 및 추적)
게임 케릭터 생성	게임 업체

기술료 수준

구 분		공동연구 참여기업		일반 기업		
		중소기업	대기업	중소기업	중견기업	대기업
A.딥러닝기반 실시간 알약 인식기술	정액기본료(원)	-	-	33,000,000	99,000,000	132,000,000

감사합니다



인공지능(딥러닝)기반 실시간 헤어, 수염 정보 추출 및 인식 기술

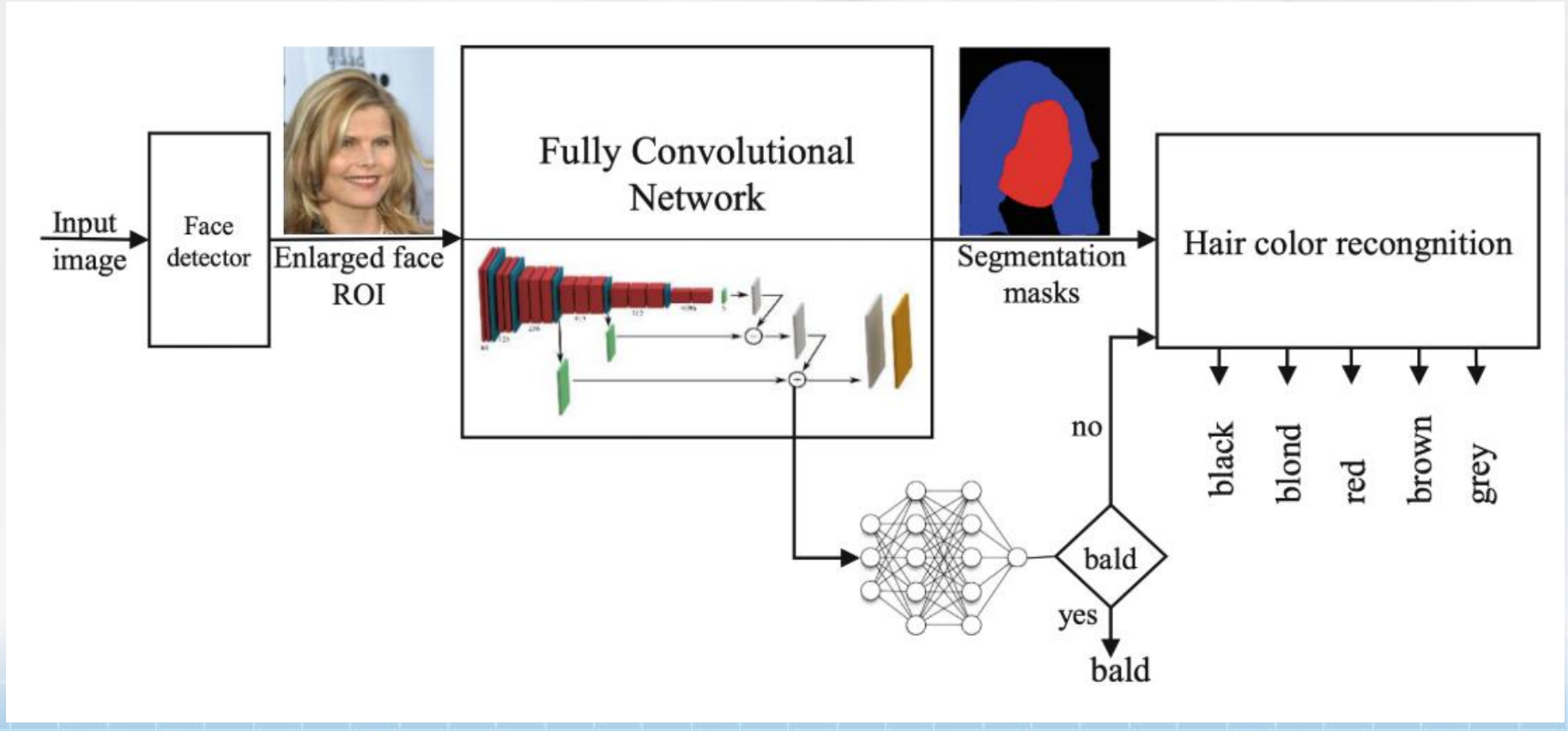
목 차

- 기술 개요
- 기술의 특성
- 기존 기술과의 차별성
- 기술 이전 범위
- 기술 응용 분야
- 시장성 예측
- 기술료 수준

기술 개요 (1): 소개

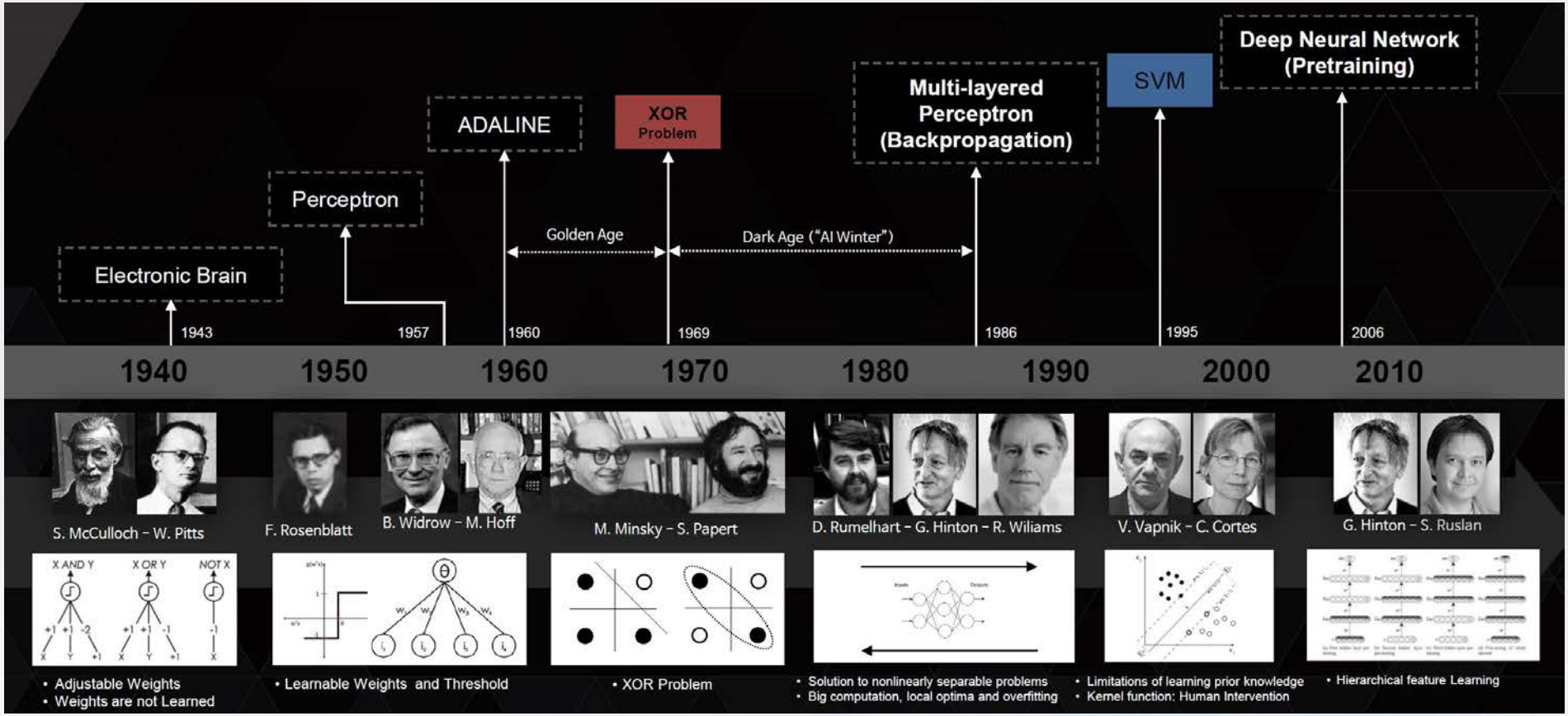
헤어, 수염 인식 기술

- 영상 내에서 헤어 및 수염을 검출하고, 검출된 영역을 분석하여 스타일, 길이, 색상을 인식하는 기술



기술 개요 (1): 소개

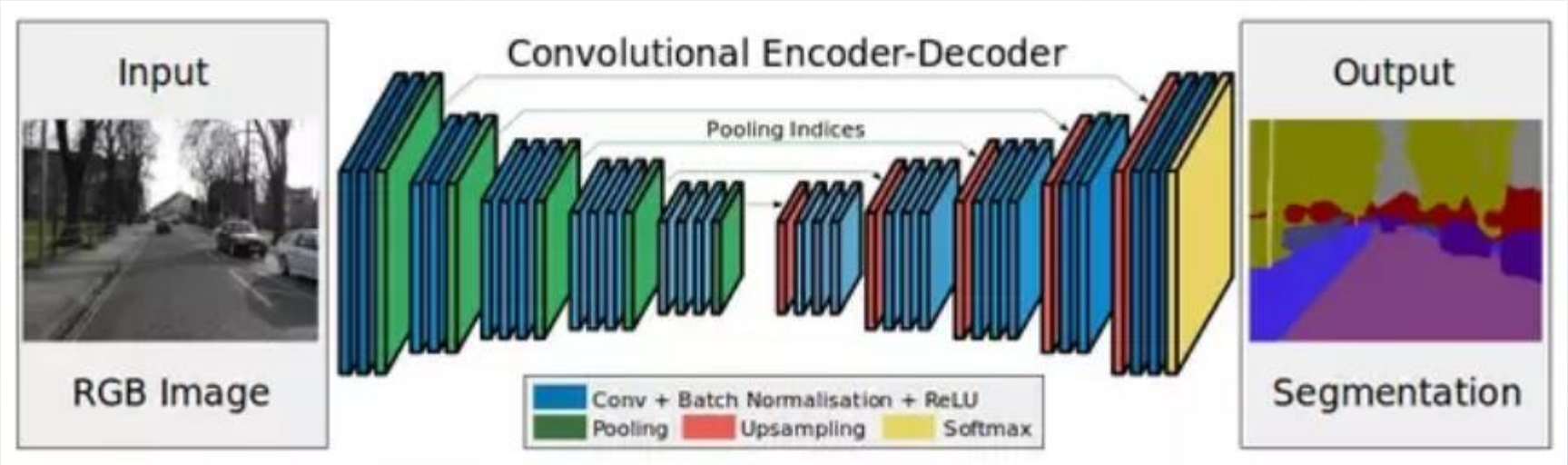
BRIEF HISTORY OF NEURAL NETWORK



*출처: 2015 GTX KOREA VUNO(이예하)

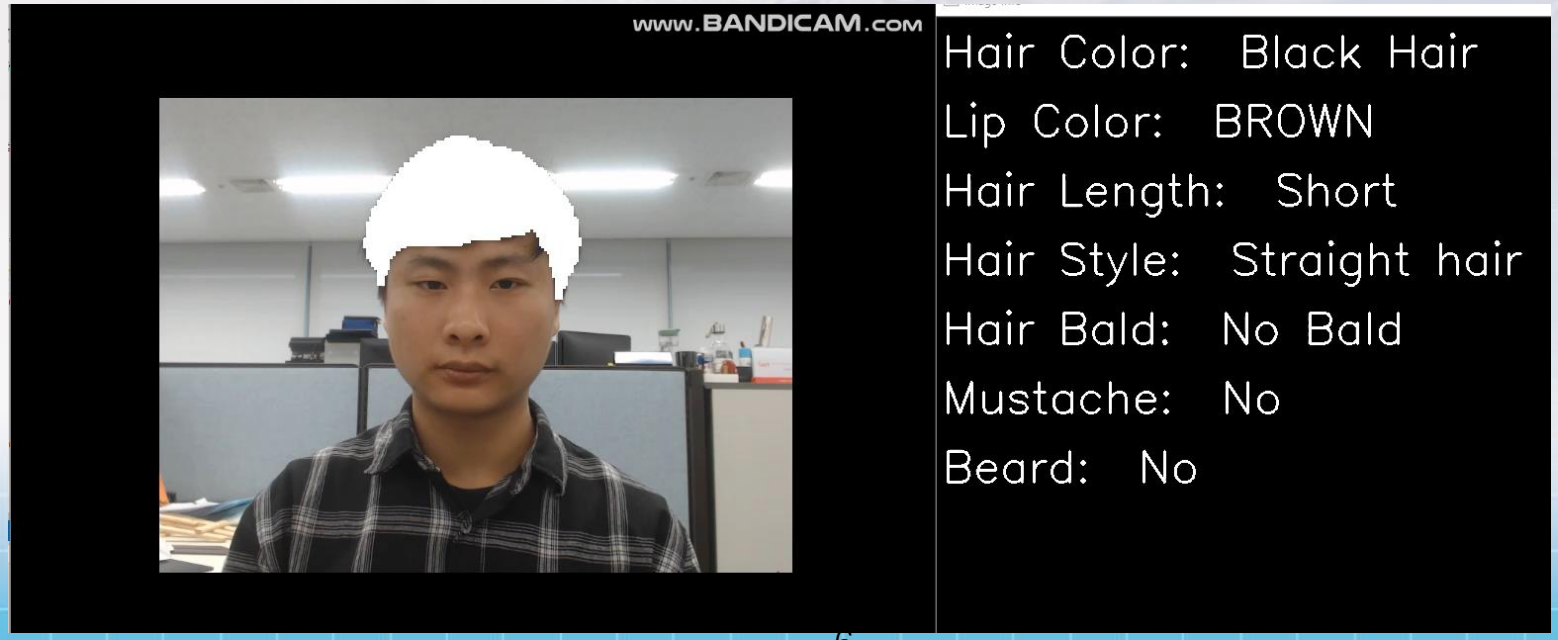
기술 개요 (1): 소개

■ 딥러닝을 이용한 Segmentation



기술의 특성

- Mobile Unet V2 기반 최적화된 분할 방법 사용해 실시간 처리 가능
- 6,800의 학습데이터를 14배로 Augmentation
- 기존의 서양인 학습데이터에 국내 얼굴 학습데이터 추가
- 헤어와 수염을 별개의 딥러닝 모델을 이용해 인식
- 딥러닝 방법과 Rule-based 방법을 융합하여 인식율 향상
- 헤어 정보는 스타일, 길이 컬러 정보 제공
- 수염 정보는 구랫나루, 턱수염, 콧수염 정보 제공



기술의 특성

- **제안된 기술 장점 및 효과**

. 일반 CCTV 카메라나 저가의 USB 카메라 환경에서도 모두 적용 가능

. 다양한 조명, 거리, 헤어스타일을 가진 얼굴을 대상으로 인식 검증 테스트시 겹침, 회전, 앞뒤 변화 변화에 모두 90% 이상 검출 및 인식 이상

. 640x480일반 사양의 컴퓨팅 환경에서 초당 1~5 프레임 이상의 실시간 얼굴 검출 및 인식

. 명암 정보를 사용한 얼굴 검출 및 인식 기술

. 헤어, 수염 검출 및 인식을 위한 학습 DB 확보

. 얼굴 검출에 기반하여 헤어, 수염 검출기를 따로 분리하여 설계

. 동시에 다수의 얼굴이 입력돼도 검출 및 인식 가능

. 얼굴이 가깝거나 멀어도 강인하게 검출 및 인식 가능

. 해외 및 국내 얼굴 데이터베이스를 이용하여 검출 및 인식 성능 테스트

기술의 특성

개발환경 및 주의사항

- **개발환경**
  - CPU : Intel i7-6700 @ 4.0GHz, 64G RAM
  - GPU : NVIDIA Geforce GTX 970
  - OS : Microsoft windows 10 Pro x64
  - Compiler : Microsoft Visual Studio 2013
  - Language : C++

- **주의사항**
  - NVIDIA의 GPU기반 CUDA 연산을 수행하면 특징 추출의 속도가 빨라지기 때문에 GPU 사용 권장
  - NVIDIA Kepler / Maxwell architecture 에서 동작(GeForce 600 series 이상)
  - 대부분 GPU에서 처리되므로, CPU 성능이 좋을 필요 없음

기존 기술과의 차별성

기존(선행)기술과 비교하여 유리한 점

- 다양한 환경의 변화에도 강인하게 인식할 수 있으며 멀리 떨어져 있어도 상대적(구글)으로 인식이 정확함
- 비교적 저 사양의 컴퓨팅 환경에서 실시간 인식이 가능함
- 헤어 및 수염의 검출과 인식기를 딥러닝 기술을 이용함으로써 정확도를 90%이상까지 확보함
- 새로운 헤어 및 수염의 검출 및 인식이 필요할 때, 쉽게 재 학습하여 인식성능 확장 가능함

기존(선행)기술과 비교하여 불리한 점

- 기존 기술 대비 성능이나 속도에서 불리한 점은 없음.
- 실행 PC에 GPU기능이 포함된 그래픽카드의 사용을 추천함

기술 이전 범위

기술이전 범위

A. 기술명: 인공지능(딥러닝) 기술을 이용한 실시간 헤어, 수염정보 검출 및 인식기술

- 내국인 혹은 외국인 얼굴 실시간 자동 검출기 소스코드
- 내국인 혹은 외국인 얼굴 실시간 헤어정보(길이, 색상, 스타일) 자동 검출 및 인식기 소스코드
- 내국인 혹은 외국인 얼굴 실시간 수염정보(턱수염, 콧수염, 구레나룻) 자동 검출 및 인식기 소스코드
- C++(Window and Linux 지원), Pytorch, ONNX

특허 및 기술문서

A. 특허명: 3D 카메라를 이용한 머리카락 영역 검출 방법 및 이를위한 장치

- 출원번호: 2020-0024067

A. 기술문서:

관리번호	기술자료 명칭	비고
1420-2017-03516	요구사항정의서	
1420-2017-05185	휴먼케어 서비스 선택 엔진 요구사항 분석	

기술 응용 분야

응용 분야

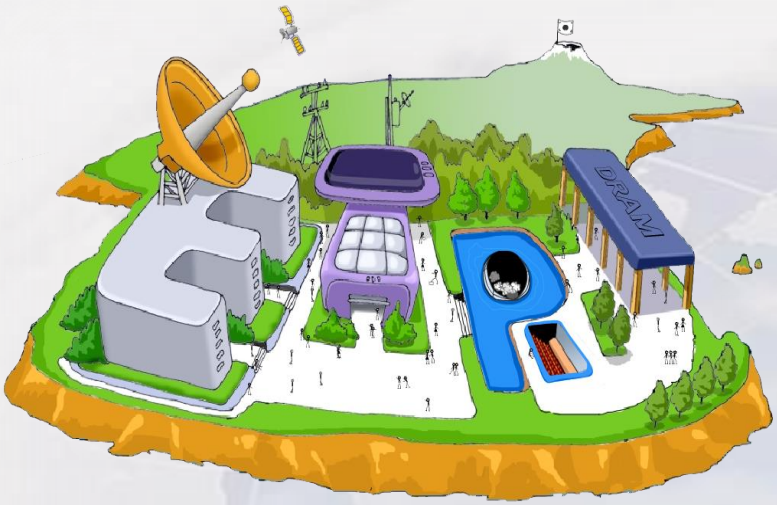
- 헤어, 수염 검출 및 인식 기술은 활용도가 한정되나 높은 수요를 가진 기술로서 지능형 로봇, 헤어샵 들도 수요자가 될 수 있음

예상 제품/서비스	예상 수요자(층)
지능형로봇/사용자 맞춤형 서비스	헬스케어, 돌보미 로봇
헤어 색상 미리보기/고객 서비스	헤어샵

기술료 수준

구 분		공동연구 참여기업		일반 기업		
		중소기업	대기업	중소기업	중견기업	대기업
A.딥러닝기반 실시간 알약 인식기술	정액기본료(원)	-	-	33,000,000	99,000,000	132,000,000

감사합니다



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0109167
(43) 공개일자 2021년09월06일

(51) 국제특허분류(Int. Cl.)
G06T 7/11 (2017.01) *G06T 5/00* (2019.01)
G06T 7/593 (2017.01)
(52) CPC특허분류
G06T 7/11 (2017.01)
G06T 5/005 (2013.01)
(21) 출원번호 10-2020-0024067
(22) 출원일자 2020년02월27일
심사청구일자 없음

(71) 출원인
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
윤호섭
대전광역시 유성구 어은로 57, 119동 701호
박성우
경기도 고양시 일산서구 대화1로 70, 701동 1701호
(74) 대리인
한양특허법인

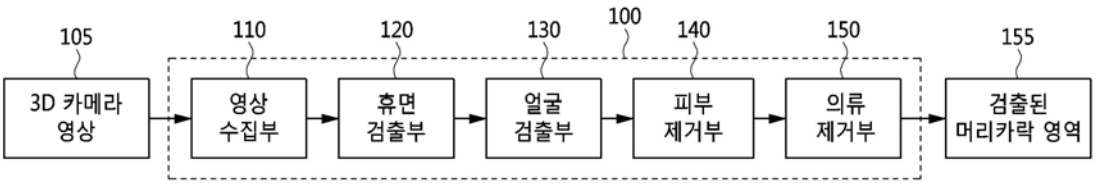
전체 청구항 수 : 총 1 항

(54) 발명의 명칭 **3D 카메라를 이용한 머리카락 영역 검출 방법 및 이를 위한 장치**

(57) 요 약

기재된 실시예는 사람의 영상으로부터 머리카락 영역을 쉽고 간단하게 검출하는 방법에 관한 것으로서, 깊이 영상을 기반으로 색상 영상으로부터 배경 영역을 제거하여 휴먼 영역을 검출하는 단계, 상기 휴먼 영역에서 얼굴 주변 영역을 검출하는 단계, 상기 휴먼 영역에서 상기 얼굴 주변 영역 내에 위치한 피부영역을 제거하여 피부 제외 영역을 검출하는 단계 및 적외선 영상을 기반으로 상기 피부 제외 영역에서 의류 영역을 제거하여 머리카락 영역을 검출하는 단계를 포함하는, 3D 카메라를 이용한 머리카락 영역 검출 방법이 제공될 수 있다.

대 표 도



(52) CPC특허분류
G06T 7/593 (2017.01)
G06T 2207/30088 (2013.01)
G06T 2207/30201 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	2017-0-00162
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원(IITP)
연구사업명	ICT융합산업원천기술개발사업
연구과제명	고령 사회에 대응하기 위한 실환경 휴먼케어 로봇 기술 개발
기 여 율	1/1
과제수행기관명	한국전자통신연구원
연구기간	2019.01.01 ~ 2019.12.31

명 세 서

청구범위

청구항 1

깊이 영상을 기반으로 색상 영상으로부터 배경 영역을 제거하여 휴먼 영역을 검출하는 단계;

상기 휴먼 영역에서 얼굴 주변 영역을 검출하는 단계;

상기 휴먼 영역에서 상기 얼굴 주변 영역 내에 위치한 피부영역을 제거하여 피부 제외 영역을 검출하는 단계; 및

적외선 영상을 기반으로 상기 피부 제외 영역에서 의류 영역을 제거하여 머리카락 영역을 검출하는 단계를 포함하는, 3D 카메라를 이용한 머리카락 영역 검출 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 3D 카메라를 이용하여 사람의 영상으로부터 쉽고 간단하게 머리카락 영역을 검출하는 기술에 관한 것이다.

배 경 기 술

[0002] 현재 가상 헤어스타일 디자인, 가상 인간 모델, 가상 이미지 디자인 등에 머리카락 영역 검출 기술을 매우 유용하다. 그러나 현대인의 다양한 헤어스타일, 머리카락의 색상 및 밝기로 인해 머리카락의 검출은 매우 도전적인 연구 주제가 되었다.

[0003] 종래의 카메라를 이용한 머리카락 검출 방법은 주로 인공지능의 머신 러닝 또는 딥러닝을 이용하여 미리 준비된 데이터 셋을 가지고 학습하여 검출하는 방식으로 진행된다.

[0004] 딥러닝을 이용한 머리카락 검출 방식은 머리카락이 분할된 데이터셋이 많이 필요하고, 딥러닝 모델을 감당할 수 있는 높은 사양의 컴퓨터가 필요하다. 그러나 현재 머리카락이 분할되어 있는 데이터셋 중 다운받아 쓸 수 있는 데이터는 약 6,000개에 불과하고, 실제 머리카락 분할을 통해 컨텐츠를 개발하기 위해서는 많은 인력을 고용해 다량의 고수준의 데이터셋을 만들어내야 한다. 따라서 딥러닝을 이용한 머리카락 검출 방식은 정확도가 보장되는 반면, 그 소요 비용이 상당하다.

[0005] 현재 머신러닝을 이용한 머리카락 검출 방식은 LTP, HOG, SIFT 등 기술자를 이용하여 이미지의 대표 기술자를 추출한 다음 머신러닝 기법(주로 Random Forest)를 사용해 어느정도 머리 영역의 윤곽을 잡은 다음에 그래프 컷 등을 이용해 분할을 하지만 정확도가 계획했던 것만큼 나오지 않는다. 한국 공개 특허 제 10-2100-0090764호에는 머리영역의 신뢰도 이미지를 획득하여, 상기 회복한 신뢰도 이미지를 처리하여 머리카락영역을 검출하는 단계를 포함하는 머리카락 영역 검출 방법이 개시되어 있다. 상기 방법은 피부 및 머리카락의 색깔, 주파수, 깊이 정보를 결합하여 머리카락 영역을 검출하는데 그래프 컷 방법를 이용하여 노이즈 배경에서 전체 머리카락 영역을 분할한다.

[0006] 위와 같은 종래 기술은 검출의 정확성은 어느 정도 보장할 수는 있지만, 잘못된 검출을 한다면 그 원인을 알기 어려워 그 오류를 고치기가 쉽지 않다. 그리고 데이터셋의 정확성에 대한 의존도가 높고, 딥러닝이나 머신러닝에 대한 고도의 지식도 필요하며, 무엇보다도 비용의 부담이 상당하다.

[0007] 따라서 위 종래 기술들이 가지는 문제점을 해결하고, 쉽고 간단하게 머리카락 영역을 검출하는 기술의 필요성이 대두된다.

발명의 내용

해결하려는 과제

[0008] 본 발명의 목적은, 3D 카메라를 이용하여 사람의 영상으로부터 머리카락 영역을 쉽고 간단하게 검출하는 방법

및 이를 위한 장치를 제공함에 있다.

과제의 해결 수단

[0009] 실시예에 따른 3D 카메라를 이용한 머리카락 영역 검출 방법은, 깊이 영상을 기반으로 색상 영상으로부터 배경 영역을 제거하여 휴먼 영역을 검출하는 단계; 상기 휴먼 영역에서 얼굴 주변 영역을 검출하는 단계; 상기 휴먼 영역에서 상기 얼굴 주변 영역 내에 위치한 피부영역을 제거하여 피부 제외 영역을 검출하는 단계; 및 적외선 영상을 기반으로 상기 피부 제외 영역에서 의류 영역을 제거하여 머리카락 영역을 검출하는 단계를 포함한다.

[0010] 상기 얼굴 주변 영역을 검출하는 단계는 상기 휴먼 영역에 대하여 랜드마크 알고리즘을 사용하여 복수개의 랜드마크 포인트를 검출하는 단계; 적어도 하나 이상의 상기 랜드마크 포인트에 기반하여 얼굴 주변 영역을 검출하는 단계를 포함할 수 있다.

[0011] 상기 피부 제외 영역을 검출하는 단계는 피부 색상의 범위에 기반하여, 색상의 값이 상기 피부 색상의 범위 내에 해당하는 영역을 피부 영역으로 생성하는 단계를 포함할 수 있다.

[0012] 상기 피부 제외 영역을 검출하는 단계는 적어도 하나 이상의 상기 랜드마크 포인트로부터 얼굴의 외곽선을 검출하는 단계를 더 포함하고, 상기 얼굴의 외곽선의 내부에 위치하는 상기 피부 영역을 제거하는 것일 수 있다.

[0013] 상기 의류 영역을 제거하는 단계는 상기 적외선 영상으로부터 정해진 밝기 이상의 값을 가지는 구조광(Structured Light)의 위치를 추출하는 단계; 상기 구조광의 크기를 초기값으로부터 증가시키며 구조광을 제거하는 단계를 포함할 수 있다.

[0014] 상기 구조광의 위치를 추출하는 단계는 상기 랜드마크 포인트에 기반하여 두 눈 사이의 거리를 추출하는 단계; 상기 두 눈 사이의 거리로부터 상기 구조광의 크기의 초기값을 정하는 단계를 포함할 수 있다.

[0015] 상기 피부 제외 영역을 검출하는 단계에서 색상 모델을 YUV 모델을 사용한 경우 상기 피부 영역을 $77 \leq U \leq 127$, $133 \leq V \leq 173$ 인 영역으로 정하는 것일 수 있다.

[0016] 3D 카메라를 이용한 머리카락 영역 검출 방법은 상기 색상 영상의 기준 좌표를 상기 깊이 영상의 기준 좌표에 맞추는 단계를 더 포함할 수 있다.

[0017] 상기 휴먼 영역을 검출하는 단계는 상기 깊이 영상의 깊이 값이 1보다 작거나 같은 경우는 휴먼 영역으로 저장하는 단계; 상기 깊이 영상의 깊이 값이 1보다 큰 경우는 배경 영역으로 저장하는 단계를 포함할 수 있다.

[0018] 실시예에 따른 3D 카메라를 이용한 머리카락 영역 검출 장치는 깊이 영상을 기반으로 색상 영상으로부터 배경 영역을 제거하여 휴먼 영역을 검출하는 휴먼 검출부; 상기 휴먼 영역에서 얼굴 주변 영역을 검출하는 얼굴 검출부; 상기 휴먼 영역에서 상기 얼굴 주변 영역 내에 위치한 피부영역을 제거하여 피부 제외 영역을 검출하는 피부 제거부; 및 적외선 영상을 기반으로 상기 피부 제외 영역에서 의류 영역을 제거하여 머리카락 영역을 검출하는 의류 제거부를 포함할 수 있다.

발명의 효과

[0019] 본 발명에 따르면, 3D 카메라를 이용하여 사람의 영상으로부터 머리카락 영역을 쉽고 간단하게 검출하는 방법 및 이를 위한 장치를 제공할 수 있다.

도면의 간단한 설명

[0020] 도 1은 실시예에 따른 3D 카메라를 이용한 머리카락 영역 검출 장치의 일 예를 나타낸 블록도이다.

도 2는 실시예에 따른 3D 카메라를 이용한 머리카락 영역 검출 방법의 일 예를 나타낸 동작 흐름도이다.

도 3은 도 2에 도시된 영상 입력 단계의 일 예를 나타낸 동작 흐름도이다.

도 4는 도 2에 도시된 휴먼 검출 단계의 일 예를 나타낸 동작 흐름도이다.

도 5는 도 2에 도시된 얼굴 검출 단계의 일 예를 나타낸 동작 흐름도이다.

도 6는 도 2에 도시된 피부 검출 및 제거 단계의 일 예를 나타낸 동작 흐름도이다.

도 7은 도 2에 도시된 의류 검출 및 제거 단계의 일 예를 나타낸 동작 흐름도이다.

도 8은 실시예에 따른 본 발명의 활용예를 나타낸 도면이다.

도 9는 실시예에 따른 컴퓨터 시스템 구성을 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0021] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 것이며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하며, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 명세서 전체에 걸쳐 동일 참조 부호는 동일 구성 요소를 지칭한다.

[0022] 비록 "제1" 또는 "제2" 등이 다양한 구성요소를 서술하기 위해서 사용되나, 이러한 구성요소는 상기와 같은 용어에 의해 제한되지 않는다. 상기와 같은 용어는 단지 하나의 구성요소를 다른 구성요소와 구별하기 위하여 사용될 수 있다. 따라서, 이하에서 언급되는 제1 구성요소는 본 발명의 기술적 사상 내에서 제2 구성요소일 수도 있다.

[0023] 본 명세서에서 사용된 용어는 실시예를 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 또는 "포함하는(comprising)"은 언급된 구성요소 또는 단계가 하나 이상의 다른 구성요소 또는 단계의 존재 또는 추가를 배제하지 않는다는 의미를 내포한다.

[0024] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어는 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 공통적으로 이해될 수 있는 의미로 해석될 수 있다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.

[0025] 이하에서는, 도 1 내지 도 9를 참조하여 실시예에 따른 3D 카메라를 이용한머리카락 영역 검출 방법 및 이를 위한 장치가 상세히 설명된다.

[0026] 도 1은 실시예에 따른 3D 카메라를 이용한 머리카락 영역 검출 장치의 일 예를 나타낸 블록도이다.

[0027] 도 1을 참조하면, 3D 카메라를 이용한 머리카락 영역 검출 장치(100)는 영상 수집부(110), 휴먼 검출부(120), 얼굴 검출부(130), 피부 제거부(140), 의류 제거부(150)를 포함하고, 상기 3D 카메라를 이용한 머리카락 영역 검출 장치의 입력으로 들어오는 3D 카메라 영상(105)과 상기 머리카락 영역 검출 장치의 출력으로 나가는 검출된 머리카락 영역(155)을 포함한다.

[0028] 3D 카메라를 이용한 머리카락 영역 검출 장치(100)는 3D 카메라 영상(105)를 입력으로 받아, 본 발명에서 제안하는 3D 카메라를 이용한 머리카락 영역 검출 기술을 적용하여, 검출된 머리카락 영역(155)을 출력하는 장치이다.

[0029] 영상 수집부(110)에서 3D 카메라 영상(105)의 색상 영상, 깊이 영상, 적외선 영상을 수집하여 좌표를 맞춘다. 휴먼 검출부(120)에서 상기 깊이 영상을 기반으로 상기 색상 영상으로부터 배경 영역을 제거하여 휴먼 영역을 검출한다. 얼굴 검출부(130)에서 상기 휴먼 영역에서 얼굴 주변 영역을 검출한다. 피부 제거부(140)에서 상기 휴먼 영역에서 상기 얼굴 주변 영역 내에 위치한 피부영역을 제거하여 피부 제외 영역을 검출한다. 의류 제거부(150)에서 상기 적외선 영상을 기반으로 상기 피부 제외 영역에서 의류 영역을 제거하여 검출된 머리카락 영역(155)를 생성한다.

[0030] 이하에서는 3D 카메라를 Intel RealSense D400 시리즈를 사용하였을 경우의 일 예를 들어, 본 발명상의 3D 카메라를 이용한 머리카락 영역 검출 장치의 동작을 보다 상세히 설명한다.

[0031] 영상 수집부(110)는 색상 영상, 깊이 영상, 적외선 영상을 수집하기 위해서는 Visual Studio 2015 professional 버전의 환경에서 github에 있는 librealsense2 라이브러리를 사용할 수 있다. 컬러 영상, 깊이 영상, 적외선 영상은 get_color_frame(), get_depth_frame(), first(RS2_STREAM_INFRARED)를 이용해 얻을 수 있다. 상기 세 영상을 얻은 후 색상 영상과 깊이 영상의 좌표를 맞추기 위해 캘리브레이션(calibration)하는 과정이 필요하다. 각 영상의 좌표를 맞추는 것은 librealsense2의 align()함수를 이용하면 쉽게 할 수 있는데, 우선 색상 영상을 깊이 영상에 맞출 것인지 깊이 영상을 색상 영상에 맞출 것인지를 정해야 한다. Intel RealSense D400 시리즈는 카메라 디자인상 출시할 때 적외선 영상이 깊이 영상에 맞춰져 있으므로, 색상 영상을 깊이 영상에 align()함수를 통해 맞춰주고, 적외선 영상은 따로 조작을 가하지 않고 그대로 쓰면 된다.

[0032] 휴먼 검출부(120)는 깊이 영상만을 기반으로 색상 영상으로부터 배경 영역을 제거하여 휴먼 영역을 검출하는 장치이다. 구체적으로 휴먼 검출부는 깊이 영상의 깊이 값이 1보다 작거나 같은 경우는 휴먼 영역으로 저장하고, 상기 깊이 영상의 깊이 값이 1보다 큰 경우는 배경영역으로 저장하는 동작을 수행한다.

[0033] 휴먼 검출에 대한 알고리즘은 다양하지만 깊이 영상만을 이용하여 휴먼만 남기고 배경은 제거하는 알고리즘이 가장 일반적이다. 배경의 깊이 값은 휴먼의 깊이 값보다 훨씬 크기 때문에 배경이 휴먼과 붙어 있지 않는 한 휴먼만 남기고 배경을 지울 수 있다. 구체적으로는 distance = 1.f로 설정하여, 1m 이내에 있지 않은 깊이 값은 모두 지워 버리거나 혹은 value = 211 과 같이 아주 큰 값으로 바꾼다. 이렇게 깊이 영상만을 이용하여 배경을 지우게 되면, 휴먼 주위에 약간의 일부 배경이 붙는 현상이 생길 수 있다. 이는 카메라가 일부 배경의 깊이 값을 인식을 못하기 때문이다. 그러나 이러한 현상은 극히 일부에서 일어나고 그 발생빈도도 크지 않기 때문에 머리카락 영역을 검출하는데 있어서는 큰 영향을 주지 않는다.

[0034] 얼굴 검출부(130)는 상기 휴먼 영역에서 얼굴 주변 영역을 검출하는 장치이다.

[0035] 얼굴 검출부(130)는 랜드마크 알고리즘을 사용하여 복수개의 랜드마크 포인트를 자동으로 검출할 수 있다. 랜드마크 포인트((Landmark Point)는 얼굴의 각 구성요소들의 위치로, 보통의 경우 총 64개가 검출된다. 그리고 적어도 하나 이상의 랜드마크 포인트로부터 얼굴 주변 영역을 검출할 수 있는데, 모든 랜드마크 포인트의 중심점(대부분 코의 위치)을 통해 대략적인 얼굴을 해당하는 얼굴 주변 영역을 검출할 수 있다. 즉, 상기 중심점을 중심으로 모든 랜드마크 포인트를 포함하는 사각형 상자 박스를 얼굴 주변 영역으로 정할 수 있다.

[0036] 그리고 새로운 정보를 얻기 위해 랜드마크 포인트를 이용해 얼굴의 외곽선을 검출하고, 상기 얼굴의 외곽선 정보는 이후 색 기반으로 얼굴 내의 피부를 지울 때 사용할 수 있다. 구체적으로는 OpenCV 함수의 Polyline으로 랜드마크 포인트의 몇 개의 지점(얼굴의 틀을 대표하는 points들의 집합)을 저장한 집합을 대입해서 넣으면 외곽선이 그려지게 되고 fillConvexPoly를 통해 Polyline 안에 있는 픽셀들을 전부 지울 수 있다.

[0037] 피부 제거부(140)는 상기 휴먼 영역에서 상기 얼굴 주변 영역 내에 위치한 피부 영역을 제거하여 피부 제외 영역을 검출하는 장치이다.

[0038] 피부를 제거하기 위한 방법에는 여러 방법이 있지만 보통 색을 기반으로 제거하는 방법이 주로 사용되고 있다. 색을 기반으로 피부를 제거할 경우, YUV 색상 모델을 사용하는 것이 바람직하다. YUV 색상 모델은 색차 신호와 밝기 신호를 분리한 색상 모델로 RGB, YIV, HSI, YIQ 등의 다른 색상 모델보다 피부 제거 효과가 뛰어나기 때문이다.

[0039] YUV 색상 모델을 사용하여 피부를 제거할 경우, Y값은 밝기이므로 피부 영역을 결정하는 데에 있어 사용하지 않고, $77 \leq U \leq 127$, $133 \leq V \leq 173$이 피부 영역으로 설정하여 상기 피부 영역을 제거할 수 있다. 이상과 같이 YUV 색상 모델을 사용하여 얼굴 주변 영역에서 피부 영역을 제거할 경우, 머리카락 영역 중 피부 영역과 겹치는 부분이 지워질 수 있다. 이 문제는 레이블링을 사용해 얼굴 주변 영역에 해당하는 레이블만 지우는 방법으로 해결할 수 있다. 구체적으로는 코 주변에 위치해 있는 임의의 랜드마크 포인트의 레이블을 가져오면 그 레이블이 바로 피부 영역을 대표하는 레이블이 된다. 이 레이블에 해당하는 피부 영역을 지우고 지워진 머리 영역을 원본을 이용하여 다시 복원하면 피부 영역만 제거할 수 있다.

[0040] 그러나 이상과 같은 레이블링 방법을 사용하여 피부를 제거하는 경우에도 피부와 겹쳐진 머리카락 영역이 함께 제거될 수 있다. 이와 같은 문제는 앞서 얼굴 검출부(130)에서 랜드마크 포인트를 이용해 검출한 얼굴의 외곽선 정보를 이용하여, 얼굴의 외곽선 내부에 위치하는 피부 영역만을 레이블링 방법을 사용하여 지우는 것으로써 해결될 수 있다.

[0041] 의류 제거부(150)는 적외선 영상을 기반으로 상기 피부 제외 영역에서 의류영역을 제거하여 머리카락 영역을 검출하는 장치이다.

[0042] 이 때 앞서 피부 제거부(140)에서 얼굴 주변 영역내의 피부를 제거하였기 때문에 레이블링 과정을 다시 진행하는 것만으로도 의류 영역을 제거할 수 있다. 즉, 남자의 경우 보통 머리카락의 길이가 짧아 머리카락 영역과 의류 영역이 서로 떨어져 있다. 이때 지워진 피부와 배경의 픽셀 값은 0이기 때문에 0을 기점으로 레이블링을 실시하면 머리카락 영역을 확실히 추출할 수 있다. 그러나 여자의 경우 보통 머리카락의 길이가 길어 머리카락 영역과 의류 영역이 겹쳐 있기 때문에 위와 같은 레이블링을 통해서만은 의류 영역 제거가 불가능하다.

[0043] 따라서 이러한 문제점을 해결하기 위해 적외선 영상을 사용할 수 있다. Intel Real RealSense D400 시리즈의 탑재된 적외선 영상의 경우 패턴화된 구조의 광(Structured Light, 이하 '구조광'이라고만 함)을 사용하는데, 특

히 의류 영역에 분명하고 구조광이 많이 분포되어 있다. 또한 머리카락 색과는 상관 없이 머리카락 영역에는 구조광이 없다. 따라서 구조광을 이용하면 의류 영역을 제거할 수 있다.

[0044] 먼저 얼굴 경계 부분 중 턱 살짝 위쪽을 가리키는 랜드마크 포인트(CHIN.Y)를 찾는다. 그리고 적외선 영상에서 상기 랜드마크 포인트의 아래 부분에 대하여만 구조광을 추출한다. 의류 영역을 나타내는 구조광은 다른 구조광보다 그 밝기가 크기 때문에 특정 값 이상인 픽셀은 구조광으로 처리하는 방법도 있다. 그러나 이 구조광의 밝기는 상황에 따라, 사람에 따라 다 달라지고, 거리가 멀어질수록 그 밝기는 점차 줄어들어 특정 값을 통해 구조광을 추출하는 것은 바람직하지 못하다. 따라서 이상과 같은 문제를 해결하기 위하여 흑백(Grayscale) 영상을 적외선 영상과 함께 사용하는 방법을 제안한다.

[0045] Intel RealSense D400의 탑재된 적외선 영상은 흑백(Grayscale) 영상과 매우비슷한 양상을 띠고 있는데, 흑백 영상에서 밝은 부분이 적외선 영상에서도 밝다. 이러한 점에 착안하여 흑백 영상과 적외선 영상의 값의 차이의 절대값($F_{StructuredLight} = F_{SL} = FSL$)이 특정 값보다 큰 픽셀을 구조광으로 추출할 수가 있다. FSL을 계산하는 수학식은 하기 수학식 1과 같다.

수학식 1

[0046]
$$F_{StructuredLight} = \sum_{i}^{h}\sum_{j}^{w}(|GRAY(i,j) - IR(i,j)|)$$

[0047] 위 식에서 GRAY(i, j)는 흑백 영상의 (i, j) 좌표에서의 값을, IR(i, j)는 적외선 영상의 (i, j) 좌표에서의 값을, h는 입력 영상의 height, w는 width, F_{SL}은 $F_{StructuredLight}$를 의미한다.

[0048] 이 때, FSL이 특정한 임계값보다 크면 구조광으로 볼 수 있다. 실험을 통해 상기 임계값이 40일 경우, 구조광 추출의 정확도가 가장 높음을 확인하였다. 따라서 하기 수학식 2와 같이 FSL이 임계값인 40보다 큰 영역은 $F_{remainlight}$을 255로 변경하고, FSL이 40보다 작거나 같은 영역은 $F_{remainlight}$을 0으로 변경할 수 있다. 이렇게 FSL값을 $F_{remainlight}$로 변경해 저장하면, $F_{remainlight}$이 0이 아니기만 하면 구조광으로 볼 수 있어 구조광 추출에 용이하다.

수학식 2

[0049]
$$F_{remainlight}\begin{cases}255 & (F_{SL} > 40)\\ 0 & (F_{SL} \leq 40)\end{cases}$$

[0050] 위 식에서 F_{SL}은 $F_{StructuredLight}$이고, $F_{remainlight}$는 변경된 FSL 값을 의미한다.

[0051] 본 발명에서 제안하는 의류 영역을 제거하는 방법은 구조광의 크기를 어느 일정한 초기값으로부터 점차 크게 하며 지워나가는 방식을 사용할 수 있다. 만약 구조광의 크기의 초기값을 13으로 정한다면, 가로가 13, 세로가 13인 정사각형 모양의 관심 영역(ROI=Region Of Interest)을 설정하고, Y > CHIN.Y에 대해서 적외선 영상과 흑백 영상으로부터 상기 수학식 1 및 수학식 2에 따라 FSL과 $F_{remainlight}$을 순차적으로 계산한다. 그리고 이렇게 계산하는 중에 같은 관심 영역안에 $F_{remainlight}$이 255인 픽셀을 발견하면, 관심 영역의 모든 값을 255로 변경하고, 다음 관심영역에 대하여 상기 동작을 똑같이 반복한다. 이때 주의해야 할 점은 FSL 및 $F_{remainlight}$를 계산하는 이미지는 적외선 영상이고, 255의 값으로 변경하는 이미지는 앞서 피부 영역을 제거한 색상 영상이다. 그리고 구조광의 크기를 점차 증가해가며 상기 동작을 반복해 의류 영역을 제거한다.

[0052] 이상과 같이 의류 영역을 제거할 때, 구조광의 크기의 초기값이 중요하다. 사람(휴먼)이 카메라와 가까이 있으면 영상에서 두 눈 사이의 거리는 멀어지지만 카메라와 멀리 있으면 두 눈 사이의 거리는 좁아지므로, 이 간격을 통해 구조광의 크기의 초기값을 결정할 수 있다. 구체적으로는 구조광의 크기의 초기값은 두 눈 사이의 거리가 30픽셀 미만의 경우 3, 30에서 50픽셀 사이의 경우 5, 50에서 70픽셀 사이의 경우는 7로, 20픽셀 간격으로

초기값을 2씩 증가시켜 두 눈 사이의 거리가 150픽셀을 초과하면 15로 할 수 있다.

[0053] 이상과 같은 구조광 추출 방식으로 구조광을 추출한 경우 구조광 외의 다른 잡티 또한 많이 추출된다는 문제가 있지만, 앞서 설명한 바와 같이 머리카락 영역에는 구조광이 존재하지 않으므로 잘못 추출된 잡티를 구조광으로 취급해 부풀려서 지운다고 해도 머리카락 영역 검출에는 문제가 되지 않는다.

[0054] 그리고 적외선 영상에서 머리카락 영역 안에 작은 노이즈가 발생하였을 경우, 구조광을 추출하는 과정에서 머리카락 영역 안에 구멍이 생길 수 있다. 이러한 구멍은 머리카락 영역의 외곽선을 그리고 그 외곽선 안쪽에 해당하는 영역을 다시 복원함으로써 해결할 수 있다.

[0056] 도 2는 실시예에 따른 3D 카메라를 이용한 머리카락 영역 검출 방법의 일 예를 나타낸 동작 흐름도이다.

[0057] 도 2를 참조하면, 우선 영상 수집부에 3D 카메라 영상의 색상 영상, 깊이 영상, 적외선 영상이 입력된다(S210). 휴먼 검출부에서 상기 깊이 영상을 기반으로 상기 색상 영상으로부터 배경 영역을 제거하여 휴먼 영역을 검출한다(S220). 얼굴 검출부에서 상기 휴먼 영역에서 얼굴 주변 영역을 검출한다(S230). 피부 제거부에서 상기 휴먼 영역에서 상기 얼굴 주변 영역 내에 위치한 피부영역을 제거하여 피부 제외 영역을 검출한다(S240). 의류 제거부에서 상기 적외선 영상을 기반으로 상기 피부 제외 영역에서 의류 영역을 제거하여 검출된 머리카락 영역을 생성한다(S250).

[0058] 이하에서는 3D 카메라를 Intel RealSense D400 시리즈를 사용하였을 경우의 일 예를 들어, 본 발명상의 3D 카메라를 이용한 머리카락 영역 검출 방법을 보다 상세히 설명한다.

[0059] 영상 수집부는 색상 영상, 깊이 영상, 적외선 영상을 수집하는데(S210), 이때 Visual Studio 2015 professional 버전의 환경에서 github에 있는 librealsense2 라이브러리를 사용하여, 컬러 영상, 깊이 영상, 적외선 영상은 get_color_frame(), get_depth_frame(), first(RS2_STREAM_INFRARED)를 이용해 얻을 수 있다. 상기 세 영상을 얻은 후 색상 영상과 깊이 영상의 좌표를 맞추기 위해 캘리브레이션(calibration)하는 과정이 필요하다. 각 영상의 좌표를 맞추는 것은 librealsense2의 align()함수를 이용하면 쉽게 할 수 있는데, 우선 색상 영상을 깊이 영상에 맞출 것인지 깊이 영상을 색상 영상에 맞출 것인지를 정해야 한다. Intel RealSense D400 시리즈는 카메라 디자인상 출시할 때 적외선 영상이 깊이 영상에 맞춰져 있으므로, 색상 영상을 깊이 영상에 align()함수를 통해 맞춰주고, 적외선 영상은 따로 조작을 가하지 않고 그대로 쓰면 된다.

[0060] 휴먼 검출부는 깊이 영상만을 기반으로 깊이 영상을 기반으로 색상 영상으로부터 배경 영역을 제거하여 휴먼 영역을 검출한다(S220). 구체적으로 휴먼 검출부는 깊이 영상의 깊이 값이 1보다 작거나 같은 경우는 휴먼 영역으로 저장하고, 상기 깊이 영상의 깊이 값이 1보다 큰 경우는 배경영역으로 저장하는 동작을 수행한다.

[0061] 휴먼 검출에 대한 알고리즘은 다양하지만 깊이 영상만을 이용하여 휴먼만 남기고 배경은 제거하는 알고리즘이 가장 일반적이다. 배경의 깊이 값은 휴먼의 깊이 값보다 훨씬 크기 때문에 배경이 휴먼과 붙어 있지 않는 한 휴먼만 남기고 배경을 지울 수 있다. 구체적으로는 distance = 1.f로 설정하여, 1m 이내에 있지 않은 깊이 값은 모두 지워 버리거나 혹은 value = 211 과 같이 아주 큰 값으로 바꾼다. 이렇게 깊이 영상만을 이용하여 배경을 지우게 되면, 휴먼 주위에 약간의 일부 배경이 붙는 현상이 생길 수 있다. 이는 카메라가 일부 배경의 깊이 값을 인식을 못하기 때문이다. 그러나 이러한 현상은 극히 일부에서 일어나고 그 발생빈도도 크지 않기 때문에 머리카락 영역을 검출하는데 있어서는 큰 영향을 주지 않는다.

[0062] 얼굴 검출부는 상기 휴먼 영역에서 얼굴 주변 영역을 검출한다(S230). 이 때, 얼굴 검출부는 랜드마크 알고리즘을 사용하여 복수개의 랜드마크 포인트를 자동으로 검출할 수 있다. 랜드마크 포인트((Landmark Point)는 얼굴의 각 구성요소들의 위치로, 보통의 경우 총 64개가 검출된다. 그리고 적어도 하나 이상의 랜드마크 포인트로부터 얼굴 주변 영역을 검출할 수 있는데, 모든 랜드마크 포인트의 중심점(대부분 코의 위치)을 통해 대략적인 얼굴을 해당하는 얼굴 주변 영역을 검출할 수 있다. 즉, 상기 중심점을 중심으로 모든 랜드마크 포인트를 포함하는 사각형 상자 박스를 얼굴 주변 영역으로 정할 수 있다.

[0063] 그리고 새로운 정보를 얻기 위해 랜드마크 포인트를 이용해 얼굴의 외곽선을 검출하고, 상기 얼굴의 외곽선 정보는 이후 색 기반으로 얼굴 내의 피부를 지울 때 사용할 수 있다. 구체적으로는 OpenCV 함수의 Polyline으로 랜드마크 포인트의 몇 개의 지점(얼굴의 틀을 대표하는 points들의 집합)을 저장한 집합을 대입해서 넣으면 외곽선이 그려지게 되고 fillConvexPoly를 통해 Polyline 안에 있는 픽셀들을 전부 지울 수 있다.

[0064] 피부 제거부는 상기 휴먼 영역에서 상기 얼굴 주변 영역 내에 위치한 피부 영역을 제거하여 피부 제외 영역을

검출한다(S240).

[0065] 피부를 제거하기 위한 방법에는 여러 방법이 있지만 보통 색을 기반으로 제거하는 방법이 주로 사용되고 있다. 색을 기반으로 피부를 제거할 경우, YUV 색상 모델을 사용하는 것이 바람직하다. YUV 색상 모델은 색차 신호와 밝기 신호를 분리한 색상 모델로 RGB, YIV, HSI, YIQ 등의 다른 색상 모델보다 피부 제거 효과가 뛰어나기 때문이다.

[0066] YUV 색상 모델을 사용하여 피부를 제거할 경우, Y값은 밝기이므로 피부 영역을 결정하는 데에 있어 사용하지 않고, $77 \leq U \leq 127$, $133 \leq V \leq 173$이 피부 영역으로 설정하여 상기 피부 영역을 제거할 수 있다. 이상과 같이 YUV 색상 모델을 사용하여 얼굴 주변 영역에서 피부 영역을 제거할 경우, 머리카락 영역 중 피부 영역과 겹치는 부분이 지워질 수 있다. 이 문제는 레이블링을 사용해 얼굴 주변 영역에 해당하는 레이블만 지우는 방법으로 해결할 수 있다. 구체적으로는 코 주변에 위치해 있는 임의의 랜드마크 포인트의 레이블을 가져오면 그 레이블이 바로 피부 영역을 대표하는 레이블이 된다. 이 레이블에 해당하는 피부 영역을 지우고 지워진 머리 영역을 원본을 이용하여 다시 복원하면 피부 영역만 제거할 수 있다.

[0067] 그러나 이상과 같은 레이블링 방법을 사용하여 피부를 제거하는 경우에도 피부와 겹쳐진 머리카락 영역이 함께 제거될 수 있다. 이와 같은 문제는 앞서 얼굴 검출부(130)에서 랜드마크 포인트를 이용해 검출한 얼굴의 외곽선 정보를 이용하여, 얼굴의 외곽선 내부에 위치하는 피부 영역만을 레이블링 방법을 사용하여 지우는 것으로써 해결될 수 있다.

[0068] 의류 제거부는 적외선 영상을 기반으로 상기 피부 제외 영역에서 의류영역을 제거하여 머리카락 영역을 검출한다(S250).

[0069] 이 때 앞서 피부 제거부(140)에서 얼굴 주변 영역내의 피부를 제거하였기 때문에 레이블링 과정을 다시 진행하는 것만으로도 의류 영역을 제거할 수 있다. 즉, 남자의 경우 보통 머리카락의 길이가 짧아 머리카락 영역과 의류 영역이 서로 떨어져 있다. 이때 지워진 피부와 배경의 픽셀 값은 0이기 때문에 0을 기점으로 레이블링을 실시하면 머리카락 영역을 확실히 추출할 수 있다. 그러나 여자의 경우 보통 머리카락의 길이가 길어 머리카락 영역과 의류 영역이 겹쳐 있기 때문에 위와 같은 레이블링을 통해서만은 의류 영역 제거가 불가능하다.

[0070] 따라서 이러한 문제점을 해결하기 위해 적외선 영상을 사용할 수 있다. Intel Real RealSense D400 시리즈의 탑재된 적외선 영상의 경우 패턴화된 구조의 광(Structured Light, 이하 '구조광'이라고만 함)을 사용하는데, 특히 의류 영역에 분명하고 구조광이 많이 분포되어 있다. 또한 머리카락 색과는 상관 없이 머리카락 영역에는 구조광이 없다. 따라서 구조광을 이용하면 의류 영역을 제거할 수 있다.

[0071] 먼저 얼굴 경계 부분 중 턱 살짝 위쪽을 가리키는 랜드마크 포인트(CHIN.Y)를 찾는다. 그리고 적외선 영상에서 상기 랜드마크 포인트의 아래 부분에 대하여만 구조광을 추출한다. 의류 영역을 나타내는 구조광은 다른 구조광보다 그 밝기가 크기 때문에 특정 값 이상인 픽셀은 구조광으로 처리하는 방법도 있다. 그러나 이 구조광의 밝기는 상황에 따라, 사람에 따라 다 달라지고, 거리가 멀어질수록 그 밝기는 점차 줄어들어 특정 값을 통해 구조광을 추출하는 것은 바람직하지 못하다. 따라서 이상과 같은 문제를 해결하기 위하여 흑백(Grayscale) 영상을 적외선 영상과 함께 사용하는 방법을 제안한다.

[0072] Intel RealSense D400의 탑재된 적외선 영상은 흑백(Grayscale) 영상과 매우 비슷한 양상을 띠고 있는데, 흑백 영상에서 밝은 부분이 적외선 영상에서도 밝다. 이러한 점에 착안하여 상기 수학식 1과 같이 흑백 영상과 적외선 영상의 값의 차이의 절대값($F_{StructuredLight} = F_{SL} = FSL$)이 특정 값보다 큰 픽셀을 구조광으로 추출할 수가 있다.

[0073] 이 때, FSL이 특정한 임계값보다 크면 구조광으로 볼 수 있다. 실험을 통해 상기 임계값이 40일 경우, 구조광 추출의 정확도가 가장 높음을 확인하였다. 따라서 상기 수학식 2와 같이 FSL이 임계값인 40보다 큰 영역은 $F_{remainlight}$을 255로 변경하고, FSL이 40보다 작거나 같은 영역은 $F_{remainlight}$을 0으로 변경할 수 있다. 이렇게 FSL값을 변경해 저장하면, $F_{remainlight}$이 0이 아니기만 하면 구조광으로 볼 수 있는 장점이 있다.

[0074] 본 발명에서 제안하는 의류 영역을 제거하는 방법은 구조광의 크기를 어느 일정한 초기값으로부터 점차 크게 하며 지워나가는 방식을 사용할 수 있다. 만약 구조광의 크기의 초기값을 13으로 정한다면, 가로가 13, 세로가 13인 정사각형 모양의 관심 영역(ROI=Region Of Interest)을 설정하고, Y > CHIN.Y에 대해서 적외선 영상과 흑백 영상으로부터 상기 수학식 1 및 수학식 2에 따라 FSL과 $F_{remainlight}$을 순차적으로 계산한다. 그리고 이렇게 계산하는 중에 같은 관심 영역안에 $F_{remainlight}$이 255인 픽셀을 발견하면, 관심 영역의 모든 값을 255로 변경하고, 다음

관심영역에 대하여 상기 동작을 똑같이 반복한다. 이때 주의해야 할 점은 FSL 및 $F_{remainlight}$를 계산하는 이미지는 적외선 영상이고, 255의 값으로 변경하는 이미지는 앞서 피부 영역을 제거한 색상 영상이다. 그리고 구조광의 크기를 점차 증가해가며 상기 동작을 반복해 의류 영역을 제거한다.

[0075] 이상과 같이 의류 영역을 제거할 때, 구조광의 크기의 초기값이 중요하다. 사람(휴먼)이 카메라와 가까이 있으면 영상에서 두 눈 사이의 거리는 멀어지지만 카메라와 멀리 있으면 두 눈 사이의 거리는 좁아지므로, 이 간격을 통해 구조광의 크기의 초기값을 결정할 수 있다. 구체적으로는 구조광의 크기의 초기값은 두 눈 사이의 거리가 30픽셀 미만의 경우 3, 30에서 50픽셀 사이의 경우 5, 50에서 70픽셀 사이의 경우는 7로, 20픽셀 간격으로 초기값을 2씩 증가시켜 두 눈 사이의 거리가 150픽셀을 초과하면 15로 할 수 있다.

[0076] 이상과 같은 구조광 추출 방식으로 구조광을 추출한 경우 구조광 외의 다른 잡티 또한 많이 추출된다는 문제가 있지만, 앞서 설명한 바와 같이 머리카락 영역에는 구조광이 존재하지 않으므로 잘못 추출된 잡티를 구조광으로 취급해 부풀려서 지운다고 해도 머리카락 영역 검출에는 문제가 되지 않는다.

[0077] 그리고 적외선 영상에서 머리카락 영역 안에 작은 노이즈가 발생하였을 경우, 구조광을 추출하는 과정에서 머리카락 영역 안에 구멍이 생길 수 있다. 이러한 구멍은 머리카락 영역의 외곽선을 그리고 그 외곽선 안쪽에 해당하는 영역을 다시 복원함으로써 해결할 수 있다.

[0079] 도 3은 도 2에 도시된 영상 입력 단계(S210)의 일 예를 나타낸 동작 흐름도이다.

[0080] 도 3을 참조하면, 상기 영상 입력 단계는 3D 카메라로부터 색상 영상, 깊이 영상, 적외선 영상을 불러오는 단계(S310), 색상 영상의 기준 좌표를 깊이 영상의 기준 좌표에 맞추는 단계(S320), 기준 좌표가 맞춰진 색상 영상, 깊이 영상, 적외선 영상을 저장하는 단계(S330)을 포함한다.

[0081] 영상 입력 단계(S210)에서 영상 수집부는 색상 영상, 깊이 영상, 적외선 영상을 수집하기 위해서는 Visual Studio 2015 professional 버전의 환경에서 github에 있는 librealsense2 라이브러리를 사용할 수 있다. 컬러 영상, 깊이 영상, 적외선 영상은 get_color_frame(), get_depth_frame(), first(RS2_STREAM_INFRARED)를 이용해 얻을 수 있다. 상기 세 영상을 얻은 후 색상 영상과 깊이 영상의 좌표를 맞추기 위해 캘리브레이션(calibration)하는 과정이 필요하다. 각 영상의 좌표를 맞추는 것은 librealsense2의 align()함수를 이용하면 쉽게 할 수 있는데, 우선 색상 영상을 깊이 영상에 맞출 것인지 깊이 영상을 색상 영상에 맞출 것인지를 정해야 한다. Intel RealSense D400 시리즈는 카메라 디자인상 출시할 때 적외선 영상이 깊이 영상에 맞춰져 있으므로, 색상 영상을 깊이 영상에 align()함수를 통해 맞춰주고, 적외선 영상은 따로 조작을 가하지 않고 그대로 쓰면 된다.

[0083] 도 4는 도 2에 도시된 휴먼 검출 단계(S220)의 일 예를 나타낸 동작 흐름도이다.

[0084] 도 4를 참조하면, 휴먼 검출부에 기준 좌표가 맞춰진 색상 영상이 입력된다(S410). 입력된 색상영상에 대하여 깊이 값이 1.0보다 작거나 같은지를 판단한다(S420). 만약 깊이 값이 1.0보다 크면 배경영역으로 저장하고(S430), 작거나 같으면 휴먼영역으로 저장한다(S440). 그리고 배경과 분리된 휴먼 영역의 영상을 저장한다(S450).

[0085] 휴먼 검출에 대한 알고리즘은 다양하지만 깊이 영상만을 이용하여 휴먼만 남기고 배경은 제거하는 알고리즘이 가장 일반적이다. 배경의 깊이 값은 휴먼의 깊이 값보다 훨씬 크기 때문에 배경이 휴먼과 붙어 있지 않는 한 휴먼만 남기고 배경을 지울 수 있다. 구체적으로는 distance = 1.f로 설정하여, 1m 이내에 있지 않은 깊이 값은 모두 지워 버리거나 혹은 value = 211 과 같이 아주 큰 값으로 바꾼다. 이렇게 깊이 영상만을 이용하여 배경을 지우게 되면, 휴먼 주위에 약간의 일부 배경이 붙는 현상이 생길 수 있다. 이는 카메라가 일부 배경의 깊이 값을 인식을 못하기 때문이다. 그러나 이러한 현상은 극히 일부에서 일어나고 그 발생빈도도 크지 않기 때문에 머리카락 영역을 검출하는데 있어서는 큰 영향을 주지 않는다.

[0087] 도 5는 도 2에 도시된 얼굴 검출 단계(S230)의 일 예를 나타낸 동작 흐름도이다.

[0088] 도 5를 참조하면, 우선 얼굴 검출부에 배경과 분리된 휴먼 영상이 입력된다(S510). 얼굴 검출부는 랜드마크 알

고리즘을 이용하여 랜드마크 포인트를 검출한다(S520). 상기 랜드마크 포인트의 중심점(주로 코의 위치)을 기준으로 얼굴 주변 영역을 검출한다(S530). 그리고 얼굴 주변 영역 검출에 성공했는지 여부를 판단하고(S540), 만약 얼굴 주변 영역 검출에 성공하지 못하였다면 다시 랜드마크 포인트 중심점을 기준으로 얼굴 주변 영역을 검출하는 단계(S530)로 돌아간다. 만약 얼굴 주변 영역 검출에 성공하였다면 얼굴 주변 영역 및 랜드마크 포인트를 저장한다(S550).

[0089] 이 때, 얼굴 검출부는 랜드마크 알고리즘을 사용하여 복수개의 랜드마크 포인트를 자동으로 검출할 수 있다. 랜드마크 포인트((Landmark Point)는 얼굴의 각 구성요소들의 위치로, 보통의 경우 총 64개가 검출된다. 그리고 적어도 하나 이상의 랜드마크 포인트로부터 얼굴 주변 영역을 검출할 수 있는데, 모든 랜드마크 포인트의 중심점(대부분 코의 위치)을 통해 대략적인 얼굴을 해당하는 얼굴 주변 영역을 검출할 수 있다. 즉, 상기 중심점을 중심으로 모든 랜드마크 포인트를 포함하는 사각형 상자 박스를 얼굴 주변 영역으로 정할 수 있다.

[0090] 그리고 새로운 정보를 얻기 위해 랜드마크 포인트를 이용해 얼굴의 외곽선을 검출하고, 상기 얼굴의 외곽선 정보는 이후 색 기반으로 얼굴 내의 피부를 지울 때 사용할 수 있다. 구체적으로는 OpenCV 함수의 Polyline으로 랜드마크 포인트의 몇 개의 지점(얼굴의 틀을 대표하는 points들의 집합)을 저장한 집합을 대입해서 넣으면 외곽선이 그려지게 되고 fillConvexPoly를 통해 Polyline 안에 있는 픽셀들을 전부 지울 수 있다.

[0092] 도 6는 도 2에 도시된 피부 검출 및 제거 단계(S240)의 일 예를 나타낸 동작 흐름도이다.

[0093] 도 6을 참조하면, 우선 피부 검출부에 배경과 분리된 휴먼 영상이 입력된다(S610). 그리고 YUV 색상 모델을 이용하여 Y, U, V를 구한다(S620). 이때 구해진 U와 V 가 피부 색상의 범위인 77≤U≤127, 133≤V≤173의 범위안에 들어가는지를 판단한다(S630). 만약 구해진 U와 V 가 위 범위안에 포함되면 피부로 결정되고(S650), 위 범위안에 포함되지 않으면 피부가 아닌 것으로 결정된다(S640). 랜드마크 포인트로부터 검출한 얼굴 외곽선 내부에서 앞서 피부로 결정된 영역만을 제거한다(S660). 그리고 피부가 제거된 휴먼 영상을 저장한다(S670).

[0094] 피부 검출 및 제거 단계(S240)에서 피부 제거부(140)는 상기 휴먼 영역에서 상기 얼굴 주변 영역 내에 위치한 피부 영역을 제거하여 피부 제외 영역을 검출할 수 있다.

[0095] 피부를 제거하기 위한 방법에는 여러 방법이 있지만 보통 색을 기반으로 제거하는 방법이 주로 사용되고 있다. 색을 기반으로 피부를 제거할 경우, YUV 색상 모델을 사용하는 것이 바람직하다. YUV 색상 모델은 색차 신호와 밝기 신호를 분리한 색상 모델로 RGB, YIV, HSI, YIQ 등의 다른 색상 모델보다 피부 제거 효과가 뛰어나기 때문이다.

[0096] YUV 색상 모델을 사용하여 피부를 제거할 경우, Y값은 밝기이므로 피부 영역을 결정하는 데에 있어 사용하지 않고, 77≤U≤127, 133≤V≤173이 피부 영역으로 설정하여 상기 피부 영역을 제거할 수 있다. 이상과 같이 YUV 색상 모델을 사용하여 얼굴 주변 영역에서 피부 영역을 제거할 경우, 머리카락 영역 중 피부 영역과 겹치는 부분이 지워질 수 있다. 이 문제는 레이블링을 사용해 얼굴 주변 영역에 해당하는 레이블만 지우는 방법으로 해결할 수 있다. 구체적으로는 코 주변에 위치해 있는 임의의 랜드마크 포인트의 레이블을 가져오면 그 레이블이 바로 피부 영역을 대표하는 레이블이 된다. 이 레이블에 해당하는 피부 영역을 지우고 지워진 머리 영역을 원본을 이용하여 다시 복원하면 피부 영역만 제거할 수 있다.

[0097] 그러나 이상과 같은 레이블링 방법을 사용하여 피부를 제거하는 경우에도 피부와 겹쳐진 머리카락 영역이 함께 제거될 수 있다. 이와 같은 문제는 앞서 얼굴 검출부(130)에서 랜드마크 포인트를 이용해 검출한 얼굴의 외곽선 정보를 이용하여, 얼굴의 외곽선 내부에 위치하는 피부 영역만을 레이블링 방법을 사용하여 지우는 것으로써 해결될 수 있다.

[0099] 도 7은 도 2에 도시된 의류 검출 및 제거 단계(S250)의 일 예를 나타낸 동작 흐름도이다.

[0100] 도 7을 참조하면, 우선 의류 제거부에 피부가 제거된 휴먼 영상이 입력된다(S710). 적외선 영상을 이용하여 FSL을 구한다(S720). 상기 FSL이 40보다 크고, Y좌표가 CHIN.Y(얼굴 경계 부분의 턱 살짝 위쪽을 가리키는 랜드마크 포인트)보다 큰지를 판단한다(S730). 만약 어떤 관심영역이 위 판단식을 통과하면 구조광으로 결정되고(S750), 위 판단식을 통과하지 못하면 구조광이 아닌 것으로 결정되어(S740), 구조광의 위치를 추출한다. 이렇게 추출된 구조광의 크기를 점점 크게 하여 구조광이 속해 있는 의류 영역을 제거한다(S760). 이렇게 의류 영역

을 제거함으로써 머리 카락 영역을 추출해 내 저장하게 된다(S770).

[0101] 의류 제거부는 앞서 피부 제거부에서 얼굴 주변 영역내의 피부를 제거하였기 때문에 레이블링 과정을 다시 진행하는 것만으로도 의류 영역을 제거할 수도 있다. 즉, 남자의 경우 보통 머리카락의 길이가 짧아 머리카락 영역과 의류 영역이 서로 떨어져 있다. 이때 지워진 피부와 배경의 픽셀 값은 0이기 때문에 0을 기점으로 레이블링을 실시하면 머리카락 영역을 확실히 추출할 수 있다. 그러나 여자의 경우 보통 머리카락의 길이가 길어 머리카락 영역과 의류 영역이 겹쳐 있기 때문에 위와 같은 레이블링을 통해서만은 의류 영역 제거가 불가능하다.

[0102] 따라서 이러한 문제점을 해결하기 위해 적외선 영상을 사용할 수 있다. Intel Real RealSense D400 시리즈의 탑재된 적외선 영상의 경우 패턴화된 구조의 광(Structured Light, 이하 '구조광'이라고만 함)을 사용하는데, 특히 의류 영역에 분명하고 구조광이 많이 분포되어 있다. 또한 머리카락 색과는 상관 없이 머리카락 영역에는 구조광이 없다. 따라서 구조광을 이용하면 의류 영역을 제거할 수 있다.

[0103] 먼저 얼굴 경계 부분 중 턱 살짝 위쪽을 가리키는 랜드마크 포인트(CHIN.Y)를 찾는다. 그리고 적외선 영상에서 상기 랜드마크 포인트의 아래 부분에 대하여만 구조광을 추출한다. 의류 영역을 나타내는 구조광은 다른 구조광보다 그 밝기가 크기 때문에 특정 값 이상인 픽셀은 구조광으로 처리하는 방법도 있다. 그러나 이 구조광의 밝기는 상황에 따라, 사람에 따라 다 달라지고, 거리가 멀어질수록 그 밝기는 점차 줄어들어 특정 값을 통해 구조광을 추출하는 것은 바람직하지 못하다. 따라서 이상과 같은 문제를 해결하기 위하여 흑백(Grayscale) 영상을 적외선 영상과 함께 사용하는 방법을 제안한다.

[0104] Intel RealSense D400의 탑재된 적외선 영상은 흑백(Grayscale) 영상과 매우 비슷한 양상을 띠고 있는데, 흑백 영상에서 밝은 부분이 적외선 영상에서도 밝다. 이러한 점에 착안하여 상기 수학식 1과 같이 흑백 영상과 적외선 영상의 값의 차이의 절대값($F_{StructuredLight} = F_{SL} = FSL$)이 특정 값보다 큰 픽셀을 구조광으로 추출할 수가 있다.

[0105] 이 때, FSL이 특정한 임계값보다 크면 구조광으로 볼 수 있다. 실험을 통해 상기 임계값이 40일 경우, 구조광 추출의 정확도가 가장 높음을 확인하였다. 따라서 상기 수학식 2와 같이 FSL이 임계값인 40보다 큰 영역은 $F_{remainlight}$을 255로 변경하고, FSL이 40보다 작거나 같은 영역은 $F_{remainlight}$을 0으로 변경할 수 있다. 이렇게 FSL값을 변경해 저장하면, $F_{remainlight}$이 0이 아니기만 하면 구조광으로 볼 수 있는 장점이 있다.

[0106] 본 발명에서 제안하는 의류 영역을 제거하는 방법은 구조광의 크기를 어느 일정한 초기값으로부터 점차 크게 하며 지워나가는 방식을 사용할 수 있다. 만약 구조광의 크기의 초기값을 13으로 정한다면, 가로가 13, 세로가 13인 정사각형 모양의 관심 영역(ROI=Region Of Interest)을 설정하고, Y > CHIN.Y에 대해서 적외선 영상과 흑백 영상으로부터 상기 수학식 1 및 수학식 2에 따라 FSL과 $F_{remainlight}$을 순차적으로 계산한다. 그리고 이렇게 계산하는 중에 같은 관심 영역안에 $F_{remainlight}$이 255인 픽셀을 발견하면, 관심 영역의 모든 값을 255로 변경하고, 다음 관심영역에 대하여 상기 동작을 똑같이 반복한다. 이때 주의해야 할 점은 FSL 및 $F_{remainlight}$를 계산하는 이미지는 적외선 영상이고, 255의 값으로 변경하는 이미지는 앞서 피부 영역을 제거한 색상 영상이다. 그리고 구조광의 크기를 점차 증가해가며 상기 동작을 반복해 의류 영역을 제거한다.

[0107] 이상과 같이 의류 영역을 제거할 때, 구조광의 크기의 초기값이 중요하다. 사람(휴먼)이 카메라와 가까이 있으면 영상에서 두 눈 사이의 거리는 멀어지지만 카메라와 멀리 있으면 두 눈 사이의 거리는 좁아지므로, 이 간격을 통해 구조광의 크기의 초기값을 결정할 수 있다. 구체적으로는 구조광의 크기의 초기값은 두 눈 사이의 거리가 30픽셀 미만의 경우 3, 30에서 50픽셀 사이의 경우 5, 50에서 70픽셀 사이의 경우는 7로, 20픽셀 간격으로 초기값을 2씩 증가시켜 두 눈 사이의 거리가 150픽셀을 초과하면 15로 할 수 있다.

[0108] 이상과 같은 구조광 추출 방식으로 구조광을 추출한 경우 구조광 외의 다른 잡티 또한 많이 추출된다는 문제가 있지만, 앞서 설명한 바와 같이 머리카락 영역에는 구조광이 존재하지 않으므로 잘못 추출된 잡티를 구조광으로 취급해 부풀려서 지운다고 해도 머리카락 영역 검출에는 문제가 되지 않는다.

[0109] 그리고 적외선 영상에서 머리카락 영역 안에 작은 노이즈가 발생하였을 경우, 구조광을 추출하는 과정에서 머리카락 영역 안에 구멍이 생길 수 있다. 이러한 구멍은 머리카락 영역의 외곽선을 그리고 그 외곽선 안쪽에 해당하는 영역을 다시 복원함으로써 해결할 수 있다.

[0111] 도 8은 실시예에 따른 본 발명의 활용예를 나타낸 도면이다.

[0112] 본 발명상의 3D 카메라를 이용한 머리카락 영역 검출 장치의 영상 수집부는 3D 카메라 영상을 입력으로 받고, 3D 카메라 영상의 색상 영상, 깊이 영상, 적외선 영상의 각 기준 좌표가 맞춰진다. 휴먼 검출부는 영상 수집부에서 기준 좌표가 맞춰진 상기 색상 영상(810)을 배경과 분리하여 배경과 분리된 휴먼 영상1(820)을 저장한다. 그리고 얼굴 검출부는 상기 배경과 분리된 휴먼 영상1으로부터 랜드마크 알고리즘을 사용하여 얼굴 주변 영역과 랜드마크 포인트를 검출하고, 배경과 분리된 휴먼 영상2(830)를 생성한다. 이 때, 상기 배경과 분리된 휴먼 영상2(830)의 각 얼굴 사진에서 사각형 상자로 표시된 부분이 검출된 얼굴 주변 영역이고, 점으로 표시된 부분이 랜드마크 포인트를 나타낸다. 피부 제거부는 배경과 분리된 휴먼 영상2(830)를 YUV 색상모델을 이용하여 피부 영역만을 제거한다. 이 때, 랜드마크 포인트로부터 검출한 얼굴 외곽선 내부의 피부 영역만을 제거함으로써 피부가 제거된 휴먼 영상(840)을 생성할 수 있다. 의류 제거부는 상기 피부가 제거된 휴먼 영상(840)를 적외선 영상을 이용하여 의류 영역을 나타내는 구조광만을 추출하고, 상기 구조광을 제거하여 의류 영역을 제거한 휴먼 영상(850)을 생성한다. 그리고 의류 영역을 제거한 휴먼 영상(850)으로부터 머리카락 영역만이 표시된 휴먼 영상(860)을 생성할 수 있다.

[0114] 도 9는 실시예에 따른 컴퓨터 시스템 구성을 나타낸 도면이다.

[0115] 실시예에 따른 3D 카메라를 이용한 머리카락 영역 검출 장치는 컴퓨터로 읽을 수 있는 기록매체와 같은 컴퓨터 시스템(900)에서 구현될 수 있다.

[0116] 컴퓨터 시스템(900)은 버스(920)를 통하여 서로 통신하는 하나 이상의 프로세서(910), 메모리(930), 사용자 인터페이스 입력 장치(940), 사용자 인터페이스 출력 장치(950) 및 스토리지(960)를 포함할 수 있다. 또한, 컴퓨터 시스템(900)은 네트워크(980)에 연결되는 네트워크 인터페이스(970)를 더 포함할 수 있다. 프로세서(910)는 중앙 처리 장치 또는 메모리(930)나 스토리지(960)에 저장된 프로그램 또는 프로세싱 인스트럭션들을 실행하는 반도체 장치일 수 있다. 메모리(930) 및 스토리지(960)는 휘발성 매체, 비휘발성 매체, 분리형 매체, 비분리형 매체, 통신 매체, 또는 정보 전달 매체 중에서 적어도 하나 이상을 포함하는 저장 매체일 수 있다. 예를 들어, 메모리(930)는 ROM(931)이나 RAM(932)을 포함할 수 있다.

[0117] 이상에서 첨부된 도면을 참조하여 본 발명의 실시예들을 설명하였지만, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다.

부호의 설명

[0118] 100: 3D 카메라를 이용한 머리카락 영역 검출 장치

105: 3D 카메라 영상 110: 영상 수집부

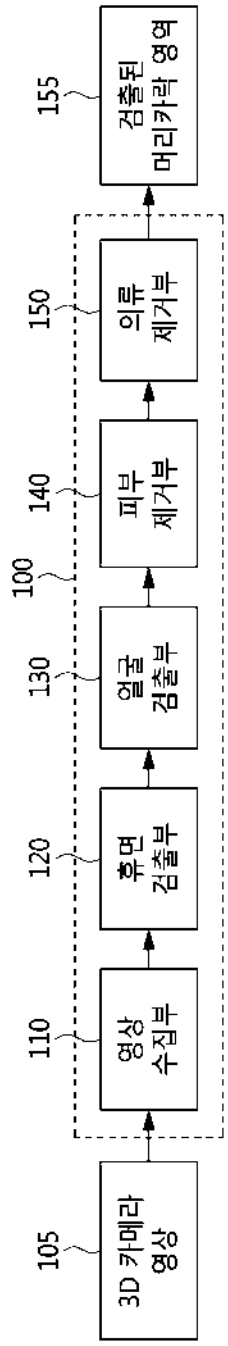
120: 휴먼 검출부 130: 얼굴 검출부

140: 피부 제거부 150: 의류 제거부

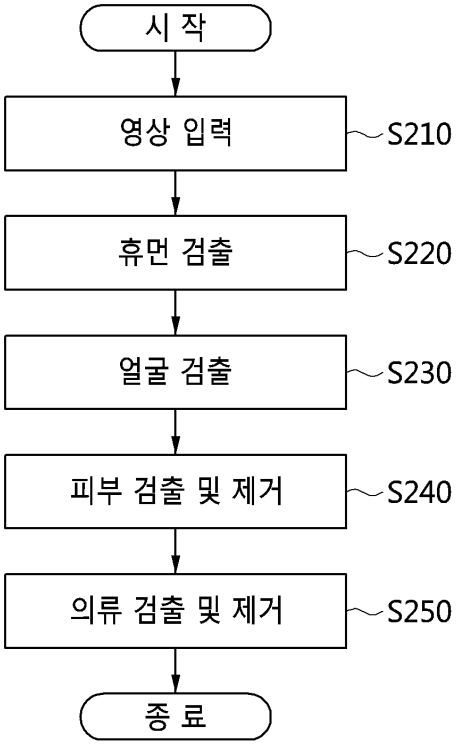
155: 검출된 머리카락 영역

도면

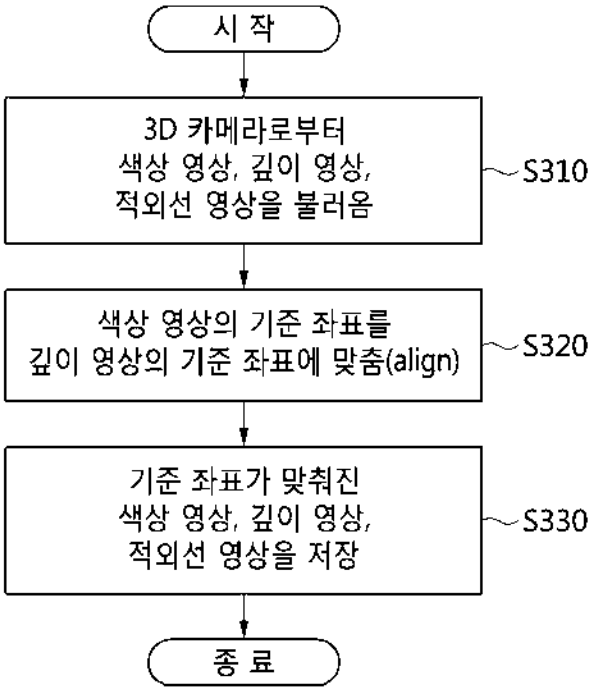
도면1


105
3D 카메라 영상
110
영상 수집부
120
후면 검출부
130
얼굴 검출부
100
140
피부 제거부
150
의류 제거부
155
검출된 머리카락 영역

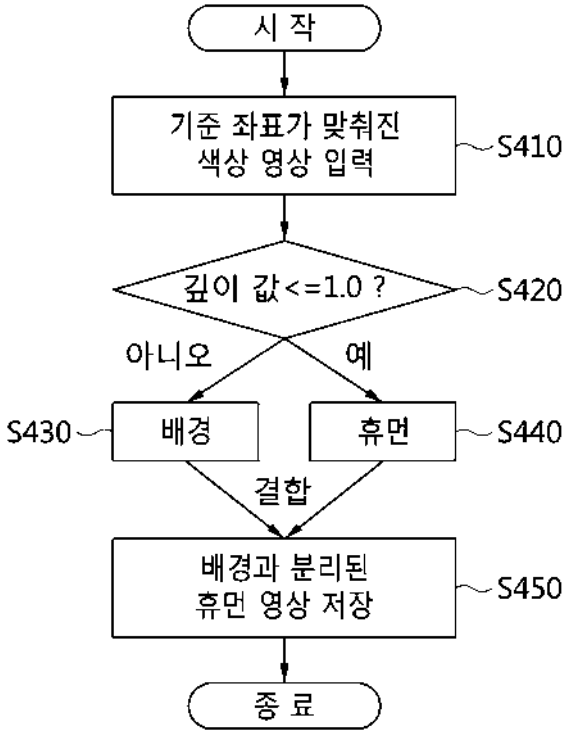
도면2


시 작
영상 입력
S210
휴먼 검출
S220
얼굴 검출
S230
피부 검출 및 제거
S240
의류 검출 및 제거
S250
종 료

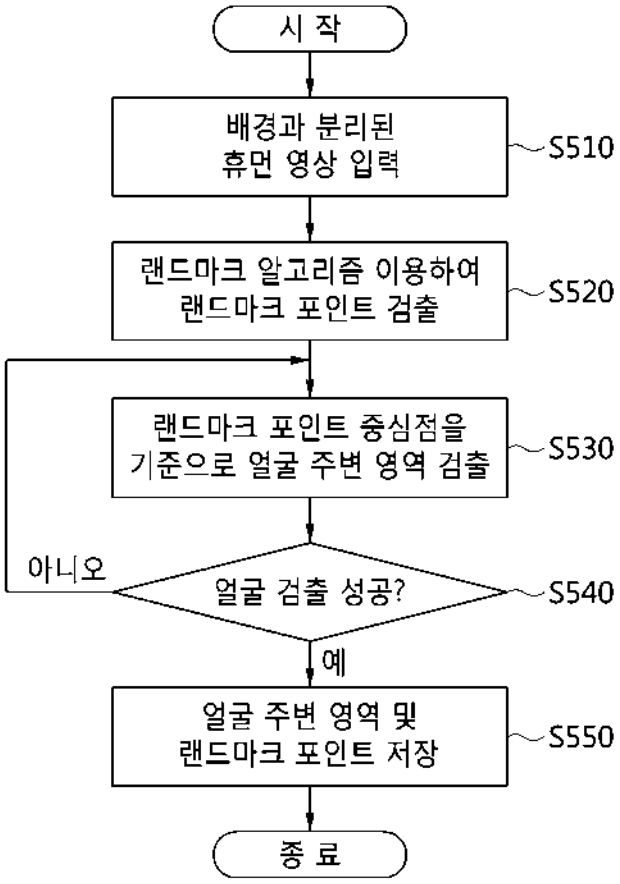
도면3


시 작
3D 카메라로부터
색상 영상, 깊이 영상,
적외선 영상을 불러옴
S310
색상 영상의 기준 좌표를
깊이 영상의 기준 좌표에 맞춤(align)
S320
기준 좌표가 맞춰진
색상 영상, 깊이 영상,
적외선 영상을 저장
S330
종 료

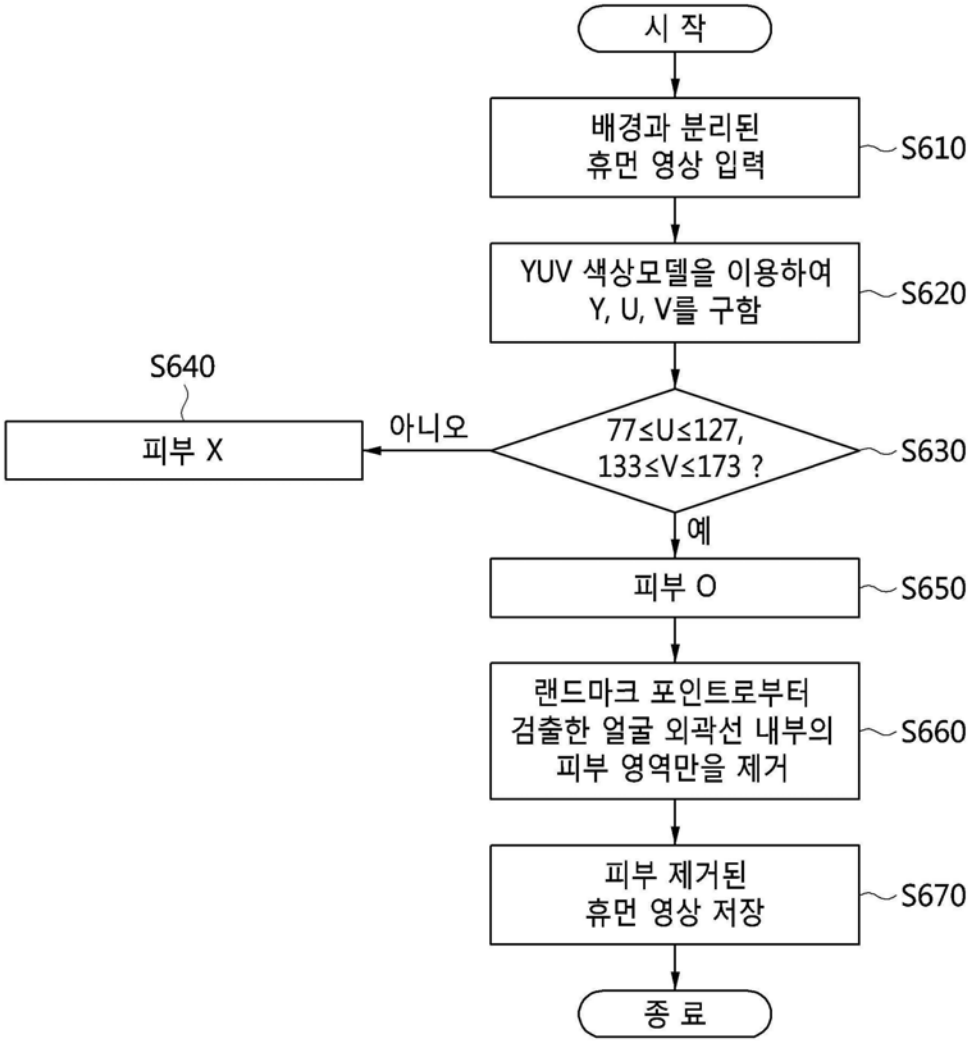
도면4


시 작
기준 좌표가 맞춰진
색상 영상 입력
S410
깊이 값<=1.0 ?
S420
아니오
예
S430
배경
휴먼
S440
결합
배경과 분리된
휴먼 영상 저장
S450
종 료

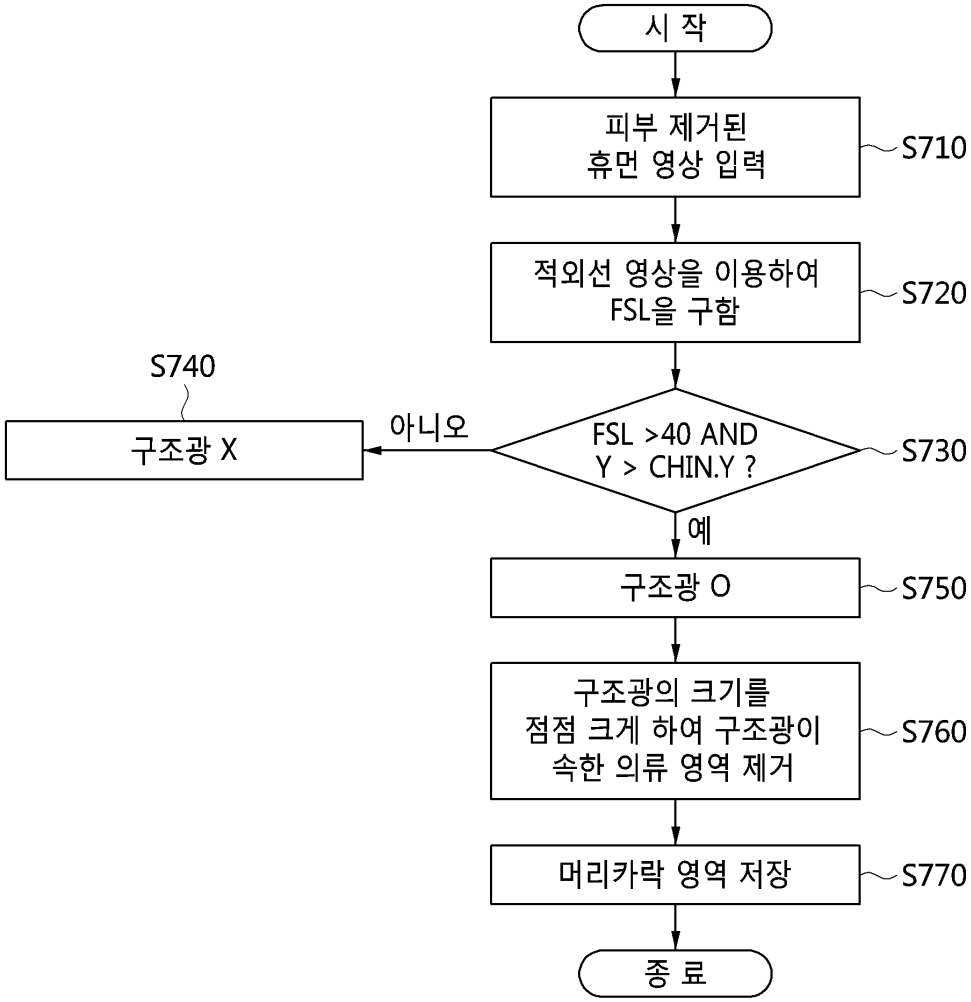
도면5


시 작
배경과 분리된
휴먼 영상 입력
S510
랜드마크 알고리즘 이용하여
랜드마크 포인트 검출
S520
랜드마크 포인트 중심점을
기준으로 얼굴 주변 영역 검출
S530
아니오
얼굴 검출 성공?
S540
예
얼굴 주변 영역 및
랜드마크 포인트 저장
S550
종 료

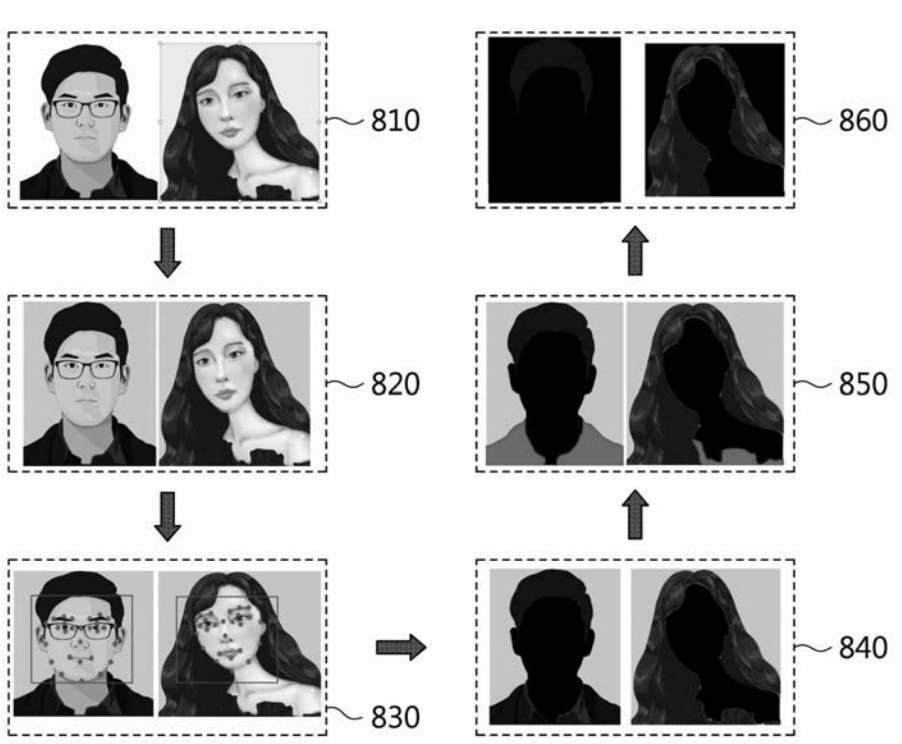
도면6


시 작
배경과 분리된
휴먼 영상 입력
S610
YUV 색상모델을 이용하여
Y, U, V를 구함
S620
77≤U≤127,
133≤V≤173 ?
S630
아니오
S640
피부 X
예
피부 O
S650
랜드마크 포인트로부터
검출한 얼굴 외곽선 내부의
피부 영역만을 제거
S660
피부 제거된
휴먼 영상 저장
S670
종 료

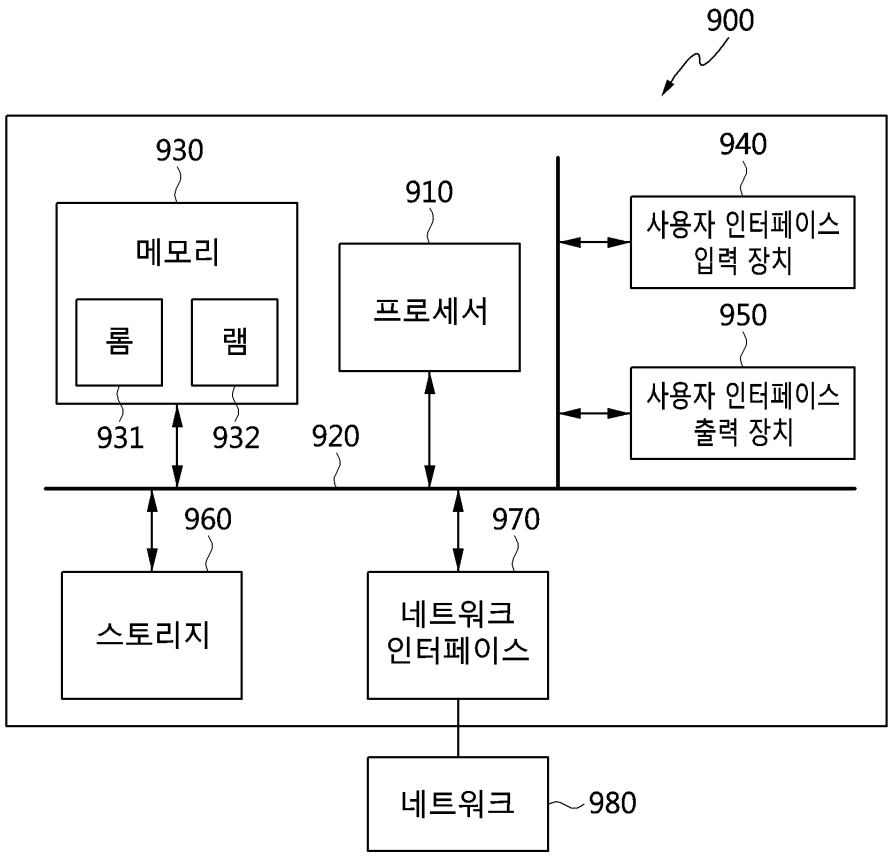
도면7


시 작
피부 제거된
휴먼 영상 입력
S710
적외선 영상을 이용하여
FSL을 구함
S720
FSL >40 AND
Y > CHIN.Y ?
S730
아니오
S740
구조광 X
예
구조광 O
S750
구조광의 크기를
점점 크게 하여 구조광이
속한 의류 영역 제거
S760
머리카락 영역 저장
S770
종 료

도면8


810
860
820
850
830
840

도면9


900
930
메모리
롬
램
931
932
910
프로세서
920
940
사용자 인터페이스
입력 장치
950
사용자 인터페이스
출력 장치
960
스토리지
970
네트워크
인터페이스
네트워크
980

대규모 딥러닝 고속처리를 위한 분산 딥러닝 플랫폼

목 차

1. 기술의 개요
2. 기술이전 내용 및 범위
3. 경쟁기술과 비교
4. 기술의 사업성
 – 활용분야 및 기대효과
5. 국내외 시장 동향

1. 기술의 개요

▣ 분산 딥러닝 플랫폼

❖ 기술 정의 : HPC(고성능컴퓨팅) 시스템 상에서 다수의 서버들을 이용하여 더 빠르게, 더 효율적으로 대규모 딥러닝 모델의 학습을 수행하는 S/W

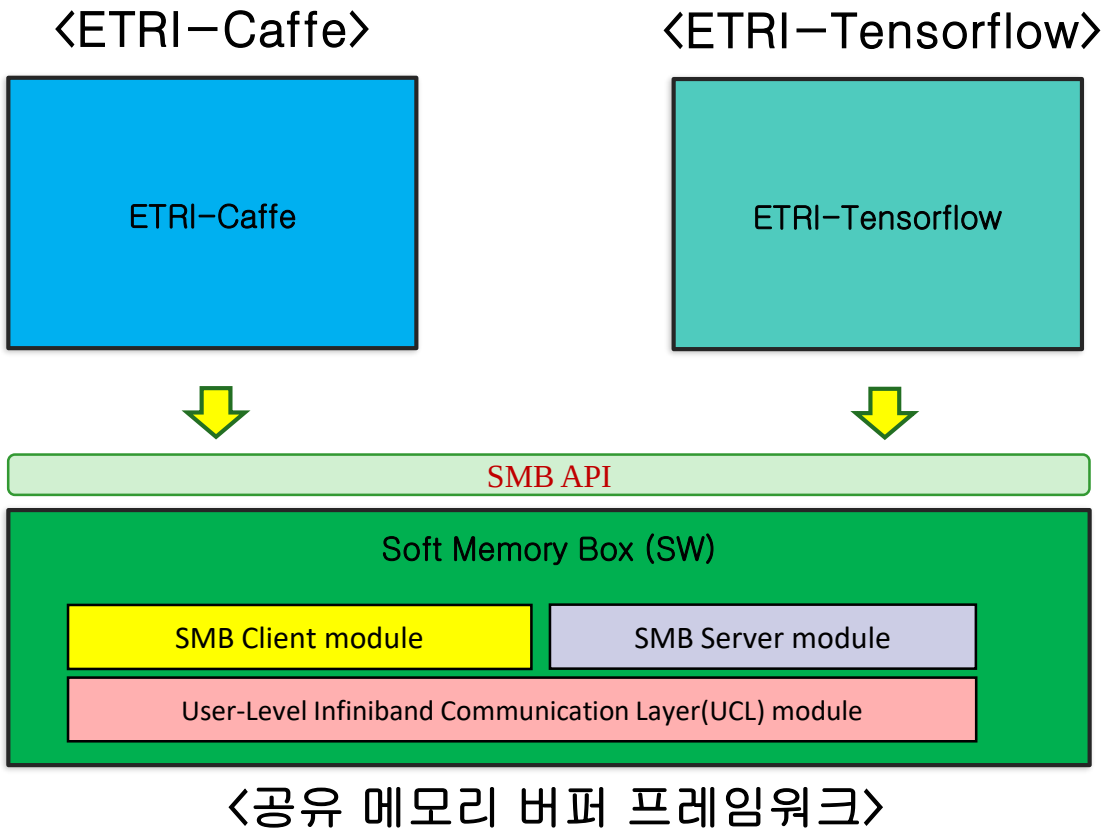
❖ 기술 분야 : 딥러닝, 분산처리, 병렬처리

❖ 기술 개발 배경

- 딥러닝 기술은 높은 정확도를 요구하는 딥러닝 모델일수록 더 많은 학습데이터와 더 높은 해상도의 학습 데이터를 요구(예, 고해상도 영상 처리 요구 증가)
- 더 높은 정확도를 가지는 모델은 기하급수적인 계산량 증가를 수반하며, HPC 시스템을 이용하여 대규모 딥러닝 모델을 분산 학습하려는 수요 증가
- 다수의 서버를 이용한 딥러닝 분산 학습은 대규모 통신이 필요하여 통신 병목이 발생함. 이를 해결하는 고속 분산 병렬 학습 기술이 필요
- 관련 용어

– HPC : High Performance Computing

▣ 기술이전 내용 및 범위



<공유 메모리 버퍼 프레임워크>

2. 기술이전 내용 및 범위

▣ 기술이전 범위

❖ [SW] 분산 딥러닝 프로그램 3종

1) 통합 공유 메모리 버퍼 프레임워크 SW 버전 1.0
2) 공유메모리 기반 분산 딥러닝 지원 에트리(ETRI) 카페(Caffe) 버전 3.0
3) 공유 메모리 기반 에트리-텐서플로우(ETRI-TensorFlow) 2.5

❖ [문서] 대시보드 설계문서 8종

1) 분산 딥러닝 플랫폼 요구사항정의서
2) 딥러닝 HPC 분산 딥러닝 플랫폼 상세설계서
3) 딥러닝 HPC 분산 딥러닝 플랫폼 시험절차서
4) 딥러닝 HPC 분산 딥러닝 플랫폼 시험결과서
5) 소프트 메모리 박스 사용자 매뉴얼
6) ShmCaffe 분산처리 확장성 분석
7) ETRI-Caffe 사용자 매뉴얼
8) ETRI-Tensorflow 사용자 매뉴얼

❖개술 개발 현황

기술성숙도(TRL : Technology Readiness Level) **단계 : (6)단계**

2. 기술이전 내용 및 범위

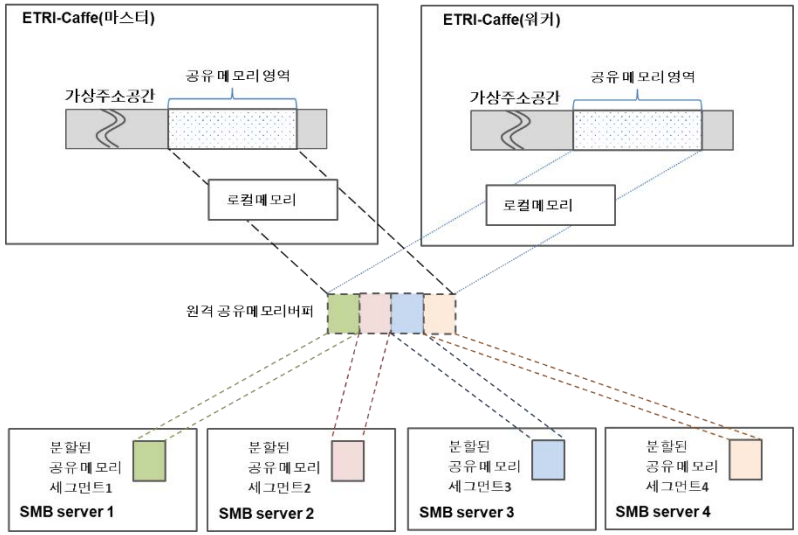
▣ 소프트 메모리 박스 기술 현황

❖ 통합 공유 메모리 버퍼 프레임워크

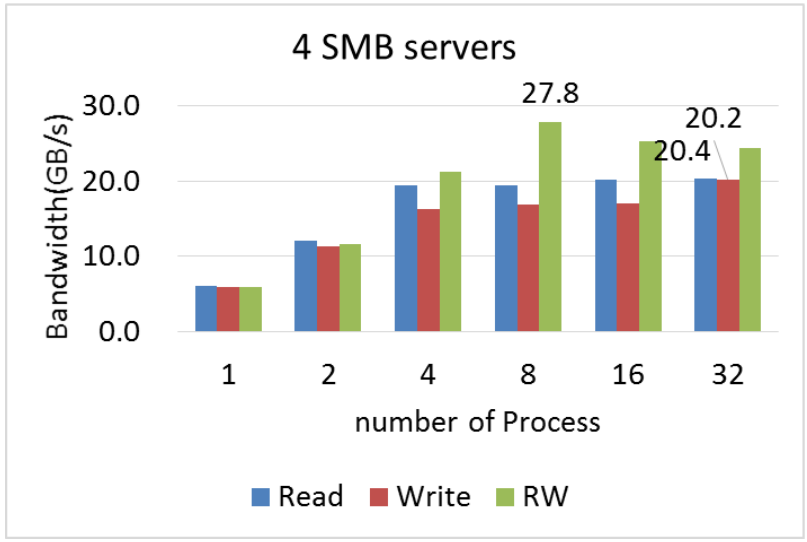
- 다중 서버 공유 메모리 통합 기술
- 원격 메모리 직접 읽기/쓰기
- 분산 공유메모리 할당/해제/접근/잠금
- 정적/공유 라이브러리 제공

❖ 고속 딥러닝 파라미터 통신 제공

- Throughput: 7GB/1SMB server
  → 27.8GB/4SMB servers
- 우수한 확장 효율: 99%/4SMB servers



<통합 분산 공유 메모리 프레임워크 구조>



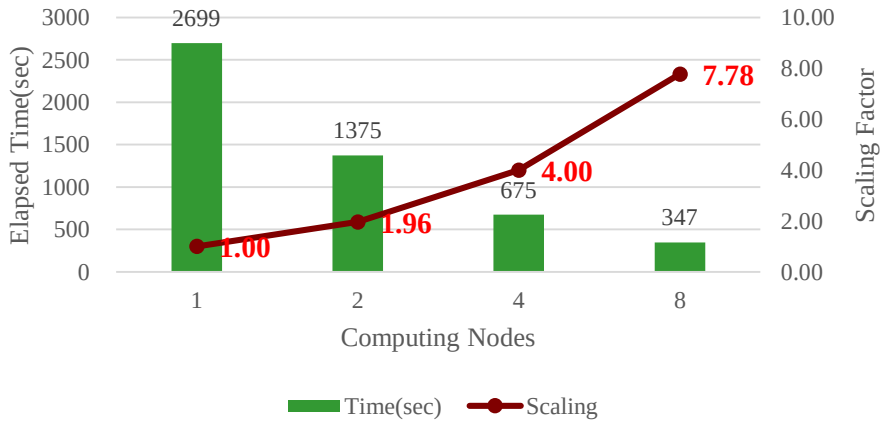
< SMB R/W Throughput>

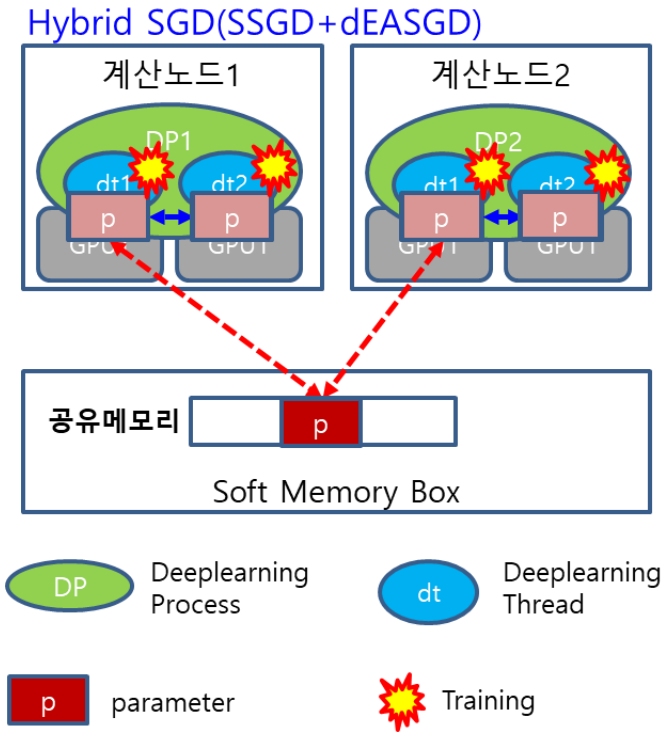
2. 기술이전 내용 및 범위

▣ ETRI-Caffe 기술 현황

- BVLC/Nvidia Caffe 모델 호환
- MPI 기반 분산 딥러닝 실행 관리
- 비동기식/하이브리드 파라미터 업데이트
- 고속 딥러닝 분산 트레이닝
- 우수한 노드 확장성(97% 확장 효율, 1→8 node)

<이미지 인식 분산 처리 확장성>





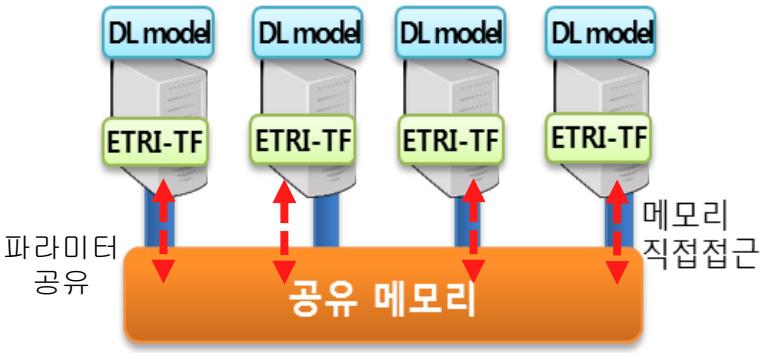
<하이브리드 분산 병렬 학습 방식>

BVLC : Berkeley Vision and Learning Center, MPI: Message Passing Interface

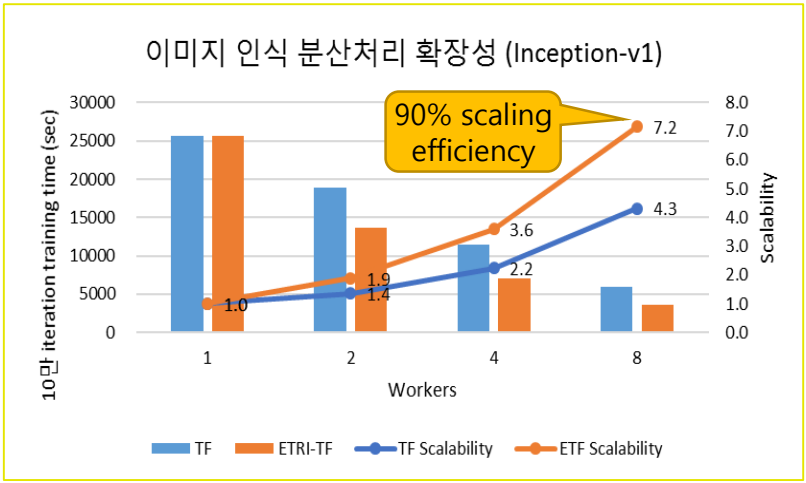
2. 기술이전 내용 및 범위

▣ ETRI-Tensorflow 기술 개발 현황

- TensorFlow 모델 호환
- CNN, RNN 및 그 외 DNN 모델 지원
- 데이터 병렬 및 모델 병렬 분산 트레이닝 지원
- 비동기식 데이터/모델 병렬 트레이닝
- 공유메모리 기반 분산 트레이닝 가속
- TensorFlow 대비 2배 빠른 학습
- 다수 이미지 인식, 음성인식 모델로 성능 검증
- 우수한 확장 성능 제공 (90% 확장 효율)



<ETRI-TensorFlow 기술 개념도>

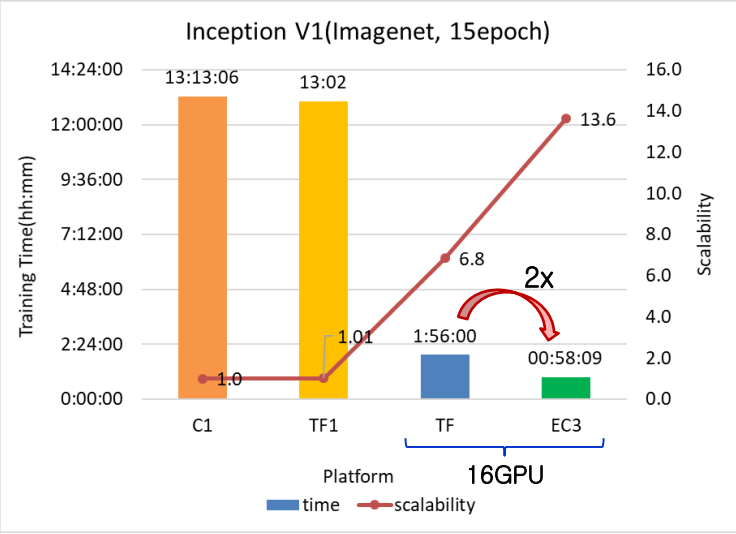


3. 경쟁기술과 비교

▣ 경쟁 분산 딥러닝 기술과의 성능 비교

❖ ETRI-Caffe

- NVIDIA Caffe(1GPU)대비 13.6배(16GPU)
- TF 대비 최대 2배 빠른 학습



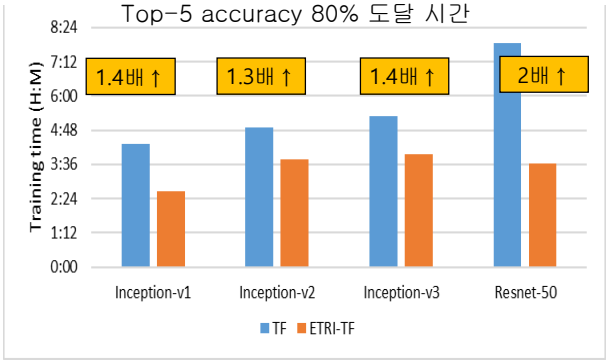
C1: Caffe(1GPU), TF1: Tensorflow(1GPU)
TF: Tensorflow(v1.13) EC3: ETRI-Caffe(v3.0)

< ETRI-Caffe 성능 비교 >

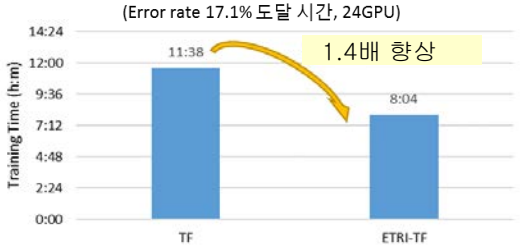
❖ ETRI-Tensorflow

- 이미지 인식 : Tensorflow 대비 최대 2배
- 음성인식 트레이닝: 최대 1.4배

<이미지 인식 모델 트레이닝 성능>



<음성인식 모델 트레이닝 성능>



< ETRI-Tensorflow 성능 비교>

4. 기술의 사업성

▣ 사업 가능 영역

- ❖ (B2B) 기업용 온프레미스 분산 딥러닝 개발 환경/인프라 구축
- ❖ (B2B/B2C) 클라우드기반 분산 딥러닝 개발 환경 서비스

(수요처)
병원, 중견기업,
공공기관,
대학교 연구실,

분산 딥러닝 플랫폼 기술

(B2B) 기업용 온프레미스
분산 딥러닝 개발환경 구축

(수요처)
클라우드
서비스 기업

분산 딥러닝 플랫폼 기술

(B2B/B2C) 클라우드기반
분산 딥러닝 개발 환경

4. 기술의 사업성

▣ 기술이전 방식

❖ 정액기술료(세부기술별 이전 가능)

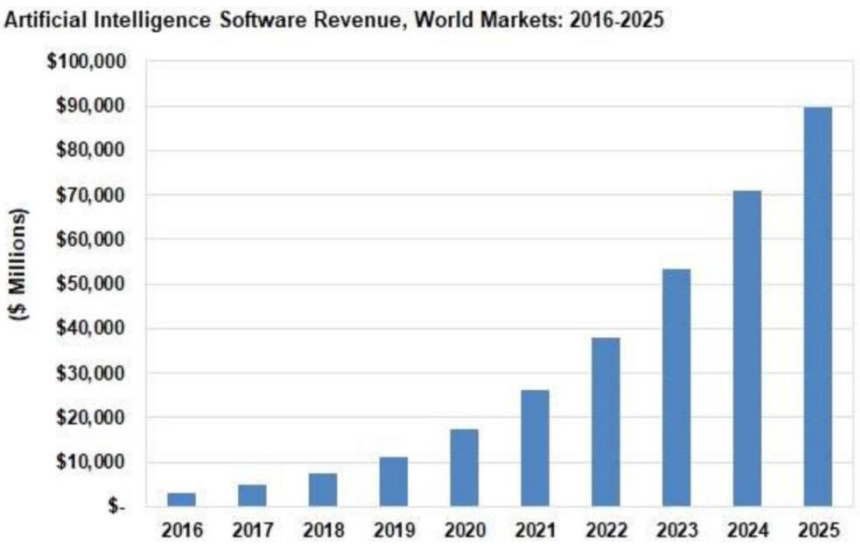
구분		착수기본료(원)		
		중소기업	중견기업	대기업
A. Soft Memory Box	특허2건실시권 소스프로그램1건 기술문서5건	50,000,000	150,000,000	200,000,000
B. ETRI-Caffe	특허1건실시권 소스프로그램1건 기술문서2건	13,000,000	39,000,000	52,000,000
C. ETRI-Tensorflow	특허1건실시권 소스프로그램1건 기술문서1건	20,000,000	60,000,000	80,000,000
합계	특허4건, 프로그램3건, 기술문서 8건	83,000,000	249,000,000	332,000,000

- A기술은 B,C 기술을 활용하기 위한 필수 기술임

5. 국내외 시장 동향

▣ 국외 AI HW/SW 시장 예측 (2017-2025)

글로벌 AI시장 규모 및 전망



Source: Tractica, 2017

인공지능 시장 전망

- 2017년 구글 연례 개발자 회의에서 구글 CEO 선다 피차이는 '약 10년 주기로 PC에서 웹, 스마트폰으로 컴퓨터 주류 형태가 변해왔는데, 이제 모바일 퍼스트에서 AI퍼스트로 옮겨가고 있다'고 밝힘. 엄청난 속도로 쌓이는 데이터를 가치 있는 비즈니스로 만들어주는 도구가 AI임
- 전 세계 인공지능 기반 스마트 머신 시장은 2014년 62억 2,900만 달러에서 2019년 152억 7,900만 달러 규모로 성장 전망됨 (BCC리서치)
- 영상처리 시장은 2015년 765억 달러에서 2017년 기준 1,090억 달러, 음성인식 시장은 같은 기간 840억 달러에서 1,130억 달러 수준으로 성장 예상됨

5. 국내외 시장 동향

▣ 국내 AI 산업 시장 예측 (2014-2020, 단위 조원)

AI 분야별 국내 시장규모 및 전망

Source : 과학기술정보통신부

영상 분석 산업 전망

- 2020년 영상처리 시스템 세계 시장규모는 약 176억 달러로 전망되며, 2013년~2015년 연 성장률 8.80%를 보였음 (ETRI 기술경제연구본부, 2016)
- 국내 영상처리 시스템 시장 규모 역시 2013~2015년 연 4% 이상의 성장률로 2020년 1,900억 원을 돌파할 것으로 예상됨
- 카메라 기술의 빠른 발전으로 영상 및 이미지 해상도가 높아지는 가운데, 이를 빠르게 분석, 처리할 수 있는 GPU의 등장으로 영상 및 이미지 분석 시장 성장이 가속될 것으로 보임

5. 국내외 시장 동향

▣ 예상 제품/서비스의 예상매출액(생산/판매부터 향후 매 5년 간 추정)

(단위: M$(국외), 십억원(국내))

관련제품/서비스	시장	2020	2021	2022	2023	2024	합계
기업용 딥러닝 개발 환경 (on-premise용 플랫폼)	국외	0.00	132.00	420.00	846.00	1551.00	2949.00
	국내	12.24	6.66	10.99	20.15	35.46	85.50
클라우드 기반 딥러닝 개발 환경 (클라우드용 플랫폼)	국외	0.00	44.00	140.00	282.00	413.60	879.60
	국내	0.41	1.06	2.52	5.64	9.93	19.55
합계	국외	0.00	176.00	560.00	1128.00	1964.60	3828.60
	국내	12.65	7.72	13.51	25.79	45.39	105.05


감사합니다.

(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년08월25일
(11) 등록번호 10-2292389
(24) 등록일자 2021년08월17일

(51) 국제특허분류(Int. Cl.)
G06F 12/084 (2016.01) *G06F 12/02* (2018.01)
G06F 13/28 (2006.01)
(52) CPC특허분류
G06F 12/084 (2013.01)
G06F 12/0292 (2013.01)
(21) 출원번호 10-2018-0006024
(22) 출원일자 2018년01월17일
심사청구일자 2019년03월14일
(65) 공개번호 10-2019-0087783
(43) 공개일자 2019년07월25일
(56) 선행기술조사문헌
KR1020160033505 A*
KR1020130079865 A*
JP2013513839 A*
KR101533405 B1*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
안신영
대전광역시 서구 둔산북로 160, 5동 701호
임은지
대전광역시 유성구 노은동로 187, 602동 1801호
(뒷면에 계속)
(74) 대리인
한양특허법인

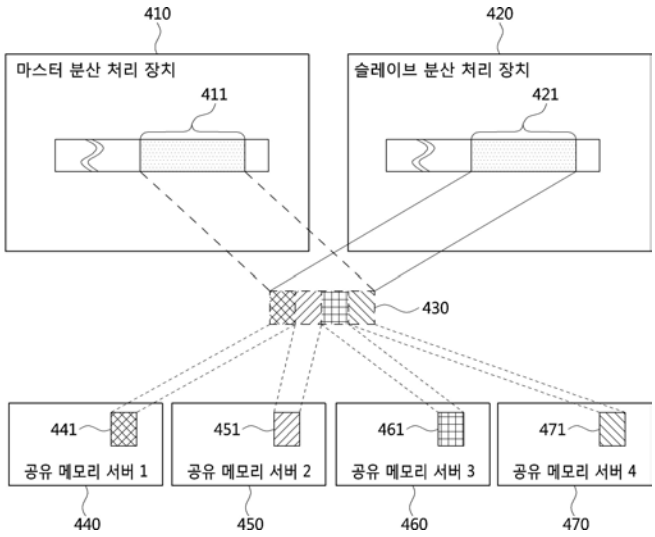
전체 청구항 수 : 총 19 항 심사관 : 김계준

(54) 발명의 명칭 **원격 직접 메모리 접근을 통한 분산 처리 장치 및 그 방법**

(57) 요 약

본 발명의 일 실시예는, 공유 메모리 서버 클러스터를 구성하는 복수의 공유 메모리 서버들에 구비된 원격 메모리들에 직접 접근하여 데이터를 송수신하는 통신부; 상기 원격 메모리들로부터 할당된 공유 메모리 버퍼 세그먼트들로 구성된 공유 메모리 버퍼에 대하여, 메모리에 상기 공유 메모리 버퍼와 동일한 크기만큼 로컬 공유 메모리 영역을 할당하고, 상기 공유 메모리 버퍼와 상기 로컬 공유 메모리 영역을 동기화하는 공유 메모리 접근 관리부; 상기 공유 메모리 버퍼와 상기 로컬 공유 메모리 영역 사이의 메모리 맵핑 테이블을 관리하는 메모리 맵핑 테이블 관리부; 및 상기 로컬 공유 메모리 영역에 대한 주어진 연산을 수행하는 연산부를 포함하는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 장치를 제공한다.

대 표 도



(52) CPC특허분류
G06F 13/28 (2013.01)
(72) 발명자
최용석
전광역시 유성구 지족북로 60, 207동 303호
우영춘
대전광역시 유성구 어은로 57, 113동 404호
최완
대전광역시 서구 관저북로 52, 108동 306호

이 발명을 지원한 국가연구개발사업

과제고유번호	2016-0-00087
부처명	미래창조과학부
과제관리(전문)기관명	정보통신기술진흥센터(IITP)
연구사업명	정보통신방송기술개발사업(SW컴퓨팅산업원천기술개발사업)
연구과제명	대규모 딥러닝 고속 처리를 위한 HPC 시스템 개발
기 여 율	1/1
과제수행기관명	한국전자통신연구원
연구기간	2017.01.01 ~ 2017.12.31

명 세 서

청구범위

청구항 1

공유 메모리 서버 클러스터를 구성하는 복수의 공유 메모리 서버들에 구비된 원격 메모리들에 직접 접근하여 데이터를 송수신하는 통신부;

상기 원격 메모리들로부터 할당된 공유 메모리 버퍼 세그먼트들로 구성된 공유 메모리 버퍼에 대하여, 로컬 메모리에 상기 공유 메모리 버퍼와 동일한 크기만큼 로컬 공유 메모리 영역을 할당하고, 상기 공유 메모리 버퍼와 상기 로컬 공유 메모리 영역을 동기화하는 공유 메모리 접근 관리부;

상기 공유 메모리 버퍼와 상기 로컬 공유 메모리 영역 사이의 메모리 맵핑 테이블을 관리하는 메모리 맵핑 테이블 관리부; 및

상기 로컬 공유 메모리 영역에 대한 주어진 연산을 수행하는 연산부를 포함하고,

상기 공유 메모리 버퍼 세그먼트들은 서로 다른 상기 공유 메모리 서버들에 의해 할당되고,

상기 공유 메모리 버퍼는 복수의 분산 처리 장치들에 의하여 공유되는 가상의 연속된 버퍼에 상응하고,

상기 로컬 공유 메모리 영역은 가상주소만을 반환하는 것이 아니고 실제 물리 메모리에 할당되는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 장치.

청구항 2

청구항 1에 있어서,

상기 공유 메모리 접근 관리부는

직접 상기 공유 메모리 버퍼를 생성하여 동일한 분산 처리 프레임워크를 구성하는 다른 분산 처리 장치들에 공유 메모리 버퍼 정보를 공유하거나, 상기 다른 분산 처리 장치에 의하여 생성된 상기 공유 메모리 버퍼에 상응하는 공유 메모리 버퍼 정보를 수신하는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 장치.

청구항 3

청구항 2에 있어서,

상기 공유 메모리 접근 관리부는

상기 공유 메모리 서버들에 각각에 상응하는 상기 공유 메모리 버퍼 세그먼트들의 크기를 계산하고, 상기 공유 메모리 서버들에 상기 공유 메모리 버퍼 세그먼트들의 생성 및 할당을 요청하여 상기 공유 메모리 버퍼를 생성하는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 장치.

청구항 4

청구항 2에 있어서,

상기 공유 메모리 접근 관리부는

상기 연산에 의하여 상기 로컬 공유 메모리 영역의 데이터가 변경된 경우에 상기 로컬 공유 메모리 영역의 데이터를 상기 공유 메모리 버퍼에 복사하여 상기 원격 메모리들과 데이터를 동기화하고, 상기 다른 분산 처리 장치들에 의하여 상기 공유 메모리 버퍼의 데이터가 변경된 경우에 상기 공유 메모리 버퍼의 데이터를 상기 로컬 공유 메모리 영역으로 복사하여 상기 원격 메모리들과 데이터를 동기화하는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 장치.

청구항 5

청구항 2에 있어서,

상기 공유 메모리 접근 관리부는

두 개 이상의 공유 메모리 버퍼들 사이의 데이터 누적 연산을 수행하는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 장치.

청구항 6

청구항 5에 있어서,

상기 공유 메모리 접근 관리부는

상기 공유 메모리 서버들에 누적 연산을 요청하고, 상기 공유 메모리 서버들로부터 상기 공유 메모리 버퍼 세그먼트들에 대하여 누적 연산을 수행한 결과를 수신하여 상기 데이터 누적 연산을 수행하는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 장치.

청구항 7

청구항 2에 있어서,

상기 공유 메모리 접근 관리부는

상기 공유 메모리 서버들에 상기 공유 메모리 버퍼 세그먼트들의 해제 및 삭제를 요청하고, 상기 공유 메모리 버퍼 세그먼트들의 해제 및 삭제 요청의 결과들을 수신함에 따라 상기 로컬 공유 메모리 영역을 해제 및 삭제하여 상기 공유 메모리 버퍼의 사용을 종료하고,

상기 메모리 맵핑 테이블 관리부는

상기 공유 메모리 버퍼의 사용이 종료되면 상기 메모리 맵핑 테이블을 삭제하는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 장치.

청구항 8

원격 직접 메모리 접근을 통하여 복수의 분산 처리 장치들과 데이터를 송수신하는 통신부;

상기 분산 처리 장치들이 직접 접근할 수 있는 메모리;

동일한 공유 메모리 서버 클러스터를 구성하는 다른 공유 메모리 서버들과 공유 메모리 버퍼를 구성하는 공유 메모리 관리부를 포함하고,

상기 공유 메모리 버퍼는 상기 공유 메모리 서버들로부터 할당된 공유 메모리 버퍼 세그먼트들로 구성되고,

상기 공유 메모리 버퍼는 상기 분산 처리 장치들 각각에 대하여 상기 공유 메모리 버퍼와 동일한 크기로 할당된 로컬 공유 메모리 영역과 메모리 맵핑 테이블을 이용하여 동기화되고,

상기 공유 메모리 버퍼는 복수의 분산 처리 장치들에 의하여 공유되는 가상의 연속된 버퍼에 상응하고,

상기 로컬 공유 메모리 영역은 가상주소만을 반환하는 것이 아니고 실제 물리 메모리에 할당되는 것을 특징으로 하는, 공유 메모리 서버.

청구항 9

삭제

청구항 10

청구항 8에 있어서,

상기 공유 메모리 관리부는

상기 분산 처리 장치로부터 상기 공유 메모리 버퍼를 생성하기 위한 공유 메모리 버퍼 세그먼트의 크기 정보와 함께 상기 공유 메모리 버퍼 세그먼트의 생성 및 할당을 요청을 수신하고, 상기 공유 메모리 버퍼 세그먼트를 생성 및 할당하여 상기 공유 메모리 버퍼를 구성하는 것을 특징으로 하는, 공유 메모리 서버.

청구항 11

청구항 8에 있어서,

상기 공유 메모리 버퍼는

연산에 의하여 특정 로컬 공유 메모리 영역의 데이터가 변경된 경우에 상기 변경된 로컬 공유 메모리 영역의 데이터와 동기화되고, 변경된 데이터로 나머지 로컬 공유 메모리 영역들과 동기화되는 것을 특징으로 하는, 공유 메모리 서버.

청구항 12

청구항 8에 있어서,

상기 공유 메모리 관리부는

상기 분산 처리 장치로부터 두 개 이상의 공유 메모리 버퍼들 사이의 데이터 누적 연산 요청을 수신하고, 상기 데이터 누적 연산의 대상이 되는 공유 메모리 버퍼 세그먼트들에 대하여 누적 연산을 수행하고, 결과를 상기 분산 처리 장치에 반환하는 것을 특징으로 하는, 공유 메모리 서버.

청구항 13

청구항 8에 있어서,

상기 공유 메모리 관리부는

상기 분산 처리 장치가 상기 공유 메모리 버퍼의 사용을 종료하기 위하여 전송한 상기 공유 메모리 버퍼 세그먼트의 해제 및 삭제 요청을 수신하여 상기 공유 메모리 버퍼 세그먼트를 해제 및 삭제하고, 결과를 상기 분산 처리 장치에 반환하여 상기 분산 처리 장치가 상기 로컬 공유 메모리 영역을 해제 및 삭제하고 상기 메모리 맵핑 테이블을 삭제하도록 하는 것을 특징으로 하는, 공유 메모리 서버.

청구항 14

공유 메모리 서버 클러스터를 구성하는 복수의 공유 메모리 서버들에 구비된 원격 메모리들로부터 할당된 공유 메모리 버퍼 세그먼트들로 구성된 공유 메모리 버퍼에 대하여, 로컬 메모리에 상기 공유 메모리 버퍼와 동일한 크기만큼 로컬 공유 메모리 영역을 할당하는 단계;

상기 공유 메모리 버퍼와 상기 로컬 공유 메모리 영역 사이의 메모리 맵핑 테이블을 관리하는 단계;

상기 원격 메모리들에 직접 접근하여 데이터를 송수신하여 상기 공유 메모리 버퍼와 상기 로컬 공유 메모리 영역을 동기화하는 단계; 및

상기 로컬 공유 메모리 영역에 대한 주어진 연산을 수행하는 단계를 포함하고,

상기 공유 메모리 버퍼 세그먼트들은 서로 다른 상기 공유 메모리 서버들에 의해 할당되고,

상기 공유 메모리 버퍼는 복수의 분산 처리 장치들에 의하여 공유되는 가상의 연속된 버퍼에 상응하고,

상기 로컬 공유 메모리 영역은 가상주소만을 반환하는 것이 아니고 실제 물리 메모리에 할당되는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 방법.

청구항 15

청구항 14에 있어서,

직접 상기 공유 메모리 버퍼를 생성하여 동일한 분산 처리 프레임워크를 구성하는 다른 분산 처리 장치들에 공유 메모리 버퍼 정보를 공유하거나, 상기 다른 분산 처리 장치에 의하여 생성된 상기 공유 메모리 버퍼에 상응하는 공유 메모리 버퍼 정보를 수신하여 상기 공유 메모리 버퍼 정보를 획득하는 단계

를 더 포함하는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 방법.

청구항 16

청구항 15에 있어서,

상기 공유 메모리 버퍼 정보를 획득하는 단계는

상기 공유 메모리 서버들에 각각에 상응하는 상기 공유 메모리 버퍼 세그먼트들의 크기를 계산하는 단계; 및

상기 공유 메모리 서버들에 상기 공유 메모리 버퍼 세그먼트들의 생성 및 할당을 요청하여 상기 공유 메모리 버퍼를 생성하는 단계

를 포함하는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 방법.

청구항 17

청구항 15에 있어서,

상기 동기화하는 단계는

상기 연산에 의하여 상기 로컬 공유 메모리 영역의 데이터가 변경된 경우에 상기 로컬 공유 메모리 영역의 데이터를 상기 공유 메모리 버퍼에 복사하여 상기 원격 메모리들과 데이터를 동기화하고, 상기 다른 분산 처리 장치들에 의하여 상기 공유 메모리 버퍼의 데이터가 변경된 경우에 상기 공유 메모리 버퍼의 데이터를 상기 로컬 공유 메모리 영역으로 복사하여 상기 원격 메모리들과 데이터를 동기화하는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 방법.

청구항 18

청구항 15에 있어서,

두 개 이상의 공유 메모리 버퍼들 사이의 데이터 누적 연산을 수행하는 단계

를 더 포함하는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 방법.

청구항 19

청구항 18에 있어서,

상기 데이터 누적 연산을 수행하는 단계는

상기 공유 메모리 서버들에 누적 연산을 요청하는 단계; 및

상기 공유 메모리 서버들로부터 상기 공유 메모리 버퍼 세그먼트들에 대하여 누적 연산을 수행한 결과를 수신하는 단계

를 포함하는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 방법.

청구항 20

청구항 15에 있어서,

상기 공유 메모리 서버들에 상기 공유 메모리 버퍼 세그먼트들의 해제 및 삭제를 요청하고, 상기 공유 메모리 버퍼 세그먼트들의 해제 및 삭제 요청의 결과들을 수신함에 따라 상기 로컬 공유 메모리 영역을 해제 및 삭제하고, 상기 메모리 맵핑 테이블을 삭제하여 상기 공유 메모리 버퍼의 사용을 종료하는 단계

를 더 포함하는 것을 특징으로 하는 하는, 원격 직접 메모리 접근을 통한 분산 처리 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 원격 직접 메모리 접근을 통한 분산 처리 장치 및 그 방법에 관한 것으로, 구체적으로 분산 처리 프레임워크에서 공유 메모리 서버들의 메모리들로 구성된 가상의 공유 메모리 버퍼에 직접 접근하는 분산 처리 장치 및 그 방법에 관한 것이다.

배 경 기 술

[0002] 분산 병렬 처리란 다수의 계산 자원을 동시에 병렬로 사용하여 대규모 데이터 분석을 빠르게 수행하는 것이다.

다수의 계산 노드에 분산 병렬 실행되는 프로세스들은 상호간에 데이터 공유가 필수적이며 대표적인 데이터 공유 방식으로 MPI(Message Passing Interface)를 들 수 있다. 그러나 분산 처리 데이터의 일부만을 일부 프로세스 간에 메시지 패싱 형태로 전달하는 형태가 아니라 지속적으로 전체 분산 처리 프로세스 간에 전체 처리 데이터를 비동기적으로 업데이트하고 참조하는 경우에는 MPI 방식보다는 공유 메모리 형태로 공유하는 것이 더 유리하다.

[0003] 분산 처리 플랫폼에서 분산 처리를 수행하는 프로세스들은 상호 간에 대규모 공유 데이터를 빈번하게 송수신해야 하며, 이에 따른 통신 오버헤드는 전체 분산 처리 성능이나 처리 시간에서 차지하는 비중이 매우 높다. 통신 오버헤드가 높을수록 계산 노드의 계산 프로세서(예컨대, CPU, GPU 등)들은 대기하는 시간이 길어지고 이는 자원 사용률 저하로 나타난다. 통신 오버헤드가 높은 이유는 TCP/IP를 포함한 대부분의 통신 프로토콜 스택은 응용 프로세스가 보내는 메시지를 다단계의 프로토콜 레이어를 통해 처리하는 프로토콜 처리 오버헤드와 프로토콜 처리중에 1회 이상 메모리 복사가 발생 때문이다. 따라서, 원격 직접 메모리 접근(RDMA: Remote Direct Memory Access)을 통하여 분산 처리에 따른 통신 오버헤드를 줄이는 것이 요구된다.

[0004] 전술한 배경기술은 발명자가 본 발명의 도출을 위해 보유하고 있었거나, 본 발명의 도출 과정에서 습득한 기술정보로서, 반드시 본 발명의 출원 전에 일반 공중에게 공개된 공지기술이라할 수는 없다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 국내 공개특허공보 제10-2006-0009244호

발명의 내용

해결하려는 과제

[0006] 본 발명의 목적은 원격 직접 메모리 접근을 통하여 분산 처리 데이터를 공유하는 원격 직접 메모리 접근을 통한 분산 처리 장치 및 그 방법을 제공하는 것이다.

[0007] 또한, 본 발명의 목적은 다수의 공유 메모리 서버들을 클러스터링하고 각각의 공유 메모리 서버들로부터 공유 메모리 버퍼 세그먼트들을 할당하여 공유 메모리 버퍼를 구성하고, 공유 메모리 버퍼에 원격 직접 메모리 접근하는 분산 처리 장치 및 그 방법을 제공하는 것이다.

과제의 해결 수단

[0008] 본 발명의 일 실시예는, 공유 메모리 서버 클러스터를 구성하는 복수의 공유 메모리 서버들에 구비된 원격 메모리들에 직접 접근하여 데이터를 송수신하는 통신부; 상기 원격 메모리들로부터 할당된 공유 메모리 버퍼 세그먼트들로 구성된 공유 메모리 버퍼에 대하여, 메모리에 상기 공유 메모리 버퍼와 동일한 크기만큼 로컬 공유 메모리 영역을 할당하고, 상기 공유 메모리 버퍼와 상기 로컬 공유 메모리 영역을 동기화하는 공유 메모리 접근 관리부; 상기 공유 메모리 버퍼와 상기 로컬 공유 메모리 영역 사이의 메모리 맵핑 테이블을 관리하는 메모리 맵핑 테이블 관리부; 및 상기 로컬 공유 메모리 영역에 대한 주어진 연산을 수행하는 연산부를 포함하는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 장치를 제공한다.

[0009] 이때, 상기 공유 메모리 접근 관리부는 직접 상기 공유 메모리 버퍼를 생성하여 동일한 분산 처리 프레임워크를 구성하는 다른 분산 처리 장치들에 공유 메모리 버퍼 정보를 공유하거나, 상기 다른 분산 처리 장치에 의하여 생성된 상기 공유 메모리 버퍼에 상응하는 공유 메모리 버퍼 정보를 수신할 수 있다.

[0010] 이때, 상기 공유 메모리 접근 관리부는 상기 공유 메모리 서버들에 각각에 상응하는 상기 공유 메모리 버퍼 세그먼트들의 크기를 계산하고, 상기 공유 메모리 서버들에 상기 공유 메모리 버퍼 세그먼트들의 생성 및 할당을 요청하여 상기 공유 메모리 버퍼를 생성할 수 있다.

[0011] 이때, 상기 공유 메모리 접근 관리부는 상기 연산에 의하여 상기 로컬 공유 메모리 영역의 데이터가 변경된 경우에 상기 로컬 공유 메모리 영역의 데이터를 상기 공유 메모리 버퍼에 복사하여 상기 원격 메모리들과 데이터를 동기화하고, 상기 다른 분산 처리 장치들에 의하여 상기 공유 메모리 버퍼의 데이터가 변경된 경우에 상기 공유 메모리 버퍼의 데이터를 상기 로컬 공유 메모리 영역으로 복사하여 상기 원격 메모리들과 데이터를 동기화

할 수 있다.

[0012] 이때, 상기 공유 메모리 접근 관리부는 두 개 이상의 공유 메모리 버퍼들 사이의 데이터 누적 연산을 수행할 수 있다.

[0013] 이때, 상기 공유 메모리 접근 관리부는 상기 공유 메모리 서버들에 누적 연산을 요청하고, 상기 공유 메모리 서버들로부터 상기 공유 메모리 버퍼 세그먼트들에 대하여 누적 연산을 수행한 결과를 수신하여 상기 데이터 누적 연산을 수행할 수 있다.

[0014] 이때, 상기 공유 메모리 접근 관리부는 상기 공유 메모리 서버들에 상기 공유 메모리 버퍼 세그먼트들의 해제 및 삭제를 요청하고, 상기 공유 메모리 버퍼 세그먼트들의 해제 및 삭제 요청의 결과들을 수신함에 따라 상기 로컬 공유 메모리 영역을 해제 및 삭제하여 상기 공유 메모리 버퍼의 사용을 종료하고, 상기 메모리 맵핑 테이블 관리부는 상기 공유 메모리 버퍼의 사용이 종료되면 상기 메모리 맵핑 테이블을 삭제할 수 있다.

[0015] 본 발명의 다른 일 실시예는, 원격 직접 메모리 접근을 통하여 복수의 분산 처리 장치들과 데이터를 송수신하는 통신부; 상기 분산 처리 장치들이 직접 접근할 수 있는 메모리; 동일한 공유 메모리 서버 클러스터를 구성하는 다른 공유 메모리 서버들과 공유 메모리 버퍼를 구성하는 공유 메모리 관리부를 포함하는 것을 특징으로 하는, 공유 메모리 서버를 제공한다.

[0016] 이때, 상기 공유 메모리 버퍼는 상기 분산 처리 장치들 각각에 대하여 상기 공유 메모리 버퍼와 동일한 크기로 할당된 로컬 공유 메모리 영역과 메모리 맵핑 테이블을 이용하여 동기화될 수 있다.

[0017] 이때, 상기 공유 메모리 관리부는 상기 분산 처리 장치로부터 상기 공유 메모리 버퍼를 생성하기 위한 공유 메모리 버퍼 세그먼트의 크기 정보와 함께 상기 공유 메모리 버퍼 세그먼트의 생성 및 할당을 요청을 수신하고, 상기 공유 메모리 버퍼 세그먼트를 생성 및 할당하여 상기 공유 메모리 버퍼를 구성할 수 있다.

[0018] 이때, 상기 공유 메모리 버퍼는 연산에 의하여 특정 로컬 공유 메모리 영역의 데이터가 변경된 경우에 상기 변경된 로컬 공유 메모리 영역의 데이터와 동기화되고, 변경된 데이터로 나머지 로컬 공유 메모리 영역들과 동기화될 수 있다.

[0019] 이때, 상기 공유 메모리 관리부는 상기 분산 처리 장치로부터 두 개 이상의 공유 메모리 버퍼들 사이의 데이터 누적 연산 요청을 수신하고, 상기 데이터 누적 연산의 대상이 되는 공유 메모리 버퍼 세그먼트들에 대하여 누적 연산을 수행하고, 결과를 상기 분산 처리 장치에 반환할 수 있다.

[0020] 이때, 상기 공유 메모리 관리부는 상기 분산 처리 장치가 상기 공유 메모리 버퍼의 사용을 종료하기 위하여 전송한 상기 공유 메모리 버퍼 세그먼트의 해제 및 삭제 요청을 수신하여 상기 공유 메모리 버퍼 세그먼트를 해제 및 삭제하고, 결과를 상기 분산 처리 장치에 반환하여 상기 분산 처리 장치가 상기 로컬 공유 메모리 영역을 해제 및 삭제하고 상기 메모리 맵핑 테이블을 삭제하도록 할 수 있다.

[0021] 본 발명의 다른 일 실시예는, 공유 메모리 서버 클러스터를 구성하는 복수의 공유 메모리 서버들에 구비된 원격 메모리들로부터 할당된 공유 메모리 버퍼 세그먼트들로 구성된 공유 메모리 버퍼에 대하여, 메모리에 상기 공유 메모리 버퍼와 동일한 크기만큼 로컬 공유 메모리 영역을 할당하는 단계; 상기 공유 메모리 버퍼와 상기 로컬 공유 메모리 영역 사이의 메모리 맵핑 테이블을 관리하는 단계; 상기 원격 메모리들에 직접 접근하여 데이터를 송수신하여 상기 공유 메모리 버퍼와 상기 로컬 공유 메모리 영역을 동기화하는 단계; 및 상기 로컬 공유 메모리 영역에 대한 주어진 연산을 수행하는 단계를 포함하는 것을 특징으로 하는, 원격 직접 메모리 접근을 통한 분산 처리 방법을 제공한다.

[0022] 이때, 직접 상기 공유 메모리 버퍼를 생성하여 동일한 분산 처리 프레임워크를 구성하는 다른 분산 처리 장치들에 공유 메모리 버퍼 정보를 공유하거나, 상기 다른 분산 처리 장치에 의하여 생성된 상기 공유 메모리 버퍼에 상응하는 공유 메모리 버퍼 정보를 수신하여 상기 공유 메모리 버퍼 정보를 획득하는 단계를 더 포함할 수 있다.

[0023] 이때, 상기 공유 메모리 버퍼 정보를 획득하는 단계는 상기 공유 메모리 서버들에 각각에 상응하는 상기 공유 메모리 버퍼 세그먼트들의 크기를 계산하는 단계; 및 상기 공유 메모리 서버들에 상기 공유 메모리 버퍼 세그먼트들의 생성 및 할당을 요청하여 상기 공유 메모리 버퍼를 생성하는 단계를 포함할 수 있다.

[0024] 이때, 상기 동기화하는 단계는 상기 연산에 의하여 상기 로컬 공유 메모리 영역의 데이터가 변경된 경우에 상기 로컬 공유 메모리 영역의 데이터를 상기 공유 메모리 버퍼에 복사하여 상기 원격 메모리들과 데이터를 동기화하

고, 상기 다른 분산 처리 장치들에 의하여 상기 공유 메모리 버퍼의 데이터가 변경된 경우에 상기 공유 메모리 버퍼의 데이터를 상기 로컬 공유 메모리 영역으로 복사하여 상기 원격 메모리들과 데이터를 동기화할 수 있다.

[0025] 이때, 두 개 이상의 공유 메모리 버퍼들 사이의 데이터 누적 연산을 수행하는 단계를 더 포함할 수 있다.

[0026] 이때, 상기 데이터 누적 연산을 수행하는 단계는 상기 공유 메모리 서버들에 누적 연산을 요청하는 단계; 및 상기 공유 메모리 서버들로부터 상기 공유 메모리 버퍼 세그먼트들에 대하여 누적 연산을 수행한 결과를 수신하는 단계를 포함할 수 있다.

[0027] 이때, 상기 공유 메모리 서버들에 상기 공유 메모리 버퍼 세그먼트들의 해제 및 삭제를 요청하고, 상기 공유 메모리 버퍼 세그먼트들의 해제 및 삭제 요청의 결과들을 수신함에 따라 상기 로컬 공유 메모리 영역을 해제 및 삭제하고, 상기 메모리 맵핑 테이블을 삭제하여 상기 공유 메모리 버퍼의 사용을 종료하는 단계를 더 포함할 수 있다.

발명의 효과

[0028] 본 발명에 따르면, 원격 직접 메모리 접근을 통한 분산 처리 장치 및 그 방법에 의해, 원격 직접 메모리 접근을 통하여 분산 처리 데이터를 공유함으로써 분산 처리시에 발생하는 통신 오버로드를 효과적으로 낮출 수 있다.

[0029] 또한, 본 발명은 원격 직접 메모리 접근을 통한 분산 처리 장치 및 그 방법에 의해, 다수의 공유 메모리 서버들을 클러스터링하고 각각의 공유 메모리 서버들로부터 공유 메모리 버퍼 세그먼트들을 할당하여 공유 메모리 버퍼를 구성하고 분산 처리 장치가 공유 메모리 버퍼에 원격 직접 메모리 접근함으로써, 공유 메모리 서버들 사이의 별도의 동기화 작업이 없이 효율적으로 메모리 데이터를 관리할 수 있다.

도면의 간단한 설명

[0030] 도 1은 본 발명의 일 실시예에 따른 원격 직접 메모리 접근을 통한 분산 처리 시스템의 구성을 나타낸 도면이다.

도 2는 도 1에 도시된 원격 직접 메모리 접근을 통한 분산 처리 장치의 일 예를 나타낸 블록도이다.

도 3은 도 1에 도시된 공유 메모리 서버의 일 예를 나타낸 블록도이다.

도 4는 본 발명의 일 실시예에 따른 공유 메모리 버퍼를 구성하는 방법을 나타낸 도면이다.

도 5는 본 발명의 일 실시예에 따른 원격 직접 메모리 접근을 통한 분산 처리 방법을 나타낸 동작 흐름도이다.

도 6은 도 5에 도시된 공유 메모리 버퍼를 생성 및 할당하는 단계의 일 예를 나타낸 동작 흐름도이다.

도 7은 도 5에 도시된 공유 메모리 버퍼를 해제 및 삭제하는 단계의 일 예를 나타낸 동작 흐름도이다.

도 8은 본 발명의 일 실시예에 따른 공유 메모리 버퍼들의 데이터 누적 연산 방법을 나타낸 동작이다.

도 9는 도 8에 도시된 공유 메모리 버퍼들의 데이터 누적 연산 방법을 나타낸 동작 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0031] 본 발명은 다양한 변환을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다. 본 발명의 효과 및 특징, 그리고 그것들을 달성하는 방법은 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 여기서, 반복되는 설명, 본 발명의 요지를 불필요하게 흐릴 수 있는 공지 기능, 및 구성에 대한 상세한 설명은 생략한다. 본 발명의 실시형태는 당 업계에서 평균적인 지식을 가진 자에게 본 발명을 보다 완전하게 설명하기 위해서 제공되는 것이다. 따라서, 도면에서의 요소들의 형상 및 크기 등은 보다 명확한 설명을 위해 과장될 수 있다.

[0032] 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 각 실시예들의 전부 또는 일부가 선택적으로 조합되어 구성되어 다양한 형태로 구현될 수 있다. 이하의 실시예에서, 제1, 제2 등의 용어는 한정적인 의미가 아니라 하나의 구성 요소를 다른 구성 요소와 구별하는 목적으로 사용되었다. 또한, 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는한 복수의 표현을 포함한다. 또한, 포함하다 또는 가지다 등의 용어는 명세서상에 기재된 특징, 또는 구성요소가 존재함을 의미하는 것이고, 하나 이상의 다른 특징들 또는 구성요소가 부가될 가능성을 미리 배제하는 것은 아니다.

[0033] 이하, 첨부된 도면을 참조하여 본 발명의 실시예들을 상세히 설명하기로 하며, 도면을 참조하여 설명할 때 동일하거나 대응하는 구성 요소는 동일한 도면 부호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다.

[0035] 도 1은 본 발명의 일 실시예에 따른 원격 직접 메모리 접근을 통한 분산 처리 시스템(100)의 구성을 나타낸 도면이다.

[0036] 도 1을 참조하면, 본 발명의 일 실시예에 따른 원격 직접 메모리 접근을 통한 분산 처리 시스템(100)에서 복수의 원격 직접 메모리 접근을 통한 분산 처리 장치들(110)은 원격 직접 메모리 접근(RDMA: Remote Direct Memory Access) 지원 네트워크(130)을 통해 복수의 공유 메모리 서버들(120)과 상호 연결된다. 여기서, 공유 메모리 서버들(120)은 하나의 공유 메모리 서버 클러스터(140)를 구성한다.

[0037] 본 발명의 일 실시예에 따른 원격 직접 메모리 접근을 통한 분산 처리 장치(110)는 공유 메모리 서버 클러스터를 구성하는 복수의 공유 메모리 서버들에 구비된 원격 메모리들로부터 할당된 공유 메모리 버퍼 세그먼트들로 구성된 공유 메모리 버퍼에 대하여, 메모리에 공유 메모리 버퍼와 동일한 크기만큼 로컬 공유 메모리 영역을 할당하고, 공유 메모리 버퍼와 로컬 공유 메모리 영역 사이의 메모리 맵핑 테이블을 이용하여 공유 메모리 버퍼와 로컬 공유 메모리 영역을 동기화하는 것을 특징으로 한다. 그리고, 로컬 공유 메모리 영역에 대하여 주어진 혹은 입력된 연산을 처리한다.

[0038] 여기서, 복수의 원격 직접 메모리 접근을 통한 분산 처리 장치들(110)은 하나의 분산 처리 프레임워크에 포함될 수 있다. 또한, 하나 이상의 원격 직접 메모리 접근을 통한 분산 처리 장치들(110)은 하나의 계산 노드로 구성되어 연산 기능을 제공할 수 있다.

[0039] 이때, 원격 직접 메모리 접근을 통한 분산 처리 장치(110)는 분산 처리 프레임워크에서 주도적으로 분산 처리 작업의 초기화 및 제어를 처리하는 마스터 분산 처리 장치 또는 마스터 분산 처리 장치의 제어를 받으며 계산을 담당하는 슬레이브 분산 처리 장치 혹은 워커 분산 처리 장치로 구분될 수 있다.

[0040] 이때, 마스터 분산 처리 장치는 공유 메모리 서버들(120)에 데이터를 저장하기 위한 공유 메모리 버퍼를 생성하고 슬레이브 분산 처리 장치들에 공유 메모리 버퍼 정보를 전달하여 모든 분산 처리 장치들(110)이 공유 메모리 서버(120)상의 동일 메모리 세그먼트 영역을 접근할 수 있도록 한다. 여기서, 공유 메모리 버퍼 정보에는 공유 메모리 버퍼 전체 크기, 공유 메모리 버퍼 생성키, 공유 메모리 서버별로 생성된 공유 메모리 버퍼 세그먼트 정보 등이 포함될 수 있다.

[0041] 이때, 원격 직접 메모리 접근을 통한 분산 처리 장치(110)는 연산에 의하여 로컬 공유 메모리 영역의 데이터가 변경된 경우에 로컬 공유 메모리 영역의 데이터를 공유 메모리 버퍼에 복사하여 원격 메모리들과 데이터를 동기화할 수 있다. 또한, 다른 원격 직접 메모리 접근을 통한 분산 처리 장치(110)의 연산에 의하여 공유 메모리 버퍼의 데이터가 변경된 경우에 공유 메모리 버퍼의 데이터를 로컬 공유 메모리 영역으로 복사하여 원격 메모리들과 데이터를 동기화할 수 있다.

[0042] 공유 메모리 서버(120)는 분산 처리 프레임워크에서 공유 메모리를 제공하는 장치이다.

[0043] 여기서, 복수의 공유 메모리 서버들(120)은 하나의 공유 메모리 서버 클러스터를 구성할 수 있다. 또한, 하나 이상의 공유 메모리 서버들(110)은 하나의 메모리 서비스 노드로 구성되어 공유 메모리 서비스를 제공할 수 있다.

[0044] 이때, 복수의 공유 메모리 서버들(120)은 각각 공유 메모리 버퍼의 생성 및 할당 요청에 따라 공유 메모리 버퍼 세그먼트를 생성 및 할당하여, 각각의 공유 메모리 버퍼 세그먼트들을 연결한 가상의 공유 메모리 버퍼를 제공할 수 있다. 여기서, 공유 메모리 버퍼는 분산 처리 장치(110)가 원격 직접 메모리 접근 지원 네트워크(130)를 통하여 직접 접근할 수 있다.

[0045] 이때, 공유 메모리 버퍼는 분산 처리 장치들(110)의 로컬 공유 메모리 영역과 동기화될 수 있다. 즉, RDMA 읽기/쓰기를 통하여 동기화할 수 있다.

[0046] 이때, 공유 메모리 서버(120)는 공유 메모리 버퍼 세그먼트들 간의 누적 연산 기능을 제공할 수 있다.

[0047] 원격 직접 메모리 접근 지원 네트워크(130)는 복수의 분산 처리 장치들(110)과 복수의 공유 메모리 서버들(120) 사이의 통신을 제공하는 네트워크로, 분산 처리 장치들(110)이 공유 메모리 서버들(120)의 메모리에 직접 접근 가능한 기능을 제공한다.

[0048] 즉, 원격 직접 메모리 접근을 지원하는 고성능 네트워크(130)로 연결된 고성능 컴퓨팅 클러스터 시스템 환경에

서 분산 처리 장치들(110)이 다수의 공유 메모리 서버들(120)의 물리 메모리 세그먼트들을 결합하여 가상의 연속된 공유 메모리 버퍼에 직접 접근할 수 있도록 함으로써, 분산 처리 장치들 간의 데이터 공유를 가속화하고 효율성을 높일 수 있다.

[0049] 이와 같은 공유 메모리 형태로 분산 처리 데이터를 공유하는 대표적인 분산 처리 방식으로는 비동기 딥러닝 트레이닝에 이용될 수 있다. 비동기 딥러닝 트레이닝 방식은 데이터 병렬 딥러닝 학습 방식의 하나로, 학습 데이터를 나누어 다수의 딥러닝 프로세스가 학습을 수행하고, 학습하는 도중에 학습한 내용을 다른 딥러닝 프로세스들과 공유 데이터 버퍼를 통해 비동기적으로 공유하는 학습 방법이다. 비동기 딥러닝 트레이닝 방식에서 각 딥러닝 분산 처리 프로세스는 딥러닝 파라미터(딥러닝 트레이닝에서 트레이닝의 대상이 되는 가중치와 특징값의 총칭)를 다른 프로세스들과 동기를 맞추지 않고 비동기적으로 파라미터를 업데이트하는데, 이 방식은 본 발명에서 제안하는 공유 메모리 구조에 적합하다.

[0050] 또한, 본 발명에서 제안하는 방식은 파라미터 서버와 일부 유사하나, 딥러닝 분산 프로세스들로부터 그래디언트를 받아 능동적으로 가중치 파라미터를 계산하여 직접 딥러닝 파라미터를 업데이트하는 파라미터 서버 방식과 달리 분산 처리 장치들이 원격 직접 메모리 접근 기능을 통해 공유 메모리 서버의 개입 없이 공유 메모리 버퍼를 직접 읽고 쓰는 것이 가능하다.

[0051] 또한, 하나의 단일 메모리 서버상의 메모리만을 공유 메모리로 사용할 경우에는 확장성에 제한이 있으나, 공유 메모리 서버 클러스터를 구성함으로써 대규모 분산 처리에도 유연한 확장성을 제공할 수 있다.

[0053] 도 2는 도 1에 도시된 원격 직접 메모리 접근을 통한 분산 처리 장치(110)의 일 예를 나타낸 블록도이다.

[0054] 도 2를 참조하면, 본 발명의 일 실시예에 따른 원격 직접 메모리 접근을 통한 분산 처리 장치(110)는 제어부(210), 통신부(220), 메모리(230), 연산 처리부(240) 및 공유 메모리 서버 접근 지원부(250) 등을 포함한다.

[0055] 상세히, 제어부(210)는 일종의 중앙처리장치로서 원격 직접 메모리 접근을 통한 분산 처리 과정을 제어한다. 즉, 제어부(210)는 메모리(230), 연산 처리부(240) 및 공유 메모리 서버 접근 지원부(250) 등을 제어하여 다양한 기능을 제공할 수 있다.

[0056] 여기서, 제어부(210)는 프로세서(processor)와 같이 데이터를 처리할 수 있는 모든 종류의 장치를 포함할 수 있다. 여기서, '프로세서(processor)'는, 예를 들어 프로그램 내에 포함된 코드 또는 명령으로 표현된 기능을 수행하기 위해 물리적으로 구조화된 회로를 갖는, 하드웨어에 내장된 데이터 처리 장치를 의미할 수 있다. 이와 같이 하드웨어에 내장된 데이터 처리 장치의 일 예로써, 마이크로프로세서(microprocessor), 중앙처리장치(central processing unit: CPU), 프로세서 코어(processor core), 멀티프로세서(multiprocessor), ASIC(application-specific integrated circuit), FPGA(field programmable gate array) 등의 처리 장치를 망라할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.

[0057] 통신부(220)는 RDMA 지원 네트워크(도 1의 130 참조)를 통하여 원격 직접 메모리 접근을 통한 분산 처리 장치(110)와 공유 메모리 서버(도 1의 120 참조) 간의 송수신 신호를 전송하는데 필요한 통신 인터페이스를 제공한다.

[0058] 여기서, 통신부(220)는 다른 네트워크 장치와 유무선 연결을 통해 제어 신호 또는 데이터 신호와 같은 신호를 송수신하기 위해 필요한 하드웨어 및 소프트웨어를 포함하는 장치일 수 있다.

[0059] 이때, 통신부(220)는 RDMA 지원 네트워크(도 1의 130 참조)를 통해 공유 메모리 서버들(도 1의 120 참조)의 원격 메모리들에 직접 접근하여 데이터를 읽고 쓸 수 있다.

[0060] 메모리(230)는 제어부(210)가 처리하는 데이터를 일시적 또는 영구적으로 저장하는 기능을 수행한다. 여기서, 메모리(230)는 자기 저장 매체(magnetic storage media) 또는 플래시 저장 매체(flash storage media)를 포함할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.

[0061] 이때, 메모리(230)는 공유 메모리 서버들(도 1의 120 참조)의 원격 메모리들로부터 구성된 공유 메모리 버퍼와 동일한 크기만큼을 로컬 공유 메모리 영역으로 할당하고, 공유 메모리 버퍼와 동기화할 수 있다.

[0062] 이에 따라, 분산 처리 장치(110)는 로컬 공유 메모리 영역에 대하여 연산을 수행하고, 이를 공유 메모리 버퍼와 동기화함으로써 다른 분산 처리 장치들과 메모리를 공유할 수 있다.

[0063] 연산 처리부(240)는 분산 처리 프레임워크에서 분산 처리 장치(110)에 주어진 연산을 수행한다.

[0064] 이때, 연산 처리부(240)는 로컬 공유 메모리 영역에 대하여 연산을 수행할 수 있다.

[0065] 이때, 연산 처리부(240)는 API(Application Programmable Interface)를 통해 공유 메모리 서버 접근 지원부(250)에 명시적으로 로컬 공유 메모리 영역과 공유 메모리 버퍼의 동기화를 요청 및 수행할 수 있다.

[0066] 이때, 연산 처리부(240)는 API(Application Programmable Interface)를 통해 공유 메모리 서버 접근 지원부(250)에 복수의 공유 메모리 버퍼들에 대한 누적 연산을 요청 및 수행할 수 있다.

[0067] 공유 메모리 서버 접근 지원부(250)는 공유 메모리 서버(도 1의 120 참조)와의 RDMA 읽기/쓰기를 통한 접근을 지원하여, 공유 메모리 버퍼와의 동기화를 지원한다.

[0068] 이때, 공유 메모리 서버 접근 지원부(250)는 공유 메모리 서버 클러스터를 등록하여 공유 메모리 서버 클러스터를 구성하는 공유 메모리 서버들(도 1의 120 참조)의 정보를 획득할 수 있다. 여기서, 공유 메모리 서버 클러스터의 등록은 사용자가 입력한 공유 메모리 서버들(도 1의 120 참조)의 접근 정보를 이용하여 해당 공유 메모리 서버들(도 1의 120 참조)과의 연결을 설정하고 공유 메모리 버퍼 세그먼트를 생성 및 할당받는 초기화 과정을 의미할 수 있다. 그리고, 공유 메모리 서버 접근 정보에는 IP 주소와 포트 번호 정보 등이 포함될 수 있다. 특히, 공유 메모리 서버 클러스터를 등록할 때, 모든 분산 처리 장치들(110)은 모든 공유 메모리 서버들(도 1의 120 참조)의 순서를 동일하게 등록할 수 있다.

[0069] 이때, 공유 메모리 서버 접근 지원부(250)는 공유 메모리 서버 클러스터에 등록된 각 공유 메모리 서버들(도 1의 120 참조)에 대하여 공유 메모리 버퍼를 구성하는 공유 메모리 버퍼 세그먼트들의 크기를 계산할 수 있다. 예컨대, 하나의 공유 메모리 서버 클러스터에 5개의 공유 메모리 서버들이 포함되어 있고, 공유 메모리 버퍼의 크기를 5GB로 구성하는 경우, 5개의 크기 1GB의 공유 메모리 버퍼 세그먼트들로 나눌 수 있다. 여기서, 각 공유 메모리 버퍼 세그먼트들의 크기는 동일하지 않을 수 있다.

[0070] 이때, 공유 메모리 서버 접근 지원부(250)는 공유 메모리 서버들(도 1의 120 참조)에 계산된 공유 메모리 서버 세그먼트들의 크기 정보와 공유 메모리 버퍼 세그먼트의 생성 및 할당을 요청을 전달하여 공유 메모리 버퍼를 구성할 수 있다.

[0071] 이때, 공유 메모리 서버 접근 지원부(250)는 다른 분산 처리 장치가 구성한 공유 메모리 버퍼에 대한 접근권을 획득하여 공유 메모리 버퍼의 생성 및 할당을 대신할 수 있다.

[0072] 이때, 공유 메모리 서버 접근 지원부(250)는 공유 메모리 버퍼를 구성하기 위하여 공유 메모리 서버들(도 1의 120 참조)에 공유 메모리 버퍼 생성키를 함께 전송할 수 있다. 여기서, 공유 메모리 버퍼 생성키는 동일한 공유 메모리 버퍼가 중복 생성을 방지하거나 유효한 요청인지 여부를 확인하거나 공유 메모리 버퍼 세그먼트를 특정하기 위한 목적으로 이용될 수 있다.

[0073] 이때, 공유 메모리 서버 접근 지원부(250)는 모든 공유 메모리 서버들(도 1의 120 참조)에 대하여 공유 메모리 버퍼 세그먼트들의 생성 및 할당이 이루어지면 메모리(230)에 공유 메모리 버퍼와 동일한 크기로 로컬 공유 메모리 영역을 할당할 수 있다. 여기서, 로컬 공유 메모리 영역은 실제 물리 메모리에 할당되며, 로컬 공유 메모리 영역이 할당됨에 따라 주소 정보(예컨대, 가상 주소)가 반환된다.

[0074] 이때, 공유 메모리 서버 접근 지원부(250)는 로컬 공유 메모리 영역을 공유 메모리 버퍼와 동기화를 수행할 수 있다. 여기서, 동기화는 메모리 맵핑 테이블을 이용하여 수행될 수 있다.

[0075] 이때, 공유 메모리 서버 접근 지원부(250)는 연산 처리부(240)에 의하여 로컬 공유 메모리 영역의 데이터가 변경된 경우, RDMA 통해 로컬 공유 메모리 영역의 데이터를 공유 메모리 버퍼에 복사하여 동기화할 수 있다. 또는, 변경된 데이터에 대하여만 복사하여 동기화할 수 있다.

[0076] 이때, 공유 메모리 서버 접근 지원부(250)는 다른 분산 처리 장치의 연산에 의하여 공유 메모리 버퍼의 데이터가 변경된 경우, RDMA를 통해 공유 메모리 버퍼의 데이터를 로컬 공유 메모리 영역에 복사하여 동기화할 수 있다. 또는, 변경된 데이터에 대하여만 복사하여 동기화할 수 있다.

[0077] 이때, 공유 메모리 서버 접근 지원부(250)는 공유 메모리 버퍼의 사용이 종료된 경우 공유 메모리 서버들(도 1의 120 참조)에 공유 메모리 버퍼 세그먼트들의 해제 및 삭제를 요청할 수 있다.

[0078] 이때, 공유 메모리 서버 접근 지원부(250)는 모든 공유 메모리 버퍼 세그먼트들의 해제 및 삭제가 이루어지면, 공유 메모리 서버들(도 1의 120 참조)과의 연결을 종료하고 공유 메모리 서버 클러스터를 등록 해제하며 정보를

삭제하여 공유 메모리 버퍼의 사용을 종료할 수 있다.

[0079] 이때, 공유 메모리 서버 접근 지원부(250)는 복수의 공유 메모리 버퍼에 대하여 데이터 누적 연산 기능을 제공할 수 있다. 예컨대, 제1 공유 메모리 버퍼와 제2 공유 메모리 버퍼에 대한 누적 연산을 수행하는 경우, 제1 공유 메모리 버퍼에 대한 데이터 동기화를 수행하고, 각 공유 메모리 서버들(도 1의 120 참조)에 제1 공유 버퍼 세그먼트들로부터 제2 공유 버퍼 세그먼트들로의 누적 연산을 요청하고, 모든 공유 메모리 서버들(도 1의 120 참조)에서 누적 연산이 완료되면 그 결과를 반환할 수 있다. 각 공유 메모리 서버들(도 1의 120 참조)에서는 누적 연산을 위하여 제2 공유 메모리 버퍼 세그먼트를 잠그고, 제1 공유 메모리 버퍼 세그먼트에서 제2 공유 메모리 버퍼 세그먼트로의 누적 연산을 수행할 수 있다.

[0080] 메모리 맵핑 테이블 관리부(260)는 로컬 공유 메모리 영역과 공유 메모리 버퍼 사이의 메모리 맵핑 테이블을 관리한다.

[0081] 이때, 메모리 맵핑 테이블 관리부(260)는 저장소를 포함하여 직접 메모리 맵핑 테이블을 저장하여 관리할 수도 있지만, 별도의 저장소나 메모리(230)에 메모리 맵핑 테이블을 저장하여 관리할 수 있다.

[0082] 이때, 메모리 맵핑 테이블 관리부(260)는 공유 메모리 버퍼가 생성되면 공유 메모리 버퍼와 로컬 공유 메모리 영역 사이의 메모리 맵핑 테이블을 생성할 수 있다.

[0083] 이때, 메모리 맵핑 테이블 관리부(260)는 공유 메모리 버퍼의 사용이 종료되면 메모리 맵핑 테이블을 삭제할 수 있다.

[0085] 도 3은 도 1에 도시된 공유 메모리 서버(120)의 일 예를 나타낸 블록도이다.

[0086] 도 3을 참조하면, 본 발명의 일 실시예에 따른 공유 메모리 서버(120)는 제어부(310), 통신부(320), 메모리(330) 및 공유 메모리 관리부(340) 등을 포함한다.

[0087] 상세히, 제어부(310)는 일종의 중앙처리장치로서 원격 직접 메모리 접근을 통한 분산 처리 과정을 제어한다. 즉, 제어부(310)는 메모리(330) 및 공유 메모리 관리부(340) 등을 제어하여 다양한 기능을 제공할 수 있다.

[0088] 여기서, 제어부(310)는 프로세서(processor)와 같이 데이터를 처리할 수 있는 모든 종류의 장치를 포함할 수 있다. 여기서, '프로세서(processor)'는, 예를 들어 프로그램 내에 포함된 코드 또는 명령으로 표현된 기능을 수행하기 위해 물리적으로 구조화된 회로를 갖는, 하드웨어에 내장된 데이터 처리 장치를 의미할 수 있다. 이와 같이 하드웨어에 내장된 데이터 처리 장치의 일 예로써, 마이크로프로세서(microprocessor), 중앙처리장치(central processing unit: CPU), 프로세서 코어(processor core), 멀티프로세서(multiprocessor), ASIC(application-specific integrated circuit), FPGA(field programmable gate array) 등의 처리 장치를 망라할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.

[0089] 통신부(320)는 RDMA 지원 네트워크(도 1의 130 참조)를 통하여 공유 메모리 서버(120)와 원격 직접 메모리 접근을 통한 분산 처리 장치들(도 1의 110 참조) 간의 송수신 신호를 전송하는데 필요한 통신 인터페이스를 제공한다.

[0090] 여기서, 통신부(320)는 다른 네트워크 장치와 유무선 연결을 통해 제어 신호 또는 데이터 신호와 같은 신호를 송수신하기 위해 필요한 하드웨어 및 소프트웨어를 포함하는 장치일 수 있다.

[0091] 이때, 통신부(320)는 RDMA 지원 네트워크(도 1의 130 참조)를 통해 원격 직접 메모리 접근을 통한 분산 처리 장치들(도 1의 110 참조)이 메모리(330)에 직접 접근하여 데이터를 읽고 쓸 수 있도록 지원할 수 있다.

[0092] 메모리(330)는 제어부(310)가 처리하는 데이터를 일시적 또는 영구적으로 저장하는 기능을 수행한다. 여기서, 메모리(330)는 자기 저장 매체(magnetic storage media) 또는 플래시 저장 매체(flash storage media)를 포함할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.

[0093] 이때, 메모리(330)는 전체 또는 일부가 공유 메모리 버퍼 세그먼트로 할당되어, 다른 공유 메모리 서버들의 공유 메모리 버퍼 세그먼트들과 함께 공유 메모리 버퍼를 구성할 수 있다. 여기서, 공유 메모리 버퍼는 공유 메모리 버퍼 세그먼트들을 연결한 가상의 메모리 버퍼로, 실체는 각 공유 메모리 서버들(120)의 메모리(330)에 할당된 공유 메모리 버퍼 세그먼트들의 영역이다.

[0094] 이때, 메모리(330)에 할당된 공유 메모리 버퍼 세그먼트는 분산 처리 장치(도 1의 110 참조)의 로컬 공유 메모리 영역의 데이터가 변경됨에 따라 변경된 데이터가 동기화될 수 있다.

[0095] 공유 메모리 관리부(340)는 공유 메모리 버퍼를 구성하기 위하여 공유 메모리 버퍼 세그먼트를 할당하고 이를 관리한다.

[0096] 이때, 공유 메모리 관리부(340)는 분산 처리 장치(도 1의 110 참조)로부터 공유 메모리 버퍼의 생성 및 할당을 요청받은 경우, 주어진 공유 메모리 버퍼 세그먼트의 크기만큼 메모리(330)에서 공유 메모리 버퍼 세그먼트를 생성 및 할당하여 할당 정보를 반환할 수 있다.

[0097] 이때, 공유 메모리 관리부(340)는 공유 메모리 버퍼 세그먼트를 생성 요청에 대하여 공유 메모리 버퍼 생성키를 수신하고, 수신한 공유 메모리 버퍼 생성키가 이미 사용중이 아닌 경우에만 공유 메모리 버퍼 세그먼트를 생성 및 할당하여 할당 정보를 반환할 수 있다.

[0098] 이때, 공유 메모리 관리부(340)는 공유 메모리 버퍼 세그먼트의 접근 요청에 대하여 공유 메모리 버퍼 생성키를 수신하고, 수신한 공유 메모리 버퍼 생성키가 접근을 요청하는 공유 메모리 버퍼 세그먼트의 정보와 일치하는 경우에 해당 공유 메모리 버퍼 세그먼트의 접근을 허용할 수 있다. 예컨대, 공유 메모리 버퍼 생성키는 공유 메모리 버퍼 세그먼트의 크기를 의미할 수 있다.

[0099] 이때, 공유 메모리 관리부(340)는 공유 메모리 버퍼 세그먼트들 사이의 데이터 누적 연산을 수행할 수 있다. 예컨대, 제1 공유 메모리 버퍼 세그먼트로부터 제2 공유 메모리 버퍼 세그먼트로의 데이터 누적은, 제2 공유 메모리 버퍼 세그먼트를 잠그고 제1 공유 메모리 버퍼 세그먼트의 데이터를 제2 공유 메모리 버퍼 세그먼트에 누적함으로써 수행될 수 있다. 그리고, 데이터 누적 연산이 완료된 경우 연산 완료를 알리는 결과를 반환할 수 있다.

[0101] 도 4는 본 발명의 일 실시예에 따른 공유 메모리 버퍼를 구성하는 방법을 나타낸 도면이다.

[0102] 도 4를 참조하면, 본 발명의 일 실시예에 따른 공유 메모리 버퍼를 구성하는 방법은, 공유 메모리 버퍼의 생성 및 할당을 주도하는 마스터 분산 처리 장치(410)가 공유 메모리 서버들(440, 450, 460 및 470)에 공유 메모리 버퍼를 구성하기 위한 공유 메모리 버퍼 세그먼트들의 생성 및 할당을 요청한다.

[0103] 이때, 공유 메모리 서버 1(440)은 공유 메모리 버퍼 세그먼트 1(441)을 생성 및 할당하여 그 정보를 반환하고, 공유 메모리 서버 2(450)는 공유 메모리 버퍼 세그먼트 2(451)를 생성 및 할당하여 그 정보를 반환하고, 공유 메모리 서버 3(460)은 공유 메모리 버퍼 세그먼트 3(461)을 생성 및 할당하여 그 정보를 반환하고, 공유 메모리 서버 4(470)는 공유 메모리 버퍼 세그먼트 4(471)를 생성 및 할당하여 그 정보를 반환할 수 있다.

[0104] 이때, 각 공유 메모리 서버들(440, 450, 460 및 470)에서 공유 메모리 버퍼 세그먼트들(441, 451, 461 및 471)이 생성 및 할당되면, 이들은 연결되어 가상의 공유 메모리 버퍼(430)를 구성할 수 있다. 즉, 공유 메모리 버퍼(430)의 실체는 각 공유 메모리 서버들(440, 450, 460 및 470)에 할당된 공유 메모리 버퍼 세그먼트들(441, 451, 461 및 471)이다. 따라서, 공유 메모리 버퍼(430)에 대한 데이터 입출력은 공유 메모리 버퍼 세그먼트들(441, 451, 461 및 471)에 대한 데이터 입출력을 의미한다.

[0105] 이때, 마스터 분산 처리 장치(410)와 동일한 분산 처리 프레임워크에 속하는 다른 분산 처리 장치들은 슬레이브 분산 처리 장치(420)로 분류되며, 슬레이브 분산 처리 장치(420)는 마스터 분산 처리 장치(410)에 의하여 구성된 공유 메모리 버퍼의 정보를 획득하여 동일한 공유 메모리 버퍼를 이용할 수 있다.

[0106] 이때, 마스터 분산 처리 장치(410)와 슬레이브 분산 처리 장치(420)는 각각 자신의 메모리에 대하여 공유 메모리 버퍼와 동일한 크기의 로컬 공유 메모리 영역(411 및 421)을 할당할 수 있다. 그리고, 로컬 공유 메모리 영역(411 및 421)에 대하여 입출력을 수행하여 분산 처리 프레임워크에서의 분산 처리를 수행할 수 있다.

[0107] 이때, 각각의 로컬 공유 메모리 영역(411 및 421)은 공유 메모리 버퍼(430)와 동기화되어 유지되고, 분산 처리 장치들(410 및 420) 사이에서 공유 메모리 버퍼(430)를 통해 분산 처리 데이터를 공유할 수 있다. 특히, 분산 처리 장치들(410 및 420)은 RDMA를 통하여 공유 메모리 버퍼(430)에 대하여 직접 입출력하여 로컬 공유 메모리 영역(411 및 421)과 공유 메모리 버퍼(430) 사이의 동기화를 수행할 수 있다.

[0109] 도 5는 본 발명의 일 실시예에 따른 원격 직접 메모리 접근을 통한 분산 처리 방법을 나타낸 동작 흐름도이다.

[0110] 도 5를 참조하면, 본 발명의 일 실시예에 따른 원격 직접 메모리 접근을 통한 분산 처리 방법은 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 공유 메모리 서버 클러스터를 등록한다(S501). 공유 메모리 서버 클러스터는 복수개의 공유 메모리 서버들(도 1의 120 참조)로 구성될 수 있다.

[0111] 또한, 본 발명의 일 실시예에 따른 원격 직접 메모리 접근을 통한 분산 처리 방법은 원격 직접 메모리 접근을

통한 분산 처리 장치(도 1의 110 참조)가, 공유 메모리 버퍼를 생성 및 할당한다(S503).

[0112] 또한, 본 발명의 일 실시예에 따른 원격 직접 메모리 접근을 통한 분산 처리 방법은 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 주어진 분산 연산을 수행하고 공유 메모리 버퍼에 대하여 RDMA를 통한 직접 데이터 읽기 및 쓰기를 통하여 분산 처리 데이터를 공유한다(S505).

[0113] 또한, 본 발명의 일 실시예에 따른 원격 직접 메모리 접근을 통한 분산 처리 방법은 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 공유 메모리 버퍼의 사용을 종료하는 경우에 공유 메모리 버퍼를 해제 및 삭제한다(S507).

[0114] 또한, 본 발명의 일 실시예에 따른 원격 직접 메모리 접근을 통한 분산 처리 방법은 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 공유 메모리 서버 클러스터의 등록을 해제한다(S509).

[0116] 도 6은 도 5에 도시된 공유 메모리 버퍼를 생성 및 할당하는 단계(S503)의 일 예를 나타낸 동작 흐름도이다.

[0117] 도 6을 참조하면, 도 5에 도시된 공유 메모리 버퍼를 생성 및 할당하는 단계(S503)는 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 공유 메모리 서버별 공유 메모리 버퍼 세그먼트 크기를 계산한다(S601).

[0118] 또한, 도 5에 도시된 공유 메모리 버퍼를 생성 및 할당하는 단계(S503)는 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 공유 메모리 서버들(도 1의 120 참조)에 공유 메모리 버퍼 세그먼트의 생성 및 할당을 요청한다(S603). 여기서, 공유 메모리 버퍼 세그먼트의 생성 및 할당 요청은 공유 메모리 버퍼의 생성 및 할당 요청과 동일한 의미로 사용될 수 있다.

[0119] 또한, 도 5에 도시된 공유 메모리 버퍼를 생성 및 할당하는 단계(S503)는 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 공유 메모리 서버들(도 1의 120 참조)에 의하여 공유 메모리 버퍼 세그먼트들이 생성 및 할당되면 반환되는 정보를 획득한다(S605).

[0120] 또한, 도 5에 도시된 공유 메모리 버퍼를 생성 및 할당하는 단계(S503)는 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 모든 공유 메모리 버퍼 세그먼트들이 생성 및 할당되었는지 여부를 확인한다(S607).

[0121] 또한, 도 5에 도시된 공유 메모리 버퍼를 생성 및 할당하는 단계(S503)는 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 모든 공유 메모리 버퍼 세그먼트들이 생성 및 할당되어 공유 메모리 버퍼가 구성되면, 메모리에 공유 메모리 버퍼와 동일한 크기만큼의 로컬 공유 메모리 영역을 할당하고, 공유 메모리 버퍼와 로컬 공유 메모리 영역 사이의 메모리 맵핑 테이블을 갱신한다(S609).

[0123] 도 7은 도 5에 도시된 공유 메모리 버퍼를 해제 및 삭제하는 단계(S507)의 일 예를 나타낸 동작 흐름도이다.

[0124] 도 7을 참조하면, 도 5에 도시된 공유 메모리 버퍼를 해제 및 삭제하는 단계(S507)는 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 공유 메모리 서버들(도 1의 120 참조)에 공유 메모리 버퍼 세그먼트의 해제 및 삭제를 요청한다(S701). 여기서, 공유 메모리 버퍼 세그먼트의 해제 및 삭제 요청은 공유 메모리 버퍼의 해제 및 삭제 요청과 동일한 의미로 사용될 수 있다.

[0125] 또한, 도 5에 도시된 공유 메모리 버퍼를 해제 및 삭제하는 단계(S507)는 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 공유 메모리 서버들(도 1의 120 참조)에 의하여 공유 메모리 버퍼 세그먼트들이 해제 및 삭제되면 반환되는 정보를 획득한다(S703).

[0126] 또한, 도 5에 도시된 공유 메모리 버퍼를 해제 및 삭제하는 단계(S507)는 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 모든 공유 메모리 버퍼 세그먼트들이 해제 및 삭제되었는지 여부를 확인한다(S705).

[0127] 또한, 도 5에 도시된 공유 메모리 버퍼를 해제 및 삭제하는 단계(S507)는 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 모든 공유 메모리 버퍼 세그먼트들이 해제 및 삭제되어 공유 메모리 버퍼의 사용이 종료되면, 할당된 로컬 공유 메모리 영역을 해제하고, 상응하는 메모리 맵핑 테이블을 삭제한다(S707).

[0129] 도 8은 본 발명의 일 실시예에 따른 공유 메모리 버퍼들의 데이터 누적 연산 방법을 나타낸 동작이다.

[0130] 도 8을 참조하면, 본 발명의 일 실시예에 따른 공유 메모리 버퍼들의 데이터 누적 연산 방법은, 각 공유 메모리 서버들(810, 820 및 830)에서 할당된 공유 메모리 버퍼 세그먼트들에 대하여 데이터 누적 연산을 수행하는 것으

로 이루어진다.

[0131] 첫 번째 공유 메모리 서버(810)에는 제1 공유 메모리 버퍼 세그먼트 1(841) 및 제2 공유 메모리 버퍼 세그먼트 1(851)이 할당되어 있으며, 두 번째 공유 메모리 서버(820)에는 제1 공유 메모리 버퍼 세그먼트 2(842) 및 제2 공유 메모리 버퍼 세그먼트 2(852)가 할당되어 있으며, n 번째 공유 메모리 서버(830)에는 제1 공유 메모리 버퍼 세그먼트 n(843) 및 제2 공유 메모리 버퍼 세그먼트 n(853)이 할당되어 있다. 그리고, 제1 공유 메모리 버퍼 세그먼트 1(841), 제1 공유 메모리 버퍼 세그먼트 2(842) 및 제1 공유 메모리 버퍼 세그먼트 n(843) 등은 제1 공유 메모리 버퍼(840)를 구성한다. 또한, 제2 공유 메모리 버퍼 세그먼트 1(851), 제2 공유 메모리 버퍼 세그먼트 2(852) 및 제2 공유 메모리 버퍼 세그먼트 n(853) 등은 제2 공유 메모리 버퍼(850)를 구성한다.

[0132] 각 공유 메모리 서버들(810, 820 및 830)은 제1 공유 메모리 버퍼(840)의 제1 공유 메모리 버퍼 세그먼트(841, 842 및 843)의 데이터를 제2 공유 메모리 버퍼(850)의 제2 공유 메모리 버퍼 세그먼트(851, 852 및 853)에 누적하여 공유 메모리 버퍼들 간의 데이터 누적 연산을 수행할 수 있다.

[0134] 도 9는 도 8에 도시된 공유 메모리 버퍼들의 데이터 누적 연산 방법을 나타낸 동작 흐름도이다.

[0135] 도 9를 참조하면, 도 8에 도시된 공유 메모리 버퍼들의 데이터 누적 연산 방법은 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 제1 공유 메모리 버퍼의 데이터를 로컬 공유 메모리 영역과 동기화한다(S901).

[0136] 또한, 도 8에 도시된 공유 메모리 버퍼들의 데이터 누적 연산 방법은 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 공유 메모리 서버들(도 1의 120 참조)에 제1 공유 메모리 버퍼로부터 제2 공유 메모리 버퍼로의 데이터 누적 연산을 요청한다(S903).

[0137] 또한, 도 8에 도시된 공유 메모리 버퍼들의 데이터 누적 연산 방법은 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 공유 메모리 서버들(도 1의 120 참조)로부터 제2 공유 메모리 버퍼 세그먼트를 잠근 이후 제1 공유 메모리 버퍼 세그먼트의 데이터를 제2 공유 메모리 버퍼 세그먼트에 누적한 결과를 수신한다(S905).

[0138] 또한, 도 8에 도시된 공유 메모리 버퍼들의 데이터 누적 연산 방법은 원격 직접 메모리 접근을 통한 분산 처리 장치(도 1의 110 참조)가, 모든 공유 메모리 버퍼 세그먼트들에 대한 누적 연산이 완료되었는지 확인하여 데이터 누적 연산의 결과를 반환한다(S907).

[0140] 본 발명에서 설명하는 특정 실행들은 실시예들로서, 어떠한 방법으로도 본 발명의 범위를 한정하는 것은 아니다. 명세서의 간결함을 위하여, 종래 전자적인 구성들, 제어시스템들, 소프트웨어, 상기 시스템들의 다른 기능적인 측면들의 기재는 생략될 수 있다. 또한, 도면에 도시된 구성 요소들 간의 선들의 연결 또는 연결 부재들은 기능적인 연결 및/또는 물리적 또는 회로적 연결들을 예시적으로 나타낸 것으로서, 실제 장치에서는 대체 가능하거나 추가의 다양한 기능적인 연결, 물리적인 연결, 또는 회로 연결들로서 나타내어질 수 있다. 또한, “필수적인”, “중요하게” 등과 같이 구체적인 언급이 없다면 본 발명의 적용을 위하여 반드시 필요한 구성 요소가 아닐 수 있다.

[0141] 따라서, 본 발명의 사상은 상기 설명된 실시예에 국한되어 정해져서는 아니되며, 후술하는 특허청구범위뿐만 아니라 이 특허청구범위와 균등한 또는 이로부터 등가적으로 변경된 모든 범위는 본 발명의 사상의 범주에 속한다고 할 것이다.

부호의 설명

[0142] 100: 원격 직접 메모리 접근을 통한 분산 처리 시스템

110: 원격 직접 메모리 접근을 통한 분산 처리 장치

120: 공유 메모리 서버 130: RDMA 지원 네트워크

210: 제어부 220: 통신부

230: 메모리 240: 연산 처리부

250: 공유 메모리 서버 접근 관리부

260: 메모리 맵핑 테이블 관리부

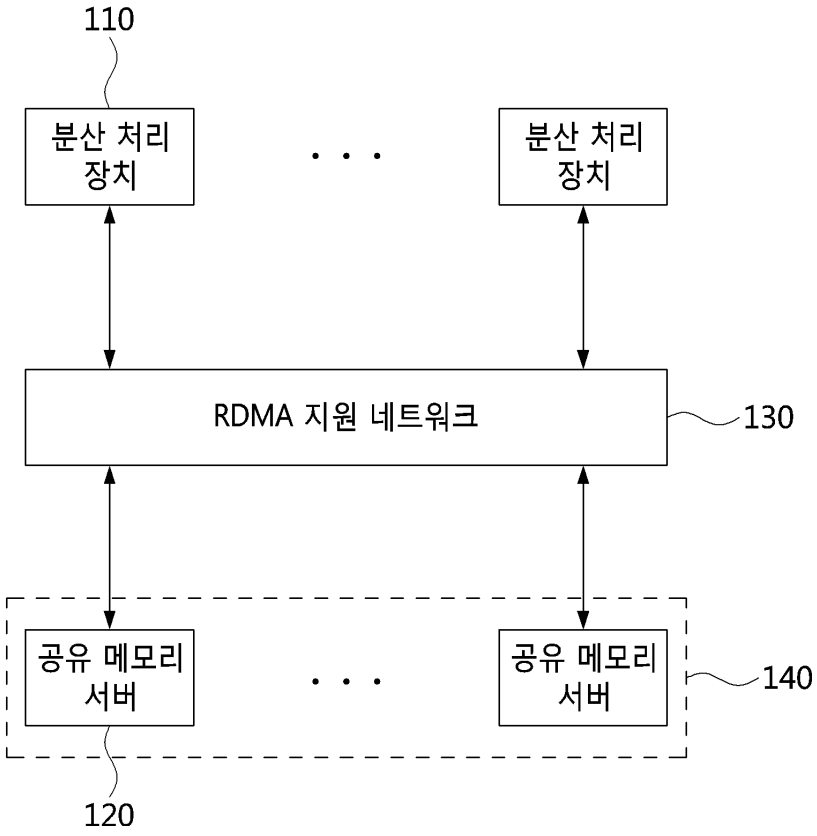
310: 제어부 320: 통신부

330: 메모리 340: 공유 메모리 관리부

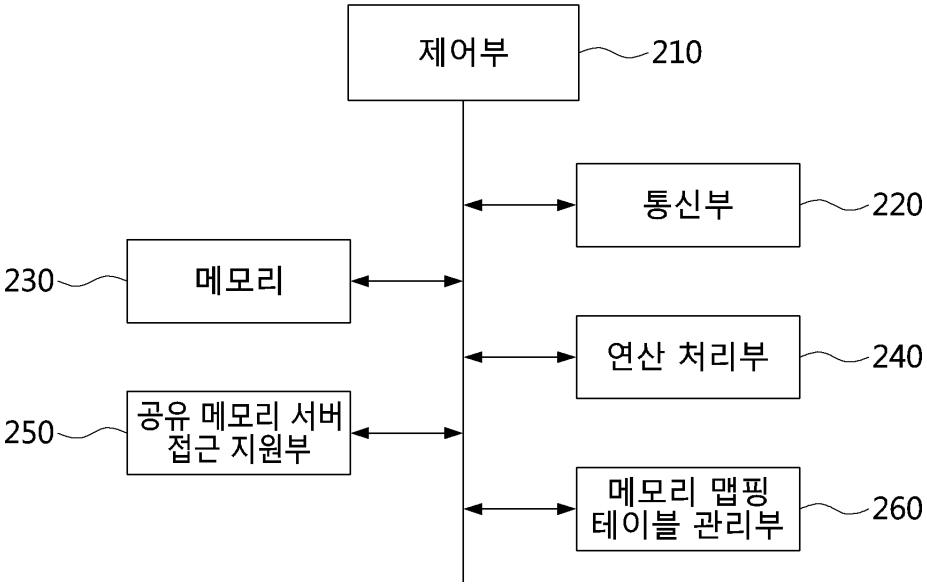
도면

도면1

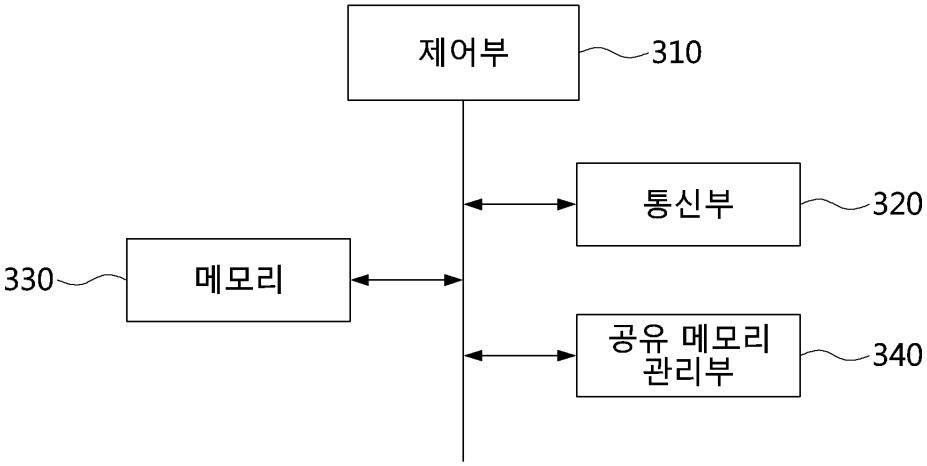
100



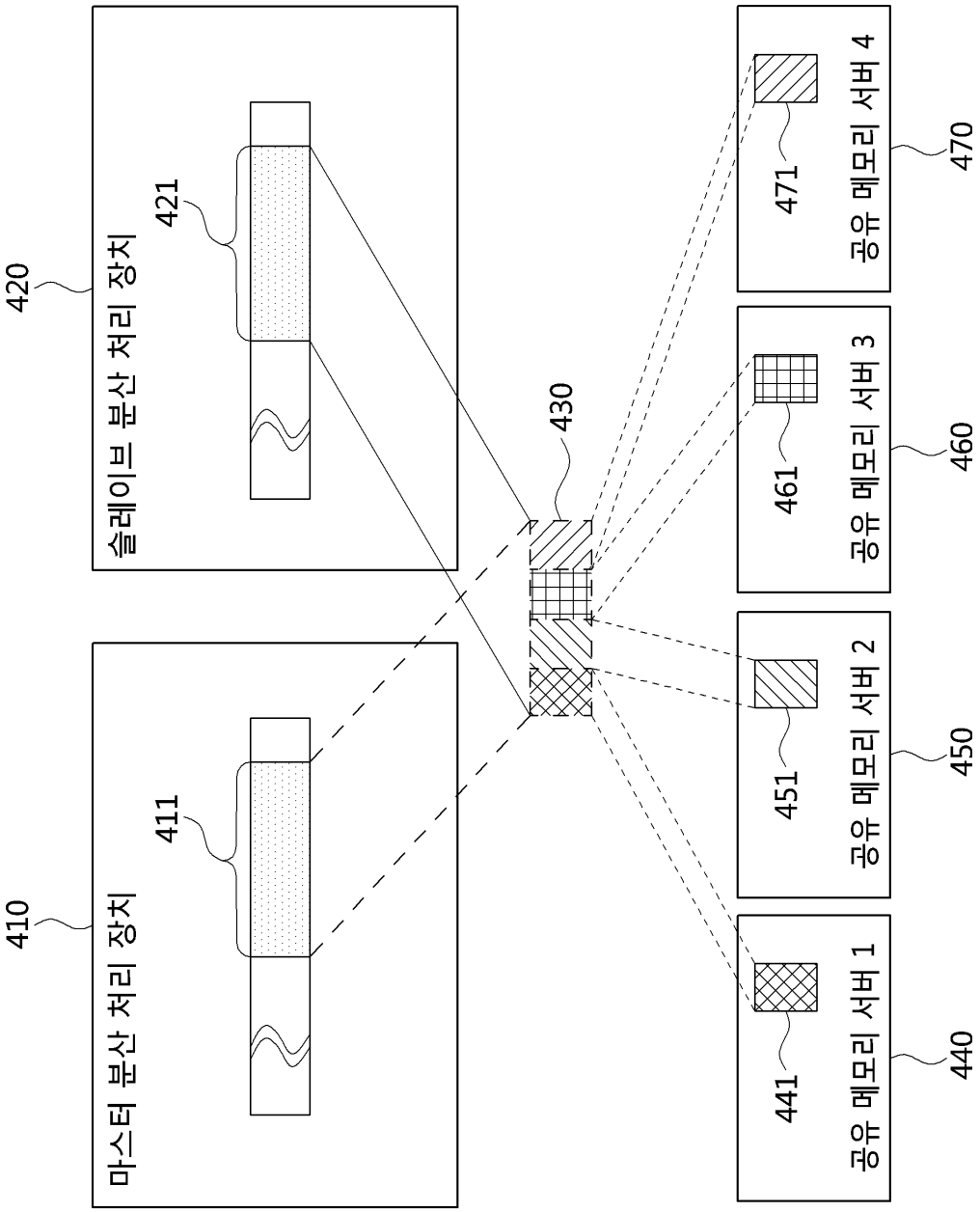
도면2


110
제어부
210
통신부
220
230
메모리
연산 처리부
240
250
공유 메모리 서버
접근 지원부
메모리 맵핑
테이블 관리부
260

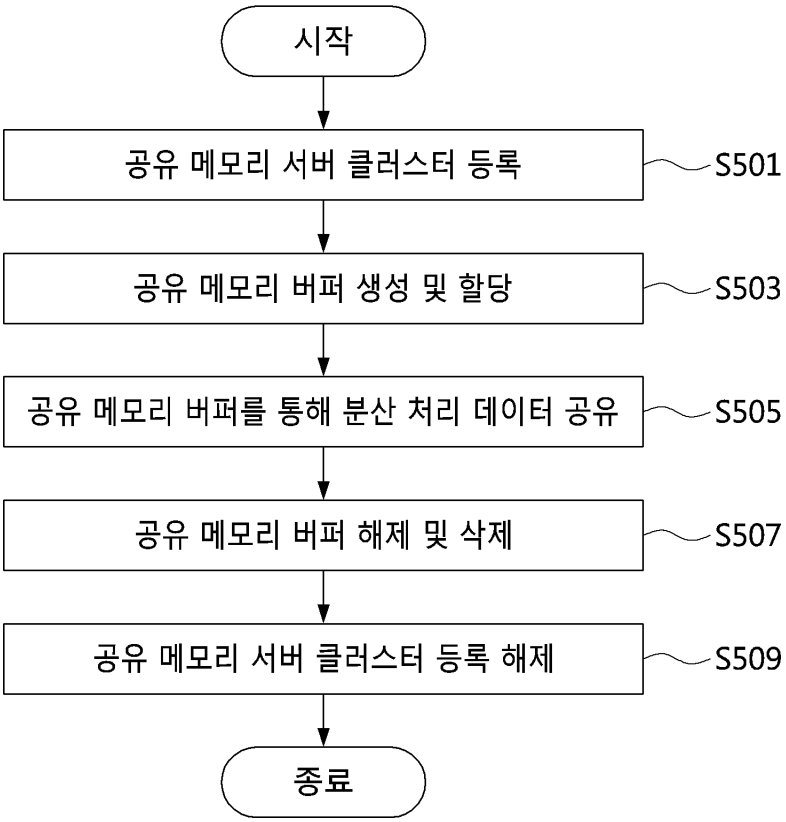
도면3


120
제어부
310
통신부
320
330
메모리
공유 메모리
관리부
340

도면4

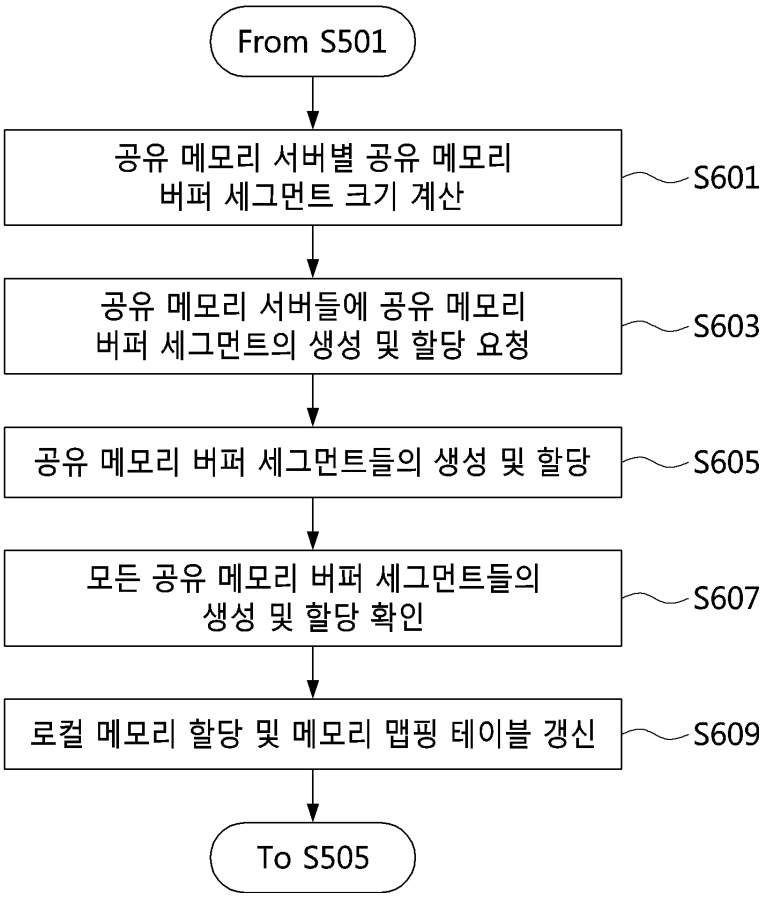

420
슬레이브 분산 처리 장치
421
410
마스터 분산 처리 장치
411
430
470
공유 메모리 서버 4
471
460
공유 메모리 서버 3
461
450
공유 메모리 서버 2
451
440
공유 메모리 서버 1
441

도면5


시작
공유 메모리 서버 클러스터 등록
S501
공유 메모리 버퍼 생성 및 할당
S503
공유 메모리 버퍼를 통해 분산 처리 데이터 공유
S505
공유 메모리 버퍼 해제 및 삭제
S507
공유 메모리 서버 클러스터 등록 해제
S509
종료

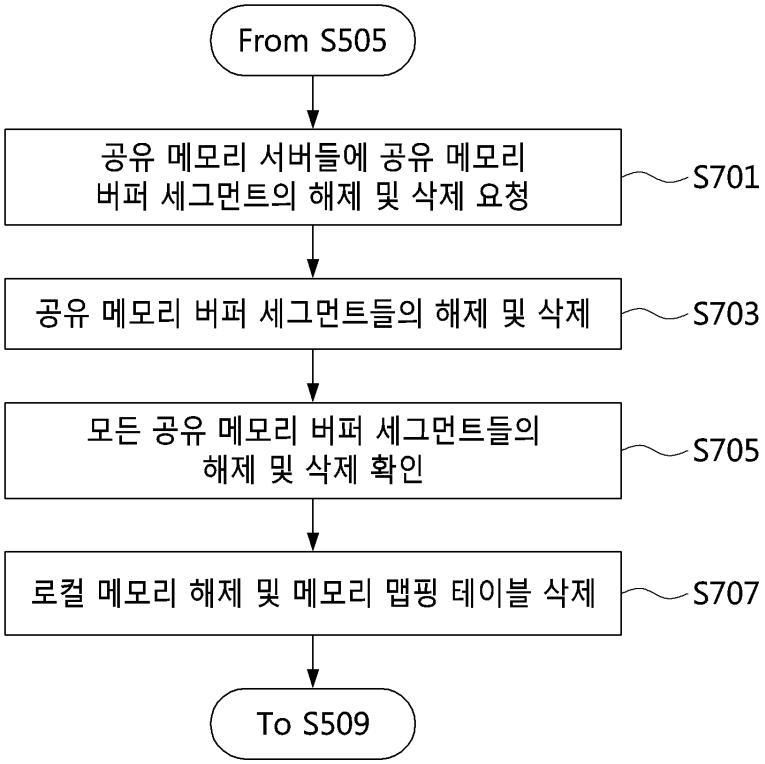
도면6

S503

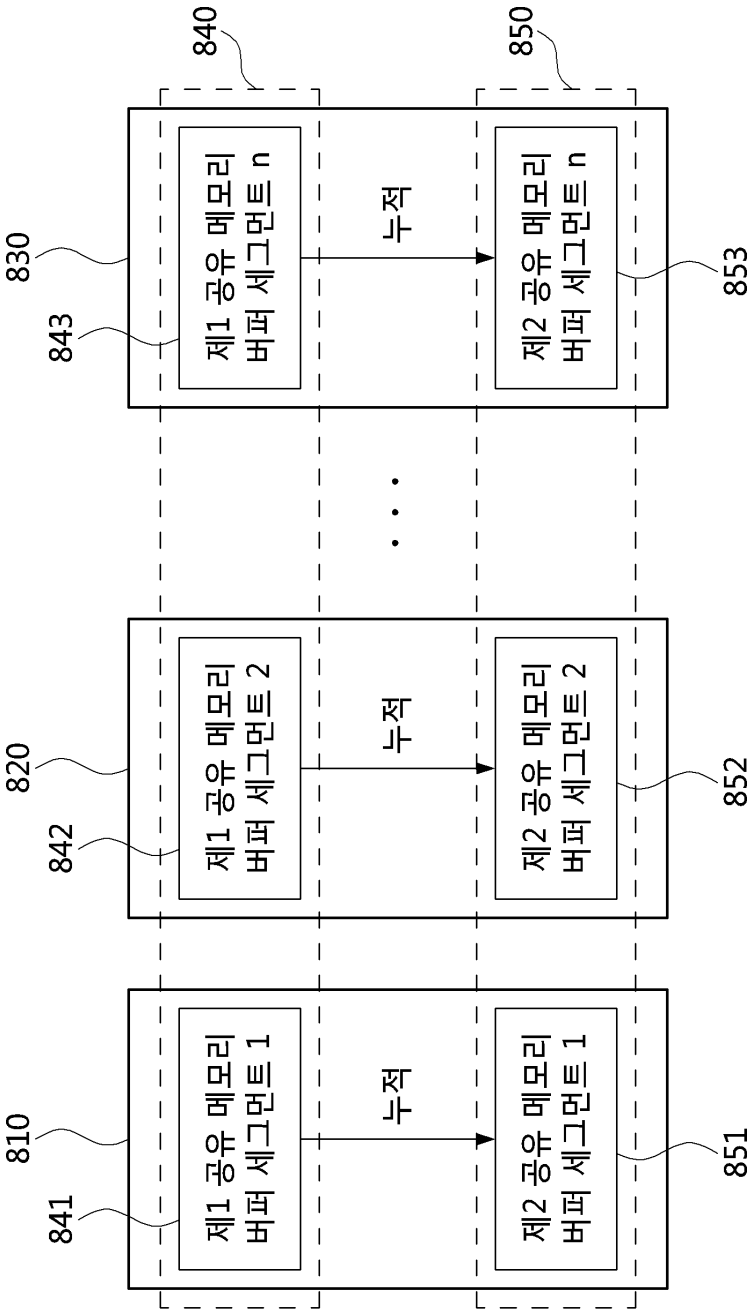

From S501
공유 메모리 서버별 공유 메모리 버퍼 세그먼트 크기 계산
S601
공유 메모리 서버들에 공유 메모리 버퍼 세그먼트의 생성 및 할당 요청
S603
공유 메모리 버퍼 세그먼트들의 생성 및 할당
S605
모든 공유 메모리 버퍼 세그먼트들의 생성 및 할당 확인
S607
로컬 메모리 할당 및 메모리 맵핑 테이블 갱신
S609
To S505

도면7

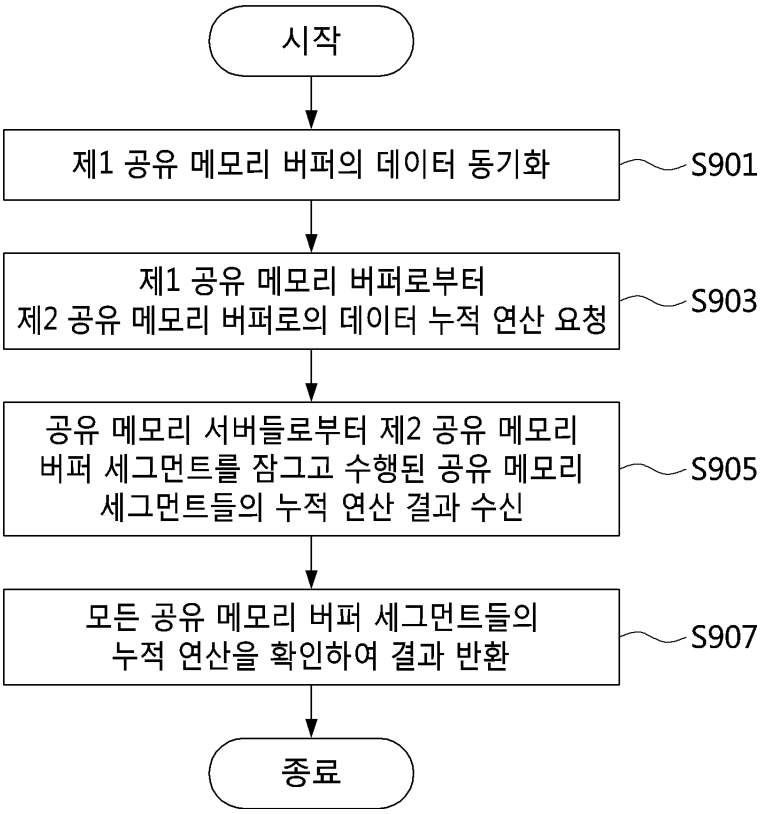
S507


From S505
공유 메모리 서버들에 공유 메모리 버퍼 세그먼트의 해제 및 삭제 요청
S701
공유 메모리 버퍼 세그먼트들의 해제 및 삭제
S703
모든 공유 메모리 버퍼 세그먼트들의 해제 및 삭제 확인
S705
로컬 메모리 해제 및 메모리 맵핑 테이블 삭제
S707
To S509

도면8


840
850
830
843
제1 공유 메모리
버퍼 세그먼트 n
누적
제2 공유 메모리
버퍼 세그먼트 n
853
820
842
제1 공유 메모리
버퍼 세그먼트 2
누적
제2 공유 메모리
버퍼 세그먼트 2
852
810
841
제1 공유 메모리
버퍼 세그먼트 1
누적
제2 공유 메모리
버퍼 세그먼트 1
851

도면9



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 10

【변경전】

청구항 9에 있어서,

상기 공유 메모리 관리부는

상기 분산 처리 장치로부터 상기 공유 메모리 버퍼를 생성하기 위한 공유 메모리 버퍼 세그먼트의 크기 정보와 함께 상기 공유 메모리 버퍼 세그먼트의 생성 및 할당을 요청을 수신하고, 상기 공유 메모리 버퍼 세그먼트를 생성 및 할당하여 상기 공유 메모리 버퍼를 구성하는 것을 특징으로 하는, 공유 메모리 서버.

【변경후】

청구항 8에 있어서,

상기 공유 메모리 관리부는

상기 분산 처리 장치로부터 상기 공유 메모리 버퍼를 생성하기 위한 공유 메모리 버퍼 세그먼트의 크기 정보와 함께 상기 공유 메모리 버퍼 세그먼트의 생성 및 할당을 요청을 수신하고, 상기 공유 메모리 버퍼 세그먼트를 생성 및 할당하여 상기 공유 메모리 버퍼를 구성하는 것을 특징으로 하는, 공유 메모리 서버.

【직권보정 2】

【보정항목】 청구범위

【보정세부항목】 청구항 11

【변경전】

청구항 9에 있어서,

상기 공유 메모리 버퍼는

연산에 의하여 특정 로컬 공유 메모리 영역의 데이터가 변경된 경우에 상기 변경된 로컬 공유 메모리 영역

의 데이터와 동기화되고, 변경된 데이터로 나머지 로컬 공유 메모리 영역들과 동기화되는 것을 특징으로 하는, 공유 메모리 서버.

【변경후】

청구항 8에 있어서,

상기 공유 메모리 버퍼는

연산에 의하여 특정 로컬 공유 메모리 영역의 데이터가 변경된 경우에 상기 변경된 로컬 공유 메모리 영역의 데이터와 동기화되고, 변경된 데이터로 나머지 로컬 공유 메모리 영역들과 동기화되는 것을 특징으로 하는, 공유 메모리 서버.

【직권보정 3】

【보정항목】 청구범위

【보정세부항목】 청구항 12

【변경전】

청구항 9에 있어서,

상기 공유 메모리 관리부는

상기 분산 처리 장치로부터 두 개 이상의 공유 메모리 버퍼들 사이의 데이터 누적 연산 요청을 수신하고, 상기 데이터 누적 연산의 대상이 되는 공유 메모리 버퍼 세그먼트들에 대하여 누적 연산을 수행하고, 결과를 상기 분산 처리 장치에 반환하는 것을 특징으로 하는, 공유 메모리 서버.

【변경후】

청구항 8에 있어서,

상기 공유 메모리 관리부는

상기 분산 처리 장치로부터 두 개 이상의 공유 메모리 버퍼들 사이의 데이터 누적 연산 요청을 수신하고, 상기 데이터 누적 연산의 대상이 되는 공유 메모리 버퍼 세그먼트들에 대하여 누적 연산을 수행하고, 결과를 상기 분산 처리 장치에 반환하는 것을 특징으로 하는, 공유 메모리 서버.

【직권보정 4】

【보정항목】 청구범위

【보정세부항목】 청구항 13

【변경전】

청구항 9에 있어서,

상기 공유 메모리 관리부는

상기 분산 처리 장치가 상기 공유 메모리 버퍼의 사용을 종료하기 위하여 전송한 상기 공유 메모리 버퍼 세그먼트의 해제 및 삭제 요청을 수신하여 상기 공유 메모리 버퍼 세그먼트를 해제 및 삭제하고, 결과를 상기 분산 처리 장치에 반환하여 상기 분산 처리 장치가 상기 로컬 공유 메모리 영역을 해제 및 삭제하고 상기 메모리 맵핑 테이블을 삭제하도록 하는 것을 특징으로 하는, 공유 메모리 서버.

【변경후】

청구항 8에 있어서,

상기 공유 메모리 관리부는

상기 분산 처리 장치가 상기 공유 메모리 버퍼의 사용을 종료하기 위하여 전송한 상기 공유 메모리 버퍼 세그먼트의 해제 및 삭제 요청을 수신하여 상기 공유 메모리 버퍼 세그먼트를 해제 및 삭제하고, 결과를 상기 분산 처리 장치에 반환하여 상기 분산 처리 장치가 상기 로컬 공유 메모리 영역을 해제 및 삭제하고 상기 메모리 맵핑 테이블을 삭제하도록 하는 것을 특징으로 하는, 공유 메모리 서버.

(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년12월14일
(11) 등록번호 10-2190511
(24) 등록일자 2020년12월07일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) *G06F 9/50* (2018.01)
G06N 3/04 (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06F 9/5027 (2013.01)
(21) 출원번호 10-2019-0025730
(22) 출원일자 2019년03월06일
심사청구일자 2019년03월19일
(65) 공개번호 10-2020-0107124
(43) 공개일자 2020년09월16일
(56) 선행기술조사문헌
EP03392825 A2
(뒷면에 계속)
전체 청구항 수 : 총 9 항

(73) 특허권자
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
안신영
대전광역시 서구 둔산북로 160, 5동 701호
박유미
대전광역시 유성구 노은서로250번길 17-3
(뒷면에 계속)
(74) 대리인
한양특허법인

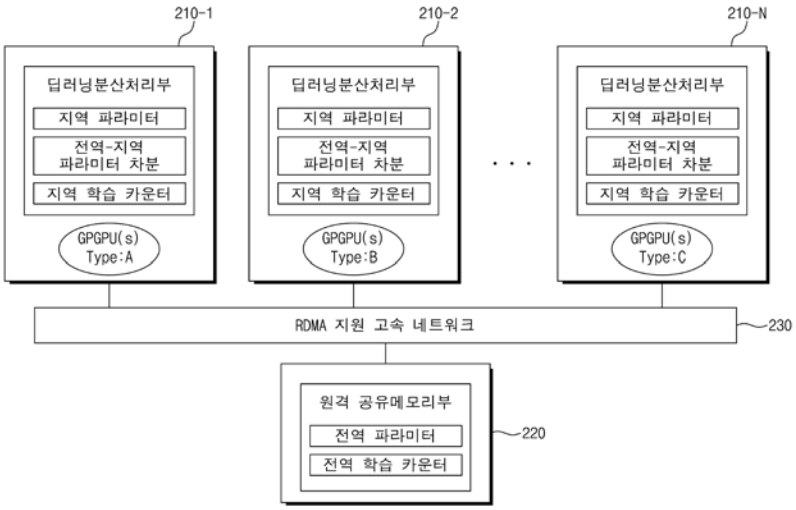
심사관 : 박성수

(54) 발명의 명칭 **이종 클러스터 기반의 분산 딥러닝 방법 및 이를 위한 장치**

(57) 요 약

이종 클러스터 기반의 분산 딥러닝 방법 및 이를 위한 장치가 개시된다. 본 발명의 일실시예에 따른 분산 딥러닝 방법은 딥러닝 성능이 상이한 복수개의 이기종 딥러닝 모듈들이 원격 공유 메모리를 기반으로 전역 파라미터와 전역 학습 카운터를 공유하는 단계; 상기 복수개의 이기종 딥러닝 모듈들이 상기 전역 학습 카운터를 기반으로 할당된 지역 학습 카운터에 상응하는 분산 딥러닝 학습과 원격 공유 메모리 업데이트를 중첩하여 수행하는 단계; 및 상기 원격 공유 메모리 업데이트에 의해 업데이트된 전역 학습 카운터를 고려하여 분산 딥러닝 프로세스를 종료하는 단계를 포함한다.

대 표 도



(52) CPC특허분류
G06N 3/0454 (2013.01)
(72) 발명자
임은지
대전광역시 유성구 노은동로 187, 602동 1801호
최용석
대전광역시 유성구 지족북로 60, 207동 303호

(56) 선행기술조사문헌
KR1020180131836 A
KR1020190087783 A
McMahan, H. B., Moore, E., Ramage, D., Hampson, S., et al. Communication-efficient learning of deep networks from decentralized data. arXiv preprint arXiv:1602.05629, 2016.
Zheng, S et al. Asynchronous stochastic gradient descent with delay compensation. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 4120-4129. JMLR. org, 2017.

이 발명을 지원한 국가연구개발사업

과제고유번호	2016-0-00087
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원(IITP)
연구사업명	정보통신방송기술개발사업(SW컴퓨팅 산업원천기술개발사업)
연구과제명	대규모 딥러닝 고속 처리를 위한 HPC 시스템 개발
기 여 율	1/1
과제수행기관명	한국전자통신연구원
연구기간	2018.01.01 ~ 2018.12.31

명 세 서

청구범위

청구항 1

복수개의 분산 딥러닝 장치들에 의해 각 단계가 수행되는 분산 딥러닝 방법에 있어서,

딥러닝 성능이 상이한 상기 복수개의 분산 딥러닝 장치들이, 원격 공유 메모리를 기반으로 전역 파라미터와 전역 학습 카운터를 공유하는 단계;

상기 복수개의 분산 딥러닝 장치들 각각이, 상기 전역 학습 카운터를 기반으로 할당된 지역 학습 카운터에 상응하도록 상기 복수개의 분산 딥러닝 장치들 각각에 기저장된 지역 파라미터와 상기 전역 파라미터의 차분 연산 결과를 이용하여 분산 딥러닝 학습을 수행하고, 상기 분산 딥러닝 학습이 수행되는 동안 상기 차분 연산 결과를 이용하여 상기 전역 파라미터를 업데이트하는 분산 딥러닝 프로세스를 수행하는 단계; 및

상기 복수개의 분산 딥러닝 장치들이, 상기 분산 딥러닝 학습의 수행 횟수에 상응하는 상기 지역 학습 카운터를 기반으로 상기 전역 학습 카운터를 업데이트하고, 상기 업데이트된 전역 학습 카운터가 기설정된 종료 카운터 이상인지 여부를 판단하여 상기 분산 딥러닝 프로세스를 종료하는 단계

를 포함하는 것을 특징으로 하는 분산 딥러닝 방법.

청구항 2

청구항 1에 있어서,

상기 수행하는 단계는

상기 복수개의 분산 딥러닝 장치들 각각이, 차분 연산 결과를 이용하여 상기 지역 파라미터를 업데이트 하고, 상기 업데이트된 지역 파라미터를 이용하여 상기 분산 딥러닝 학습을 수행하고, 상기 분산 딥러닝 학습을 수행한 결과를 이용하여 상기 업데이트된 지역 파라미터를 재업데이트하는 것을 특징으로 하는 분산 딥러닝 방법.

청구항 3

청구항 1에 있어서,

상기 분산 딥러닝 방법은

상기 복수개의 분산 딥러닝 장치들에 각각 상기 지역 파라미터, 상기 차분 연산 결과 및 상기 지역 학습 카운터의 영역을 생성하는 단계;

상기 원격 공유 메모리에 전역 파라미터 영역 및 전역 학습 카운터 영역을 생성하는 단계; 및

상기 복수개의 분산 딥러닝 장치들 중 어느 하나가 상기 전역 파라미터 및 상기 전역 학습 카운터를 초기화하는 단계를 더 포함하는 것을 특징으로 하는 분산 딥러닝 방법.

청구항 4

청구항 1에 있어서,

상기 수행하는 단계는

상기 복수개의 분산 딥러닝 장치들이 각각 상기 분산 딥러닝 학습을 위한 딥러닝 학습 스레드(THREAD) 및 상기 원격 공유 메모리 업데이트를 위한 업데이트 스레드(THREAD)를 생성하는 단계를 포함하는 것을 특징으로 하는 분산 딥러닝 방법.

청구항 5

삭제

청구항 6

청구항 4에 있어서,

상기 수행하는 단계는

상기 딥러닝 학습 스레드 및 상기 업데이트 스레드 중 어느 하나로 할당되는 흐름제어락을 기반으로 하는 것을 특징으로 하는 분산 딥러닝 방법.

청구항 7

청구항 6에 있어서,

상기 흐름제어락은 상기 분산 딥러닝 프로세스가 시작된 이후에 상기 딥러닝 학습 스레드로 먼저 할당되는 것을 특징으로 하는 분산 딥러닝 방법.

청구항 8

청구항 1에 있어서,

상기 복수개의 분산 딥러닝 장치들은 원격 직접 메모리 접근(REMOTE DIRECT MEMORY ACCESS, RDMA)을 지원하는 고속 네트워크를 기반으로 상기 원격 공유 메모리에 접근하는 것을 특징으로 하는 분산 딥러닝 방법.

청구항 9

원격 공유 메모리의 전역 학습 카운터를 기반으로 할당된 지역 학습 카운터에 상응하도록 기저장된 지역 파라미터와 전역 파라미터의 차분 연산 결과를 이용하여 분산 딥러닝 학습을 수행하고,

상기 분산 딥러닝 학습이 수행되는 동안 상기 차분 연산 결과를 이용하여 상기 전역 파라미터를 업데이트하는 분산 딥러닝 프로세스를 수행하고,

상기 분산 딥러닝 학습의 수행 횟수에 상응하는 상기 지역 학습 카운터를 기반으로 상기 전역 학습 카운터를 업데이트하고, 상기 업데이트된 전역 학습 카운터가 기설정된 종료 카운터 이상인지 여부를 판단하여 상기 분산 딥러닝 프로세스를 종료하는 프로세서; 및

상기 지역 파라미터, 상기 전역 파라미터와 상기 지역 파라미터의 차분 연산 결과 및 상기 지역 학습 카운터를 저장하는 메모리

를 포함하는 것을 특징으로 하는 분산 딥러닝 장치.

청구항 10

청구항 9에 있어서,

상기 프로세서는

상기 차분 연산 결과를 이용하여 상기 지역 파라미터를 업데이트 하고, 상기 업데이트된 지역 파라미터를 이용하여 상기 분산 딥러닝 학습을 수행하고, 상기 분산 딥러닝 학습을 수행한 결과를 이용하여 상기 업데이트된 지역 파라미터를 재업데이트하는 것을 특징으로 하는 분산 딥러닝 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 분산 딥러닝 기술에 관한 것으로, 특히 이기종의 컴퓨팅 모듈로 구성되는 이종 HPC 클러스터 환경에서 효율적으로 분산 딥러닝을 수행할 수 있는 기술에 관한 것이다.

배 경 기 술

[0002] 딥러닝이란 사람의 신경세포(BIOLOGICAL NEURON)를 모사하여 기계가 학습하도록 하는 인공신경망(ARTIFICIAL NEURAL NETWORK) 기반의 기계 학습법이다. 최근 딥러닝 모델들은 응용의 인식 성능을 높이기 위해 대규모 모델로 진화하고 있으나 점차 대형화되는 딥러닝 모델과 대규모 학습 데이터를 단일 머신에서 처리하기에는 한계가 있다. 그래서 대규모 분산 컴퓨팅 자원을 활용하려는 노력의 일환으로 딥러닝 분산 플랫폼 기술이 개발되고 있

다.

[0003] 기존의 딥러닝 분산 처리는 대부분 동일한 규격과 성능의 클러스터를 가정하는 경우가 많다. 그러나 실제로 딥러닝 분산 처리를 하려고 할 때, 동일한 규격과 성능의 컴퓨팅 서버들로 구성된 클러스터를 구비하는 경우는 많지 않다. 따라서 이종 클러스터 환경에서 다른 규격의 서버들로 구성된 이종 클러스터를 동시에 모두 이용하여 효율적으로 딥러닝 분산처리를 수행하는 것은 쉬운 일이 아니다.

[0004] 일반적으로 동종 컴퓨터로 구성된 클러스터 환경에서는 동기식 파라미터 업데이트 방식을 이용한다. 그러나 동종 클러스터 환경에서 동시에 실행되는 분산 프로세스들도 시간이 지남에 따라 다양한 원인으로 인해 속도 차가 발생하기 때문에 동기식 트레이닝의 효율을 떨어뜨리게 된다. 이에 대한 대안으로 사용되는 것이 비동기식 파라미터 업데이트 방식이다. 비동기식 업데이트 방식은 파라미터 서버가 분산 컴퓨터들로부터 늦거나 빨리 도착하는 파라미터들의 동기를 맞추지 않고 트레이닝을 진행하는 방법이다. 비동기 방식은 동기식에 비해 정확성을 크게 희생시키지 않으면서 빠르게 트레이닝 할 수 있는 장점이 있다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 한국 등록 특허 제10-1559089호, 2015년 10월 2일 공개(명칭: 장치의 컴포넌트들 간에 메모리 자원들을 공유하기 위한 통신 프로토콜)

발명의 내용

해결하려는 과제

[0006] 본 발명의 목적은 이기종의 GPU를 이용한 분산 딥러닝 수행 시 통신 오버헤드를 감소시킬 수 있는 효과적인 분산 딥러닝 방법을 제공하는 것이다.

[0007] 또한, 본 발명의 목적은 동시에 성능이 다른 GPU들을 효과적으로 사용할 수 있는 분산 딥러닝 방법을 제공하는 것이다.

[0008] 또한, 본 발명의 목적은 학습 속도가 다른 분산 프로세스들이 전체 학습을 효과적으로 나누어 수행할 수 있도록 하는 것이다.

[0009] 또한, 본 발명의 목적은 학습한 파라미터의 업데이트를 지연하는 방식으로 계산과 통신을 중첩하여 각각의 GPU 활용률을 극대화함으로써 우수한 분산 처리 확장성을 제공하는 것이다.

과제의 해결 수단

[0010] 상기한 목적을 달성하기 위한 본 발명에 따른 분산 딥러닝 방법은 딥러닝 성능이 상이한 복수개의 이기종 딥러닝 모듈들이 원격 공유 메모리를 기반으로 전역 파라미터와 전역 학습 카운터를 공유하는 단계; 상기 복수개의 이기종 딥러닝 모듈들이 상기 전역 학습 카운터를 기반으로 할당된 지역 학습 카운터에 상응하는 분산 딥러닝 학습과 원격 공유 메모리 업데이트를 중첩하여 수행하는 단계; 및 상기 원격 공유 메모리 업데이트에 의해 업데이트된 전역 학습 카운터를 고려하여 분산 딥러닝 프로세스를 종료하는 단계를 포함한다.

[0011] 이 때, 분산 딥러닝 방법은 상기 복수개의 이기종 딥러닝 모듈들에 각각 지역 파라미터, 전역-지역 파라미터 차분 및 지역 학습 카운터 영역을 생성하는 단계; 및 상기 원격 공유 메모리에 전역 파라미터 영역 및 전역 학습 카운터 영역을 생성하는 단계를 더 포함할 수 있다.

[0012] 이 때, 분산 딥러닝 방법은 상기 복수개의 이기종 딥러닝 모듈들 중 어느 하나의 마스터 모듈을 통해 상기 전역 파라미터 및 상기 전역 학습 카운터를 초기화하는 단계를 더 포함할 수 있다.

[0013] 이 때, 수행하는 단계는 상기 복수개의 이기종 딥러닝 모듈들이 각각 상기 분산 딥러닝 학습을 위한 딥러닝 학습 스레드(THREAD) 및 상기 원격 공유 메모리 업데이트를 위한 업데이트 스레드(THREAD)를 생성하는 단계를 포함할 수 있다.

[0014] 이 때, 초기화하는 단계는 상기 업데이트 스레드의 웨이크업 시점을 기준으로 수행될 수 있다.

[0015] 이 때, 수행하는 단계는 상기 딥러닝 학습 스레드 및 상기 업데이트 스레드 중 어느 하나로 할당되는 흐름제어락을 기반으로 할 수 있다.

[0016] 이 때, 흐름제어락은 상기 분산 딥러닝 프로세스가 시작된 이후에 상기 딥러닝 학습 스레드로 먼저 할당될 수 있다.

[0017] 이 때, 복수개의 이기종 딥러닝 모듈들은 원격 직접 메모리 접근(REMOTE DIRECT MEMORY ACCESS, RDMA)을 지원하는 고속 네트워크를 기반으로 상기 원격 공유 메모리에 접근할 수 있다.

[0018] 또한, 본 발명의 일실시예에 따른 분산 딥러닝 장치는, 원격 공유 메모리의 전역 학습 카운터를 기반으로 할당된 지역 학습 카운터에 상응하는 분산 딥러닝 학습과 원격 공유 메모리 업데이트를 중첩하여 수행하고, 상기 원격 공유 메모리 업데이트를 기반으로 업데이트된 전역 학습 카운터를 고려하여 분산 딥러닝 프로세스를 종료하는 프로세서; 및 지역 파라미터, 전역-지역 파라미터 차분 및 지역 학습 카운터를 저장하는 메모리를 포함한다.

[0019] 이 때, 프로세서는 지역 파라미터, 전역-지역 파라미터 차분 및 지역 학습 카운터 영역을 생성하고, 상기 원격 공유 메모리에 전역 파라미터 영역 및 전역 학습 카운터 영역을 생성할 수 있다.

[0020] 이 때, 프로세서는 상기 전역 파라미터 및 상기 전역 학습 카운터를 초기화할 수 있다.

[0021] 이 때, 프로세서는 상기 분산 딥러닝 학습을 위한 딥러닝 학습 스레드(THREAD) 및 상기 원격 공유 메모리 업데이트를 위한 업데이트 스레드(THREAD)를 생성할 수 있다.

[0022] 이 때, 프로세서는 상기 업데이트 스레드의 웨이크업 시점을 기준으로 상기 초기화를 수행할 수 있다.

[0023] 이 때, 프로세서는 상기 딥러닝 학습 스레드 및 상기 업데이트 스레드 중 어느 하나로 흐름제어락을 할당하여 상기 분산 딥러닝 학습 및 상기 원격 공유 메모리 업데이트를 수행할 수 있다.

[0024] 이 때, 흐름제어락은 상기 분산 딥러닝 프로세스가 시작된 이후에 상기 딥러닝 학습 스레드로 먼저 할당될 수 있다.

[0025] 이 때, 프로세서는 원격 직접 메모리 접근(REMOTE DIRECT MEMORY ACCESS, RDMA)을 지원하는 고속 네트워크를 기반으로 상기 원격 공유 메모리에 접근할 수 있다.

발명의 효과

[0026] 본 발명에 따르면, 이기종의 GPU를 이용한 분산 딥러닝 수행 시 통신 오버헤드를 감소시킬 수 있는 효과적인 분산 딥러닝 방법을 제공할 수 있다.

[0027] 또한, 본 발명은 동시에 성능이 다른 GPU들을 효과적으로 사용할 수 있는 분산 딥러닝 방법을 제공할 수 있다.

[0028] 또한, 본 발명은 학습 속도가 다른 분산 프로세스들이 전체 학습을 효과적으로 나누어 수행할 수 있도록 할 수 있다.

[0029] 또한, 본 발명은 학습한 파라미터의 업데이트를 지연하는 방식으로 계산과 통신을 중첩하여 각각의 GPU 활용률을 극대화함으로써 우수한 분산 처리 확장성을 제공할 수 있다.

도면의 간단한 설명

[0030] 도 1은 본 발명의 일실시예에 따른 분산 딥러닝 방법을 나타낸 동작흐름도이다.

도 2는 본 발명의 일실시예에 따른 분산 딥러닝 시스템을 나타낸 도면이다.

도 3은 본 발명의 일실시예에 따른 분산 딥러닝 과정을 상세하게 나타낸 동작흐름도이다.

도 4는 본 발명에 따른 딥러닝 학습 스레드를 기반으로 분산 딥러닝을 수행하는 과정의 일 예를 상세하게 나타낸 동작흐름도이다.

5는 본 발명에 따른 업데이트 스레드를 기반으로 원격 공유 메모리를 업데이트하는 과정의 일 예를 상세하게 나타낸 동작흐름도이다.

도 6은 본 발명의 일실시예에 따른 분산 딥러닝 장치를 나타낸 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0031] 본 발명을 첨부된 도면을 참조하여 상세히 설명하면 다음과 같다. 여기서, 반복되는 설명, 본 발명의 요지를 불필요하게 흐릴 수 있는 공지 기능, 및 구성에 대한 상세한 설명은 생략한다. 본 발명의 실시형태는 당 업계에서 평균적인 지식을 가진 자에게 본 발명을 보다 완전하게 설명하기 위해서 제공되는 것이다. 따라서, 도면에서의 요소들의 형상 및 크기 등은 보다 명확한 설명을 위해 과장될 수 있다.

[0032] 이하, 본 발명에 따른 바람직한 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0034] 도 1은 본 발명의 일실시예에 따른 분산 딥러닝 방법을 나타낸 동작흐름도이다.

[0035] 도 1을 참조하면, 본 발명의 일실시예에 따른 분산 딥러닝 방법은 딥러닝 성능이 상이한 복수개의 이기종 딥러닝 모듈들이 원격 공유 메모리를 기반으로 전역 파라미터와 전역 학습 카운터를 공유한다(S110).

[0036] 이 때, 원격 공유 메모리에 저장된 전역 파라미터와 전역 학습 카운터는 배타적으로 업데이트가 가능한 데이터에 상응하는 것으로, 처리 성능이 서로 상이한 복수개의 이기종 딥러닝 모듈들이 전체 학습을 효과적으로 나누어 수행할 수 있도록 할 수 있다.

[0037] 이 때, 복수개의 이기종 딥러닝 모듈들은 원격 직접 메모리 접근(REMOTE DIRECT MEMORY ACCESS, RDMA)을 지원하는 고속 네트워크를 기반으로 원격 공유 메모리에 접근할 수 있다. 따라서, 원격 공유 메모리는 전역 파라미터와 전역 학습 카운터를 복수개의 이기종 딥러닝 모듈들에게 제공하여 직접 접근할 수 있도록 지원할 수 있다.

[0038] 예를 들어 도 2를 참조하면, 본 발명의 일실시예에 따른 복수개의 이기종 딥러닝 모듈들(210-1~210-N)은 RDMA 고속 네트워크(230)를 통해 원격 공유 메모리(220)에 접근할 수 있다. 이 때, 도 2에 도시된 것처럼 본 발명의 일실시예에 따른 복수개의 이기종 딥러닝 모듈들(210-1~210-N)은 딥러닝 학습을 수행하는 계산노드에 해당할 수 있으며, 상호간에 서로 다른 성능의 GPGPU(GENERAL PURPOSE COMPUTING ON GRAPHICS PROCESSING UNITS))들을 포함할 수 있다.

[0039] 이 때, 도 1에는 도시하지 아니하였으나, 본 발명의 일실시예에 따른 분산 딥러닝 방법은 복수개의 이기종 딥러닝 모듈들이 각각 지역 파라미터, 전역-지역 파라미터 차분 및 지역 학습 카운터 영역을 생성한다. 예를 들어, 도 2에 도시된 것처럼, 본 발명의 일실시예에 따른 복수개의 이기종 딥러닝 모듈들(210-1~210-N)은 각각 지역 파라미터, 전역-지역 파라미터 차분, 지역 학습 카운터를 포함할 수 있다.

[0040] 이 때, 학습 카운터란, 분산 딥러닝 프로세스들이 딥러닝 학습을 수행할 때 학습한 미니배치(MINI-BATCH)의 전체 횟수를 카운팅하는데 사용될 수 있으며, 각각의 분산 딥러닝 프로세스들이 학습해야 할 미니배치의 순서 번호를 할당 받는데 활용될 수 있다. 이와 같이 각각의 딥러닝 모듈로 할당된 학습 카운터는 지역 학습 카운터로써 저장될 수 있다. 이 때, 각각의 딥러닝 모듈에 포함된 분산 프로세스들이 수행해야 할 전체 미니배치의 횟수는 분산 딥러닝 프로세스를 이용하는 사용자가 지정할 수 있다.

[0041] 또한, 본 발명의 일실시예에 따른 분산 딥러닝 방법은 복수개의 이기종 딥러닝 모듈들이 전역 학습 카운터를 기반으로 할당된 지역 학습 카운터에 상응하는 분산 딥러닝 학습과 원격 공유 메모리 업데이트를 중첩하여 수행한다(S120).

[0042] 이 때, 복수개의 이기종 딥러닝 모듈들은 분산 딥러닝 학습을 통해 지역 파라미터를 자체적으로 학습시킬 수 있고, 원격 공유 메모리에 보관되는 전역 학습 카운터는 복수개의 이기종 딥러닝 모듈들 각각에 저장된 지역 학습 카운터와 비교하여 동일하면 변경하는 방식(COMPARE AND SWAP)으로 업데이트될 수 있다.

[0043] 일반적으로 분산 딥러닝 플랫폼에서 분산 딥러닝 학습을 수행하는 프로세스들은 상호간에 대규모 파라미터를 빈번하게 송수신해야 하므로 이 과정에서 발생하는 통신 오버헤드는 전체 분산 딥러닝 학습 성능에서 차지하는 비중이 매우 높은 형편이다. 따라서, 효과적인 분산 딥러닝 학습을 위해서는 통신시간을 감소시키거나 또는 통신시간과 계산시간을 중첩함으로써 통신시간을 숨길 필요가 있다.

[0044] 이와 같은 문제점을 해결하기 위해, 본 발명에서는 학습된 파라미터의 업데이트를 즉각적으로 수행하지 않고 지연하는 방식으로 계산과 통신을 중첩시키는 분산 딥러닝 방법을 제안하고자 한다.

[0045] 이 때, 복수개의 이기종 딥러닝 모듈들이 각각 분산 딥러닝 학습을 위한 딥러닝 학습 스레드(THREAD) 및 원격 공유 메모리 업데이트를 위한 업데이트 스레드(THREAD)를 생성할 수 있다. 일반적으로 복수개의 이기종 딥러닝 모듈들 각각의 메인 스레드가 분산 딥러닝 학습 스레드에 상응할 수 있다.

[0046] 또한, 분산 딥러닝 학습 및 원격 공유 메모리 업데이트는 딥러닝 학습 스레드 및 업데이트 스레드 중 어느 하나

로 할당되는 흐름제어락을 기반으로 수행될 수 있다.

[0047] 이하에서는 도 4 내지 도 5를 기반으로 분산 딥러닝 학습과 원격 공유 메모리 업데이트를 중첩 수행하는 두 개의 스레드들의 세부 절차를 설명하도록 한다.

[0048] 먼저, 도 4에 도시된 것처럼, 딥러닝 학습 스레드는 시작되면(S410) 먼저 흐름제어락을 획득할 수 있다(S420).

[0049] 이 때, 흐름제어락은 딥러닝 학습 스레드와 업데이트 스레드 간의 흐름제어를 위해 사용되는 것으로, 흐름제어락은 분산 딥러닝 프로세스가 시작된 이후에 딥러닝 학습 스레드로 먼저 할당될 수 있다.

[0050] 이 후, 전역-지역 파라미터 차분으로부터 지역 파라미터를 업데이트할 수 있다(S430). 이 후, 딥러닝 학습 스레드로 할당되었던 흐름제어락을 해제하고, 업데이트 스레드로 웨이크업 신호를 보내 깨워줄 수 있다(S440). 즉, 분산 딥러닝 학습 스레드는 분산 딥러닝 학습을 시작하기 전에 업데이트 스레드를 깨워준다.

[0051] 이 후, 업데이트된 지역 파라미터를 이용하여 하나의 미니배치 데이터에 대한 분산 딥러닝 학습 수행한 뒤(S450), 학습 결과를 기반으로 지역 파라미터 업데이트할 수 있다(S460).

[0052] 이 후, 원격 공유 메모리에 저장된 전역 학습 카운터 고려하여 추가적인 분산 딥러닝 프로세스가 필요한 경우에는 딥러닝 학습 스레드 반복 수행하되, 전역 학습 카운터가 만료되어 사용자가 지정한 미니배치에 도달하였을 경우에는 딥러닝 분산 프로세스를 종료할 수 있다(S470).

[0053] 또한, 도 5를 참조하면, 업데이트 스레드는 생성된 이후에 딥러닝 학습 스레드 또는 메인 스레드가 깨워줄 때까지 대기할 수 있다(S510).

[0054] 따라서, 웨이크업 신호가 발생하는지 여부를 판단하고(S515), 웨이크업 신호가 발생하여 업데이트 스레드가 깨어나면, 종료변수가 참인지 여부를 확인할 수 있다(S525).

[0055] 단계(S525)의 판단결과 종료변수가 참이면, 업데이트 스레드 종료할 수 있다(S570).

[0056] 또한, 단계(S525)의 판단결과 종료변수가 참이 아니면, 업데이트 스레드로 흐름제어락을 할당할 수 있다(S530).

[0057] 이 후, 원격 공유 메모리에 저장되어 있는 전역 파라미터를 딥러닝 모듈의 지역 버퍼로 읽어와서 전역 파라미터와 지역 파라미터의 차분을 계산할 수 있다(S540).

[0058] 이 후, 단계(S540)을 통해 산출된 전력-지역 파라미터 차분을 이용하여 원격 공유 메모리의 전역 파라미터가 증가하도록 업데이트한 뒤(S550) 업데이트 스레드로 할당된 흐름제어락을 해제할 수 있다(S560).

[0059] 이 때, 단계(S530) 내지 단계(S560)의 절차는 분산 딥러닝 학습이 완료될 때까지 반복적으로 수행될 수 있으며, 대체로 분산 딥러닝 학습시간이 원격 공유 메모리 업데이트 시간보다 길기 때문에 업데이트 스레드는 전역 파라미터 업데이트 완료 후에 대기상태로 회귀할 수 있다.

[0060] 이와 같이, 본 발명에서는 복수개의 이기종 딥러닝 모듈들 각각에서 N번째로 학습한 파라미터가 N+1번째 학습 도중에 전역 파라미터로 업데이트될 수 있고, N번째 학습 도중에 읽어온 전역 파라미터를 N+1번째 학습 전에 지역 파라미터로써 업데이트하여 학습을 수행할 수 있다. 따라서, 종래에 파라미터 서버를 이용하는 비동기 방식처럼 새로운 전역 파라미터가 업데이트될 때까지 대기할 필요가 없으므로 통신에 의해 지체되는 시간을 절약할 수 있다. 즉, 본 발명에서는 딥러닝 학습 스레드와 업데이트 스레드가 흐름제어락과 대기/웨이크업 방식을 이용하여 학습과 통신(전역 파라미터 업데이트)을 중첩 수행할 수 있다.

[0061] 또한, 본 발명의 일실시예에 따른 분산 딥러닝 방법은 원격 공유 메모리 업데이트에 의해 업데이트된 전역 학습 카운터를 고려하여 분산 딥러닝 프로세스를 종료한다(S130).

[0062] 예를 들어, 도 2에 도시된 것과 같은 본 발명의 딥러닝 모듈(210-1~210-N)은 원격 공유 메모리에 저장된 전역 학습 카운터를 배타적으로 증가시키면서 분산 딥러닝 학습을 수행하므로, 전역 학습 카운터가 사용자가 지정한 값에 도달하였을 때에 분산 딥러닝 프로세스를 종료할 수 있다.

[0063] 이와 같이 함으로써 저속 GPU는 전체 미니배치 횟수 중에서 더 적은 횟수를 학습하고, 고속 GPU는 더 많은 미니배치를 학습할 수 있으므로 분산 딥러닝 프로세스의 사용자가 지정한 미니배치에 도달하였을 때에 이기종의 분산 딥러닝 프로세스들은 거의 동시에 분산 딥러닝 학습을 종료할 수 있다.

[0064] 또한, 도 1에는 도시하지 아니하였으나, 본 발명의 일실시예에 따른 분산 딥러닝 방법은 원격 공유 메모리에 전역 파라미터 영역 및 전역 학습 카운터 영역을 생성한다.

[0065] 예를 들어, 도 2에 도시된 것과 같은 본 발명의 일실시예에 따른 복수개의 이기종 딥러닝 모듈들(210-1~210-N)은 RDMA 고속 네트워크(230)를 기반으로 원격 공유 메모리(220)로 접근하여 전역 파라미터 영역과 전역 학습 카운터 영역을 생성할 수 있다. 이 때, 복수개의 이기종 딥러닝 모듈들(210-1~210-N) 중 어느 하나의 마스터 모듈을 설정하고, 설정된 마스터 모듈을 이용하여 전역 파라미터 영역과 전역 학습 카운터 영역을 생성할 수도 있다.

[0066] 또한, 도 1에는 도시하지 아니하였으나, 본 발명의 일실시예에 따른 분산 딥러닝 방법은 복수개의 이기종 딥러닝 모듈들 중 어느 하나의 마스터 모듈을 통해 전역 파라미터 및 전역 학습 카운터를 초기화한다.

[0067] 이 때, 전역 파라미터는 딥러닝 모듈별 또는 딥러닝 플랫폼별로 다양한 방식으로 초기화될 수 있으며, 전역 학습 카운터는 0으로 초기화될 수 있다.

[0068] 또한, 이와 같은 초기화 과정은 업데이트 스레드가 최초로 웨이크업되는 시점을 기준으로 수행될 수도 있다.

[0069] 또한, 도 1에는 도시하지 아니하였으나, 본 발명의 일실시예에 따른 분산 딥러닝 방법은 상술한 바와 같이 본 발명의 실시예에 따른 분산 딥러닝 과정에서 발생되는 다양한 정보를 저장할 수 있다.

[0070] 이와 같은 이종 클러스터 기반의 분산 딥러닝 방법을 통해 이기종의 GPU를 이용한 분산 딥러닝 수행 시 통신 오버헤드를 감소시킬 수 있다.

[0071] 또한, 동시에 성능이 다른 GPU들을 효과적으로 사용할 수 있는 분산 딥러닝 방법을 제공할 수 있다.

[0072] 또한, 학습 속도가 다른 분산 프로세스들이 전체 학습을 효과적으로 나누어 수행할 수 있도록 할 수 있다.

[0073] 또한, 학습한 파라미터의 업데이트를 지연하는 방식으로 계산과 통신을 중첩하여 각각의 GPU 활용률을 극대화함으로써 우수한 분산 처리 확장성을 제공할 수 있다.

[0075] 도 3은 본 발명의 일실시예에 따른 분산 딥러닝 과정을 상세하게 나타낸 동작흐름도이다.

[0076] 도 3을 참조하면, 본 발명의 일실시예에 따른 분산 딥러닝 과정은 먼저 복수개의 이기종 딥러닝 모듈들에 각각 지역 파라미터, 전역-지역 파라미터 차분, 지역 학습 카운터 영역을 생성한다(S310).

[0077] 이 후, 복수개의 이기종 딥러닝 모듈들을 통해 원격 공유 메모리에 전역 파라미터 영역, 전역 학습 카운터 영역을 생성한다(S320).

[0078] 이 후, 복수개의 이기종 딥러닝 모듈들 중 어느 하나의 마스터 모듈을 통해 전역 파라미터와 전역 학습 카운터를 초기화한다(S330).

[0079] 이 때, 전역 파라미터는 딥러닝 모듈별 또는 딥러닝 플랫폼별로 다양한 방식으로 초기화될 수 있으며, 전역 학습 카운터는 0으로 초기화될 수 있다.

[0080] 이 후, 복수개의 이기종 딥러닝 모듈들은 각각 분산 딥러닝 학습을 위한 딥러닝 학습 스레드와 원격 공유 메모리 업데이트를 위한 업데이트 학습 스레드를 생성한다(S340).

[0081] 이 후, 복수개의 이기종 딥러닝 모듈들은 딥러닝 학습 스레드를 통해 분산 딥러닝 학습을 수행하기 이전에 웨이크업 신호를 발생시켜 업데이트 스레드를 깨운다(S350).

[0082] 이 후, 복수개의 이기종 딥러닝 모듈들은 각각 분산 딥러닝 학습과 원격 공유 메모리의 업데이트를 중첩 수행한다(S360).

[0083] 이 후, 원격 공유 메모리에 업데이트되는 전역 학습 카운터가 목표 종료 카운터 이상인지 여부를 판단하고(S365), 전역 학습 카운터가 목표 종료 카운터 이상이면 분산 딥러닝 프로세스를 종료한다(S370).

[0084] 또한, 단계(S365)의 판단결과 전역 학습 카운터가 목표 종료 카운터 미만이면 지속적으로 분산 딥러닝 학습을 수행할 수 있도록 단계(S360)부터 반복 수행할 수 있다.

[0086] 도 6은 본 발명의 일실시예에 따른 분산 딥러닝 장치를 나타낸 블록도이다.

[0087] 도 6을 참조하면, 본 발명의 일실시예에 따른 분산 딥러닝 장치는 프로세서(610) 및 메모리(620)를 포함한다.

[0088] 프로세서(610)는 원격 공유 메모리를 기반으로 전역 파라미터와 전역 학습 카운터를 공유한다.

[0089] 이 때, 원격 공유 메모리에 저장된 전역 파라미터와 전역 학습 카운터는 배타적으로 업데이트가 가능한 데이터

에 상응하는 것으로, 처리 성능이 서로 상이한 복수개의 이기종 분산 딥러닝 장치들이 전체 학습을 효과적으로 나누어 수행할 수 있도록 할 수 있다.

[0090] 이 때, 원격 직접 메모리 접근(REMOTE DIRECT MEMORY ACCESS, RDMA)을 지원하는 고속 네트워크를 기반으로 원격 공유 메모리에 접근할 수 있다. 따라서, 원격 공유 메모리는 프로세서(610)가 전역 파라미터와 전역 학습 카운터에 직접 접근할 수 있도록 지원할 수 있다.

[0091] 또한, 프로세서(610)는 지역 파라미터, 전역-지역 파라미터 차분 및 지역 학습 카운터 영역을 생성한다. 예를 들어, 본 발명의 일실시예에 따른 복수개의 이기종 딥러닝 장치들은 각각 지역 파라미터, 전역-지역 파라미터 차분, 지역 학습 카운터를 포함할 수 있다.

[0092] 이 때, 학습 카운터란, 분산 딥러닝 프로세스들이 분산 딥러닝 학습을 수행할 때 학습한 미니배치(MINI-BATCH)의 전체 횟수를 카운팅하는데 사용될 수 있으며, 각각의 분산 딥러닝 프로세스들이 학습해야 할 미니배치의 순서 번호를 할당 받는데 활용될 수 있다. 이와 같이 각각의 분산 딥러닝 장치로 할당된 학습 카운터는 지역 학습 카운터로써 저장될 수 있다. 이 때, 각각의 분산 딥러닝 장치에 포함된 분산 프로세스들이 수행해야 할 전체 미니배치의 횟수는 분산 딥러닝 프로세스를 이용하는 사용자가 지정할 수 있다.

[0093] 또한, 프로세서(610)는 원격 공유 메모리의 전역 학습 카운터를 기반으로 할당된 지역 학습 카운터에 상응하는 분산 딥러닝 학습과 원격 공유 메모리 업데이트를 중첩하여 수행한다.

[0094] 이 때, 분산 딥러닝 학습을 통해 지역 파라미터를 자체적으로 학습시킬 수 있고, 원격 공유 메모리에 보관되는 전역 학습 카운터는 메모리(620)에 저장된 지역 학습 카운터와 비교하여 동일하면 변경하는 방식(COMPARE AND SWAP)으로 업데이트될 수 있다.

[0095] 일반적으로 분산 딥러닝 플랫폼에서 분산 딥러닝 학습을 수행하는 프로세스들은 상호간에 대규모 파라미터를 빈번하게 송수신해야 하므로 이 과정에서 발생하는 통신 오버헤드는 전체 분산 딥러닝 학습 성능에서 차지하는 비중이 매우 높은 형편이다. 따라서, 효과적인 분산 딥러닝 학습을 위해서는 통신시간을 감소시키거나 또는 통신시간과 계산시간을 중첩함으로써 통신시간을 숨길 필요가 있다.

[0096] 이와 같은 문제점을 해결하기 위해, 본 발명에서는 학습된 파라미터의 업데이트를 즉각적으로 수행하지 않고 지연하는 방식으로 계산과 통신을 중첩시키는 분산 딥러닝 방법을 제안하고자 한다.

[0097] 이 때, 프로세서(610)는 분산 딥러닝 학습을 위한 딥러닝 학습 스레드(THREAD) 및 원격 공유 메모리 업데이트를 위한 업데이트 스레드(THREAD)를 생성할 수 있다. 일반적으로 메인 스레드가 분산 딥러닝 학습 스레드에 상응할 수 있다.

[0098] 또한, 분산 딥러닝 학습 및 원격 공유 메모리 업데이트는 딥러닝 학습 스레드 및 업데이트 스레드 중 어느 하나로 할당되는 흐름제어락을 기반으로 수행될 수 있다.

[0099] 이하에서는 도 4 내지 도 5를 기반으로 분산 딥러닝 학습과 원격 공유 메모리 업데이트를 중첩 수행하는 두 개의 스레드들의 세부 절차를 설명하도록 한다.

[0100] 먼저, 도 4에 도시된 것처럼, 딥러닝 학습 스레드는 시작되면(S410) 먼저 흐름제어락을 획득할 수 있다(S420).

[0101] 이 때, 흐름제어락은 딥러닝 학습 스레드와 업데이트 스레드 간의 흐름제어를 위해 사용되는 것으로, 흐름제어락은 분산 딥러닝 프로세스가 시작된 이후에 딥러닝 학습 스레드로 먼저 할당될 수 있다.

[0102] 이 후, 전역-지역 파라미터 차분으로부터 지역 파라미터를 업데이트할 수 있다(S430). 이 후, 딥러닝 학습 스레드로 할당되었던 흐름제어락을 해제하고, 업데이트 스레드로 웨이크업 신호를 보내 깨워줄 수 있다(S440). 즉, 분산 딥러닝 학습 스레드는 분산 딥러닝 학습을 시작하기 전에 업데이트 스레드를 깨워준다.

[0103] 이 후, 업데이트된 지역 파라미터를 이용하여 하나의 미니배치 데이터에 대한 분산 딥러닝 학습 수행한 뒤(S450), 학습 결과를 기반으로 지역 파라미터 업데이트할 수 있다(S460).

[0104] 이 후, 원격 공유 메모리에 저장된 전역 학습 카운터 고려하여 추가적인 분산 딥러닝 프로세스가 필요한 경우에는 딥러닝 학습 스레드 반복 수행하되, 전역 학습 카운터가 만료되어 사용자가 지정한 미니배치에 도달하였을 경우에는 딥러닝 분산 프로세스를 종료할 수 있다(S470).

[0105] 또한, 도 5를 참조하면, 업데이트 스레드는 생성된 이후에 딥러닝 학습 스레드 또는 메인 스레드가 깨워줄 때까지 대기할 수 있다(S510).

[0106] 따라서, 웨이크업 신호가 발생하는지 여부를 판단하고(S515), 웨이크업 신호가 발생하여 업데이트 스레드가 깨어나면, 종료변수가 참인지 여부를 확인할 수 있다(S525).

[0107] 단계(S525)의 판단결과 종료변수가 참이면, 업데이트 스레드 종료할 수 있다(S570).

[0108] 또한, 단계(S525)의 판단결과 종료변수가 참이 아니면, 업데이트 스레드로 흐름제어락을 할당할 수 있다(S530).

[0109] 이 후, 원격 공유 메모리에 저장되어 있는 전역 파라미터를 딥러닝 모듈의 지역 버퍼로 읽어와서 전역 파라미터와 지역 파라미터의 차분을 계산할 수 있다(S540).

[0110] 이 후, 단계(S540)을 통해 산출된 전력-지역 파라미터 차분을 이용하여 원격 공유 메모리의 전역 파라미터가 증가하도록 업데이트한 뒤(S550) 업데이트 스레드로 할당된 흐름제어락을 해제할 수 있다(S560).

[0111] 이 때, 단계(S530) 내지 단계(S560)의 절차는 분산 딥러닝 학습이 완료될 때까지 반복적으로 수행될 수 있으며, 대체로 분산 딥러닝 학습시간이 원격 공유 메모리 업데이트 시간보다 길기 때문에 업데이트 스레드는 전역 파라미터 업데이트 완료 후에 대기상태로 회귀할 수 있다.

[0112] 이와 같이, 본 발명에서는 분산 딥러닝 장치가 N번째로 학습한 파라미터가 N+1번째 학습 도중에 전역 파라미터로 업데이트될 수 있고, N번째 학습 도중에 읽어온 전역 파라미터를 N+1번째 학습 전에 지역 파라미터로써 업데이트하여 학습을 수행할 수 있다. 따라서, 종래에 파라미터 서버를 이용하는 비동기 방식처럼 새로운 전역 파라미터가 업데이트될 때까지 대기할 필요가 없으므로 통신에 의해 지체되는 시간을 절약할 수 있다. 즉, 본 발명에서는 딥러닝 학습 스레드와 업데이트 스레드가 흐름제어락과 대기/웨이크업 방식을 이용하여 학습과 통신(전역 파라미터 업데이트)을 중첩 수행할 수 있다.

[0113] 또한, 프로세서(610)는 원격 공유 메모리 업데이트를 기반으로 업데이트된 전역 학습 카운터를 고려하여 분산 딥러닝 프로세스를 종료한다.

[0114] 예를 들어, 프로세서(610)는 원격 공유 메모리에 저장된 전역 학습 카운터를 배타적으로 증가시키면서 분산 딥러닝 학습을 수행하므로, 전역 학습 카운터가 사용자가 지정한 값에 도달하였을 때에 분산 딥러닝 프로세스를 종료할 수 있다.

[0115] 이와 같이 함으로써 저속 GPU는 전체 미니배치 횟수 중에서 더 적은 횟수를 학습하고, 고속 GPU는 더 많은 미니배치를 학습할 수 있으므로 분산 딥러닝 프로세스의 사용자가 지정한 미니배치에 도달하였을 때에 복수개의 분산 딥러닝 장치들은 거의 동시에 분산 딥러닝 학습을 종료할 수 있다.

[0116] 또한, 프로세서(610)는 원격 공유 메모리에 전역 파라미터 영역 및 전역 학습 카운터 영역을 생성한다.

[0117] 예를 들어, 프로세서(610)는 RDMA 기반의 고속 네트워크를 기반으로 원격 공유 메모리로 접근하여 전역 파라미터 영역과 전역 학습 카운터 영역을 생성할 수 있다.

[0118] 또한, 프로세서(610)는 전역 파라미터 및 전역 학습 카운터를 초기화한다.

[0119] 이 때, 전역 파라미터는 분산 딥러닝 장치 별로 다양한 방식으로 초기화될 수 있으며, 전역 학습 카운터는 0으로 초기화될 수 있다.

[0120] 또한, 이와 같은 초기화 과정은 업데이트 스레드가 최초로 웨이크업되는 시점을 기준으로 수행될 수도 있다.

[0121] 메모리(620)는 지역 파라미터, 전역-지역 파라미터 차분 및 지역 학습 카운터를 저장한다.

[0122] 또한, 메모리(620)는 상술한 바와 같이 본 발명의 실시예에 따른 이종 클러스터 기반의 분산 딥러닝 과정에서 발생되는 다양한 정보를 저장한다.

[0123] 실시예에 따라, 메모리(620)는 분산 딥러닝 장치와 독립적으로 구성되어 분산 딥러닝 수행을 위한 기능을 지원할 수 있다. 이 때, 메모리(620)는 별도의 대용량 스토리지로 동작할 수 있고, 동작 수행을 위한 제어 지능을 포함할 수 있다.

[0124] 한편, 분산 딥러닝 장치는 메모리가 탑재되어 그 장치 내에서 정보를 저장할 수 있다. 일 구현예의 경우, 메모리는 컴퓨터로 판독 가능한 매체이다. 일 구현 예에서, 메모리는 휘발성 메모리 유닛일 수 있으며, 다른 구현예의 경우, 메모리는 비휘발성 메모리 유닛일 수도 있다. 일 구현예의 경우, 저장장치는 컴퓨터로 판독 가능한 매체이다. 다양한 서로 다른 구현 예에서, 저장장치는 예컨대 하드디스크 장치, 광학디스크 장치, 혹은 어떤 다른 대용량 저장장치를 포함할 수도 있다.

[0125] 이와 같은 분산 딥러닝 장치를 이용함으로써 이기종의 GPU를 이용한 분산 딥러닝 수행 시 통신 오버헤드를 감소시킬 수 있다.

[0126] 또한, 동시에 성능이 다른 GPU들을 효과적으로 사용할 수 있는 분산 딥러닝 방법을 제공할 수 있다.

[0127] 또한, 학습 속도가 다른 분산 프로세스들이 전체 학습을 효과적으로 나누어 수행할 수 있도록 할 수 있다.

[0128] 또한, 학습한 파라미터의 업데이트를 지연하는 방식으로 계산과 통신을 중첩하여 각각의 GPU 활용률을 극대화함으로써 우수한 분산 처리 확장성을 제공할 수 있다.

[0130] 이상에서와 같이 본 발명에 따른 이종 클러스터 기반의 분산 딥러닝 방법 및 이를 위한 장치는 상기한 바와 같이 설명된 실시예들의 구성과 방법이 한정되게 적용될 수 있는 것이 아니라, 상기 실시예들은 다양한 변형이 이루어질 수 있도록 각 실시예들의 전부 또는 일부가 선택적으로 조합되어 구성될 수도 있다.

부호의 설명

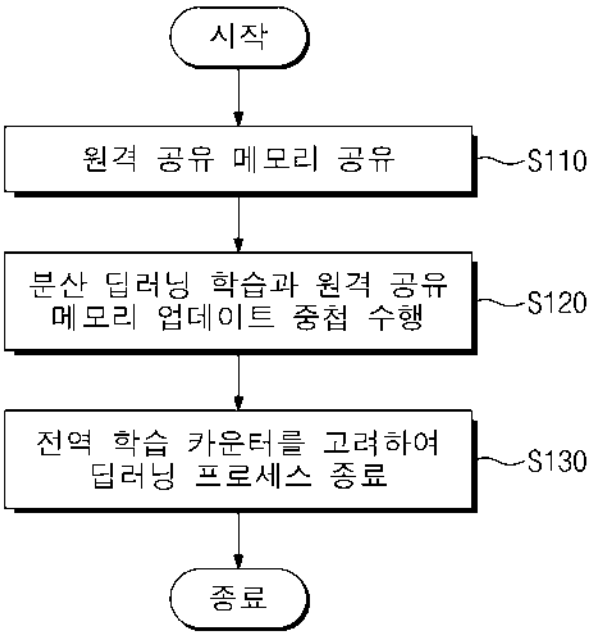
[0131] 210-1~210-N: 딥러닝 모듈 220: 원격 공유 메모리

230: RDMA 고속 네트워크 610: 프로세서

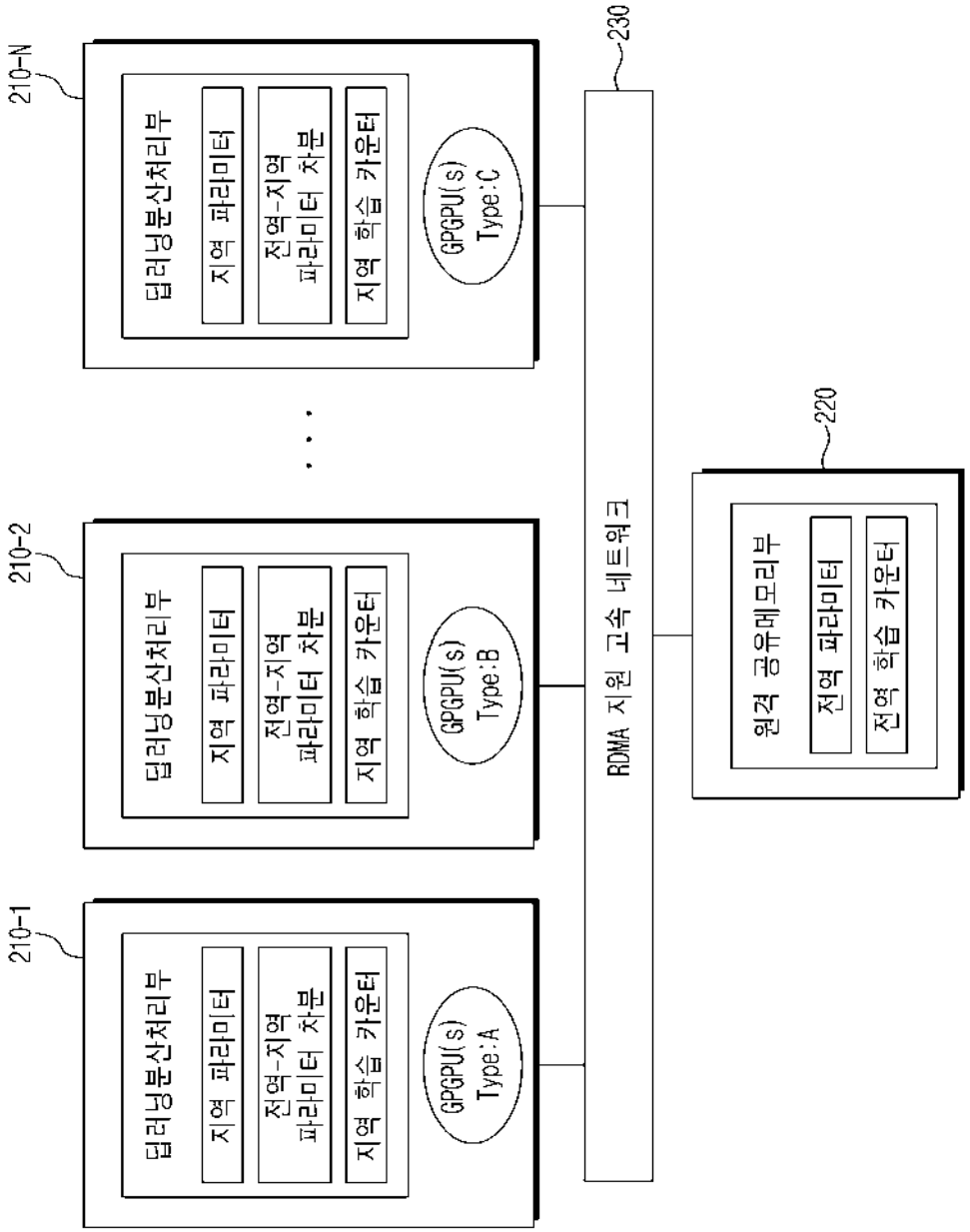
620: 메모리

도면

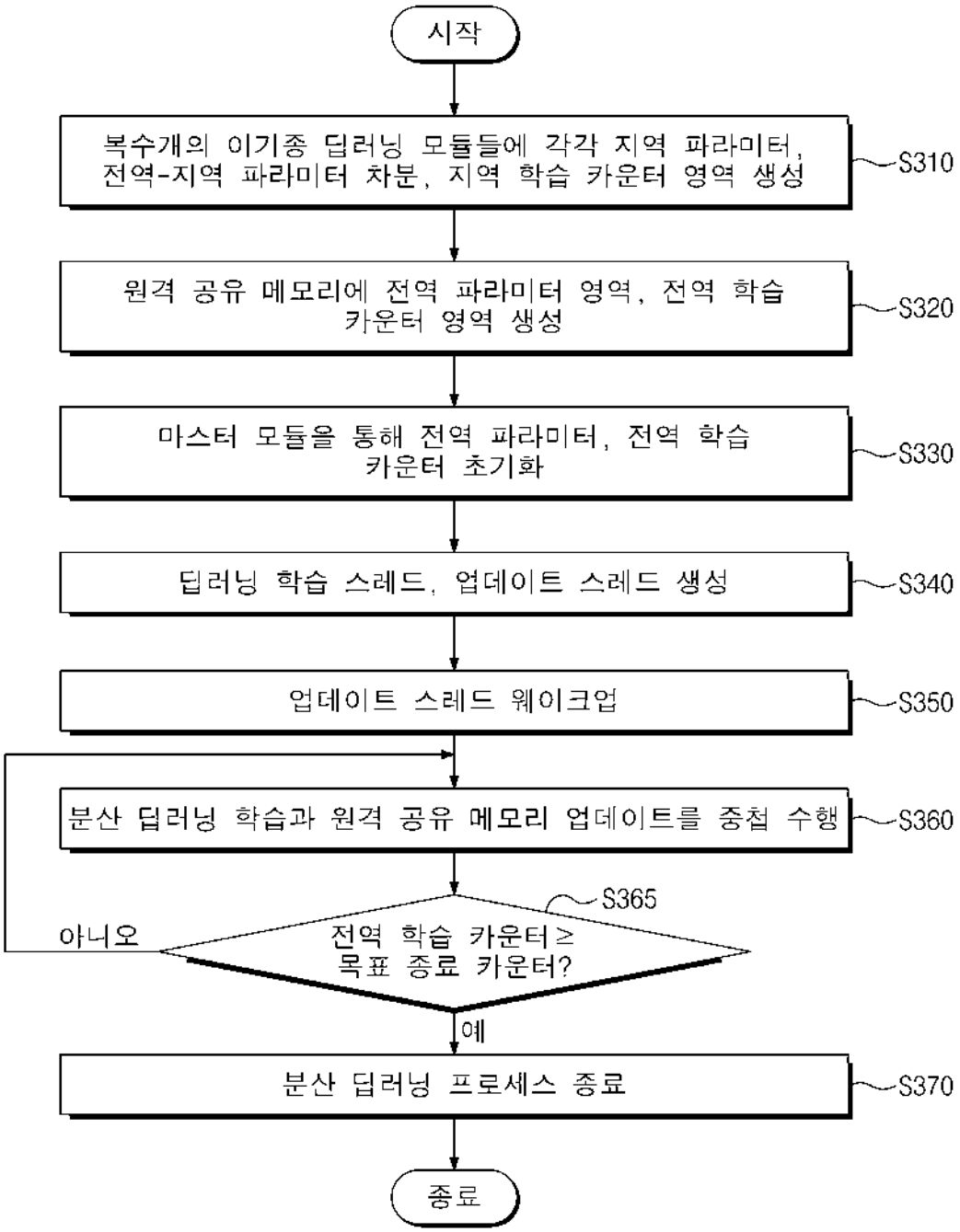
도면1



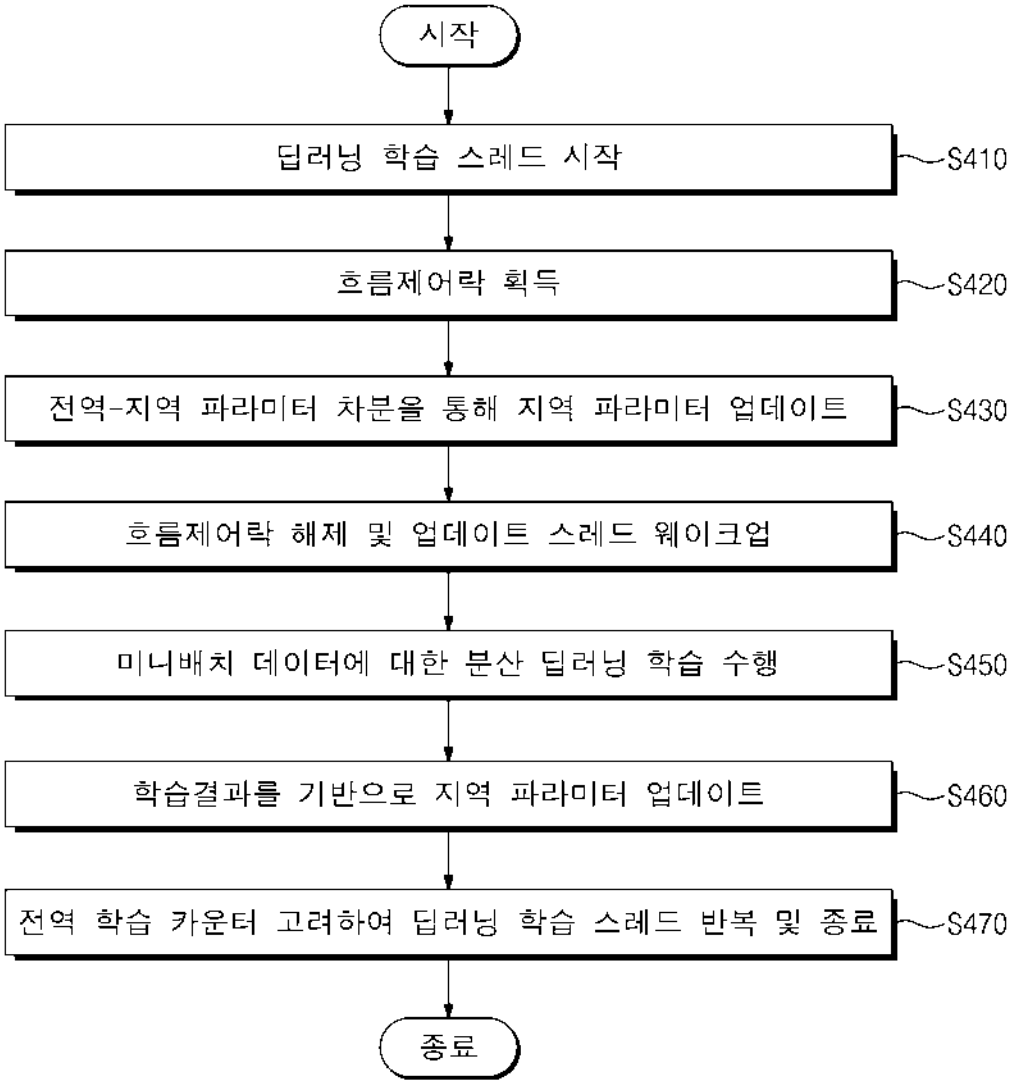
도면2


210-N
딥러닝분산처리부
지역 파라미터
전역-지역 파라미터 차분
지역 학습 카운터
GPGPU(s) Type:C
210-2
딥러닝분산처리부
지역 파라미터
전역-지역 파라미터 차분
지역 학습 카운터
GPGPU(s) Type:B
210-1
딥러닝분산처리부
지역 파라미터
전역-지역 파라미터 차분
지역 학습 카운터
GPGPU(s) Type:A
230
RDMA 지원 고속 네트워크
220
원격 공유메모리부
전역 파라미터
전역 학습 카운터

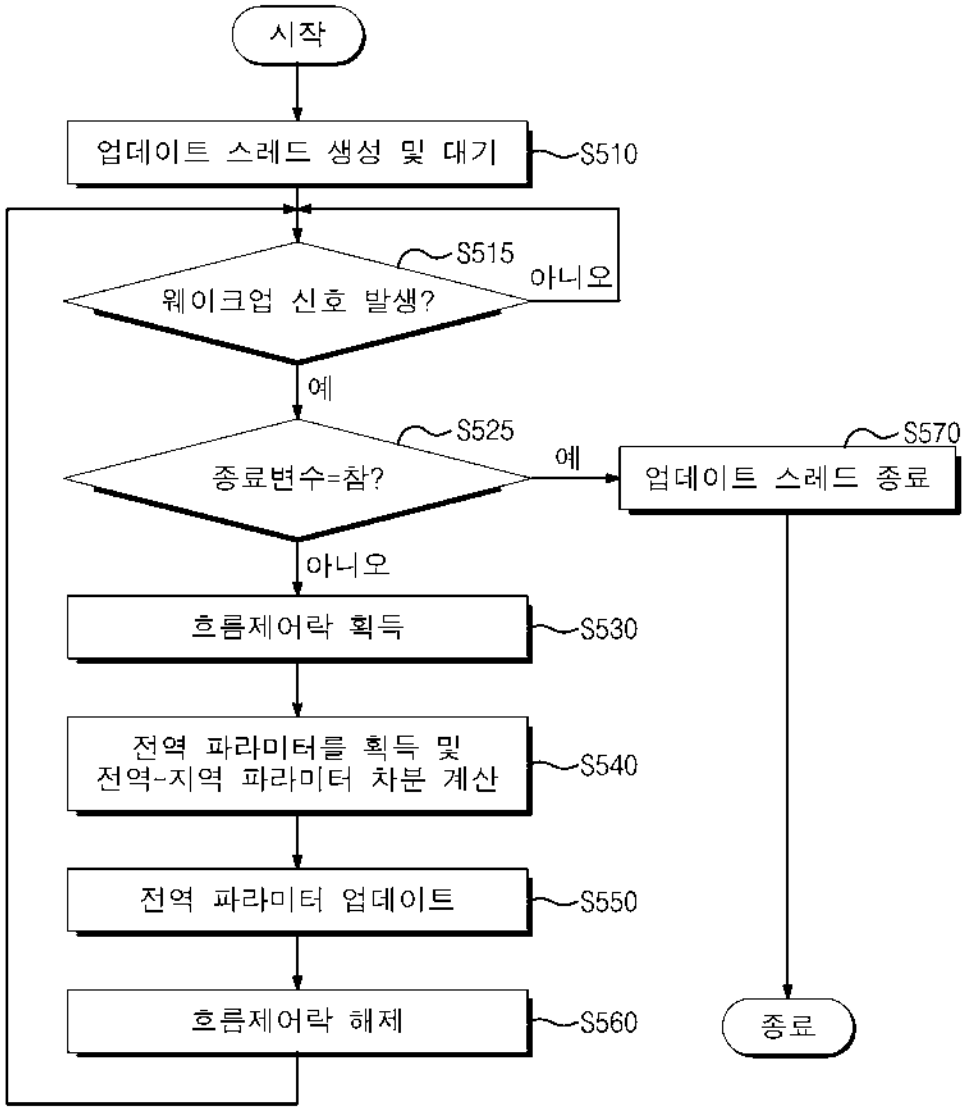
도면3


시작
복수개의 이기종 딥러닝 모듈들에 각각 지역 파라미터, 전역-지역 파라미터 차분, 지역 학습 카운터 영역 생성
S310
원격 공유 메모리에 전역 파라미터 영역, 전역 학습 카운터 영역 생성
S320
마스터 모듈을 통해 전역 파라미터, 전역 학습 카운터 초기화
S330
딥러닝 학습 스레드, 업데이트 스레드 생성
S340
업데이트 스레드 웨이크업
S350
분산 딥러닝 학습과 원격 공유 메모리 업데이트를 중첩 수행
S360
S365
전역 학습 카운터≥ 목표 종료 카운터?
아니오
예
분산 딥러닝 프로세스 종료
S370
종료

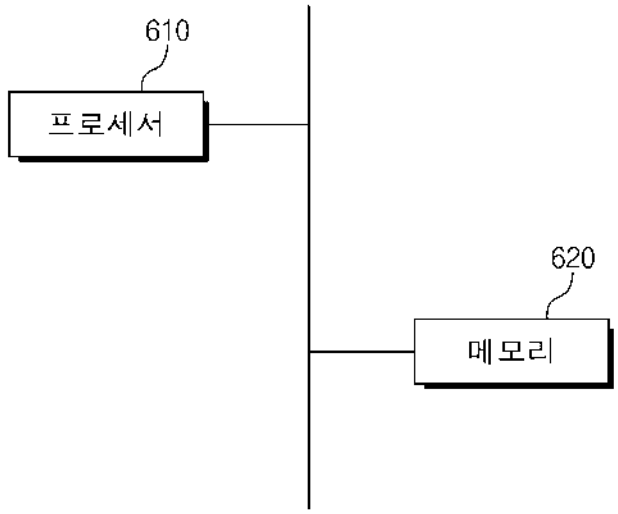
도면4


시작
딥러닝 학습 스레드 시작
S410
흐름제어락 획득
S420
전역-지역 파라미터 차분을 통해 지역 파라미터 업데이트
S430
흐름제어락 해제 및 업데이트 스레드 웨이크업
S440
미니배치 데이터에 대한 분산 딥러닝 학습 수행
S450
학습결과를 기반으로 지역 파라미터 업데이트
S460
전역 학습 카운터 고려하여 딥러닝 학습 스레드 반복 및 종료
S470
종료

도면5


시작
업데이트 스레드 생성 및 대기
S510
S515
웨이크업 신호 발생?
아니오
예
S525
종료변수=참?
예
S570
업데이트 스레드 종료
아니오
흐름제어락 획득
S530
전역 파라미터를 획득 및
전역-지역 파라미터 차분 계산
S540
전역 파라미터 업데이트
S550
흐름제어락 해제
S560
종료

도면6



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 9

【변경전】

원격 공유 메모리의 전역 학습 카운터를 기반으로 할당된 지역 학습 카운터에 상응하도록 기저장된 지역 파라미터와 상기 전역 파라미터의 차분 연산 결과를 이용하여 분산 딥러닝 학습을 수행하고,

상기 분산 딥러닝 학습이 수행되는 동안 상기 차분 연산 결과를 이용하여 상기 전역 파라미터를 업데이트하는 분산 딥러닝 프로세스를 수행하고,

상기 분산 딥러닝 학습의 수행 횟수에 상응하는 상기 지역 학습 카운터를 기반으로 상기 전역 학습 카운터를 업데이트하고, 상기 업데이트된 전역 학습 카운터가 기설정된 종료 카운터 이상인지 여부를 판단하여 상기 분산 딥러닝 프로세스를 종료하는 프로세서; 및

상기 지역 파라미터, 상기 전역 파라미터와 상기 지역 파라미터의 차분 연산 결과 및 상기 지역 학습 카운터를 저장하는 메모리

를 포함하는 것을 특징으로 하는 분산 딥러닝 장치.

【변경후】

원격 공유 메모리의 전역 학습 카운터를 기반으로 할당된 지역 학습 카운터에 상응하도록 기저장된 지역 파라미터와 전역 파라미터의 차분 연산 결과를 이용하여 분산 딥러닝 학습을 수행하고,

상기 분산 딥러닝 학습이 수행되는 동안 상기 차분 연산 결과를 이용하여 상기 전역 파라미터를 업데이트하는 분산 딥러닝 프로세스를 수행하고,

상기 분산 딥러닝 학습의 수행 횟수에 상응하는 상기 지역 학습 카운터를 기반으로 상기 전역 학습 카운터를 업데이트하고, 상기 업데이트된 전역 학습 카운터가 기설정된 종료 카운터 이상인지 여부를 판단하여 상기 분산 딥러닝 프로세스를 종료하는 프로세서; 및

상기 지역 파라미터, 상기 전역 파라미터와 상기 지역 파라미터의 차분 연산 결과 및 상기 지역 학습 카운터를 저장하는 메모리

를 포함하는 것을 특징으로 하는 분산 딥러닝 장치.

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2018-0125734
(43) 공개일자 2018년11월26일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2006.01) *G06F 12/02* (2018.01)
G06N 3/08 (2006.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
G06F 12/0292 (2013.01)
(21) 출원번호 10-2017-0060400
(22) 출원일자 2017년05월16일
심사청구일자 없음

(71) 출원인
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
임은지
대전광역시 유성구 노은동로 187, 602동 1801호
안신영
대전광역시 서구 둔산북로 160, 5동 701호
(뒷면에 계속)
(74) 대리인
한양특허법인

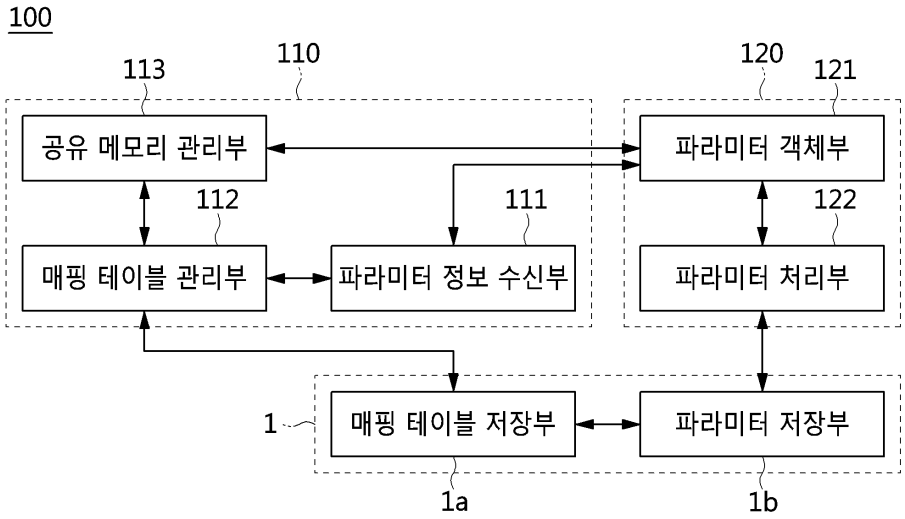
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **파라미터 공유 장치 및 방법**

(57) 요 약

파라미터 공유 장치 및 방법이 개시된다. 본 발명의 일실시예에 따른 파라미터 공유 장치는 메모리 박스에 파라미터가 저장될 메모리 영역의 할당 관리를 수행하고, 상기 메모리 영역의 할당 관리에 따라 상기 메모리 박스에 저장된 매핑 테이블을 업데이트 하는 메모리 할당부 및 상기 파라미터가 저장될 메모리 영역의 할당 관리를 위한 파라미터 정보를 상기 메모리 할당부에 제공하고, 상기 메모리 박스에 저장된 파라미터를 공유하는 연산 처리부를 포함한다.

대 표 도 - 도5



(52) CPC특허분류
G06N 3/08 (2013.01)
(72) 발명자
최용석
전광역시 유성구 지족북로 60, 207동 303호
우영춘
대전광역시 유성구 어은로 57, 113동 404호
최완
대전광역시 서구 관저북로 52, 108동 306호

이 발명을 지원한 국가연구개발사업
과제고유번호 R7117-16-0235
부처명 미래창조과학부
연구관리전문기관 정보통신기술진흥센터(IITP)
연구사업명 정보통신방송기술개발사업(SW컴퓨팅 산업원천기술개발사업)
연구과제명 대규모 딥러닝 고속 처리를 위한 HPC 시스템 개발
기 여 율 1/1
주관기관 한국전자통신연구원
연구기간 2016.04.01 ~ 2016.12.31

명 세 서

청구범위

청구항 1

파라미터 공유 장치를 이용하는 방법에 있어서,

메모리 박스에 저장될 파라미터의 메모리 영역을 할당하기 위한 파라미터 정보를 수신하는 단계;

상기 메모리 박스의 매핑 테이블에 잠금을 걸고, 매핑 테이블을 읽어오는 단계;

상기 파라미터 정보에 기반하여 상기 매핑 테이블에서 상기 메모리 박스에 파라미터를 저장할 메모리 영역의 할당 여부를 확인하는 단계;

상기 메모리 영역의 할당 여부에 따라 매핑 정보를 수정한 매핑 테이블을 상기 메모리 박스에 쓰고, 상기 매핑 테이블의 잠금을 해제하는 단계; 및

상기 메모리 영역을 할당한 메모리 주소를 고려하여 상기 파라미터를 공유하는 단계;

를 포함하는 것을 특징으로 하는 파라미터 공유 방법.

청구항 2

청구항 1에 있어서,

상기 수신하는 단계는

상기 파라미터의 파라미터 식별자 및 상기 파라미터 크기 중 적어도 하나를 포함하는 파라미터 정보를 수신하는 것을 특징으로 하는 파라미터 공유 방법.

청구항 3

청구항 2에 있어서,

상기 매핑 테이블은

각각의 엔트리에 파라미터의 식별자, 메모리 영역에 대한 메모리 주소 및 참조 카운트를 포함하는 매핑 정보가 기록된 것을 특징으로 하는 파라미터 공유 방법.

청구항 4

청구항 3에 있어서,

상기 할당 여부를 확인하는 단계는

상기 매핑 테이블의 엔트리를 확인하여 상기 메모리 박스에 상기 파라미터의 메모리 영역이 기 할당되어 있는 경우,

상기 매핑 테이블에 상기 파라미터에 상응하는 엔트리에 참조 카운트를 증가시켜 상기 매핑 테이블을 업데이트하는 것을 특징으로 하는 파라미터 공유 방법.

청구항 5

청구항 4에 있어서,

상기 할당 여부를 확인하는 단계는

상기 매핑 테이블의 엔트리를 확인하여 상기 메모리 박스에 상기 파라미터의 메모리 영역이 할당되지 않은 경우,

상기 파라미터 크기만큼 상기 메모리 박스에 메모리 영역을 할당하고, 상기 메모리 영역이 할당된 파라미터에 대한 매핑 정보를 상기 매핑 테이블의 새로운 엔트리에 추가하여 상기 매핑 테이블을 업데이트하는 것을 특징으

로 하는 파라미터 공유 방법.

청구항 6

청구항 5에 있어서,

상기 매핑 테이블의 잠금을 해제하는 단계는

상기 메모리 영역이 할당된 파라미터에 대한 상기 메모리 박스의 메모리 주소를 상기 파라미터 공유 장치에 기록하는 것을 특징으로 하는 파라미터 공유 방법.

청구항 7

청구항 6에 있어서,

상기 공유하는 단계는

상기 파라미터 공유 장치에 기록된 메모리 주소를 참조하여 상기 메모리 박스에 저장된 파라미터 값을 읽어오는(read) 단계;

모델 알고리즘을 이용하여 상기 메모리 박스의 파라미터 값에 상응하는 파라미터 차분 값을 계산하는 단계; 및

상기 파라미터 차분 값을 이용하여 상기 메모리 박스의 파라미터 값을 수정하는 단계;

를 포함하는 것을 특징으로 하는 파라미터 공유 방법.

청구항 8

청구항 7에 있어서,

상기 파라미터 값을 수정하는 단계는

상기 메모리 박스가 합저장 기능 수행이 가능한 경우,

상기 파라미터 차분 값을 이용하여 상기 메모리 박스의 합저장 기능을 통해 상기 메모리 박스의 파라미터 값을 수정하는 것을 특징으로 하는 파라미터 공유 방법.

청구항 9

청구항 7에 있어서,

상기 파라미터 값을 수정하는 단계는

상기 메모리 박스가 합저장 기능 수행이 불가능한 경우,

상기 메모리 박스로부터 파라미터 값을 다시 읽어오고(read), 상기 파라미터 차분 값과 다시 읽어온 파라미터 값을 이용하여 산출한 파라미터 수정 값을 상기 메모리 박스에 쓰는(write) 것을 특징으로 하는 파라미터 공유 방법.

청구항 10

파라미터 공유 장치를 이용하는 방법에 있어서,

메모리 박스에 파라미터가 저장된 메모리 영역을 해제하기 위한 파라미터 정보를 수신하는 단계;

상기 메모리 박스의 매핑 테이블에 잠금을 걸고, 매핑 테이블을 읽어오는 단계;

상기 매핑 테이블에 기반하여 상기 메모리 박스에 상기 파라미터가 할당된 메모리 영역의 해제 여부를 확인하는 단계;

상기 메모리 영역의 해제 여부에 따라 매핑 정보를 수정한 매핑 테이블을 상기 메모리 박스에 쓰고, 상기 매핑 테이블의 잠금을 해제하는 단계; 및

상기 메모리 영역을 해제한 메모리 주소를 고려하여 상기 파라미터를 공유하는 단계;

를 포함하는 것을 특징으로 하는 파라미터 공유 방법.

청구항 11

청구항 10에 있어서,

상기 수신하는 단계는

상기 파라미터의 파라미터 식별자 및 상기 파라미터가 저장된 메모리 영역에 대한 메모리 주소 중 적어도 하나를 포함하는 파라미터 정보를 수신하는 것을 특징으로 하는 파라미터 공유 방법.

청구항 12

청구항 11에 있어서,

상기 읽어오는 단계는

상기 매핑 테이블에 상기 파라미터에 상응하는 엔트리에 참조 카운트를 감소시켜 상기 매핑 테이블을 업데이트하는 단계; 및

상기 파라미터 공유 장치에 기록된 상기 파라미터에 상응하는 메모리 주소를 삭제하는 단계를 포함하는 파라미터 공유 방법.

청구항 13

청구항 12에 있어서,

상기 해제 여부를 확인하는 단계는

상기 매핑 테이블의 참조 카운트가 최소값인 경우, 상기 메모리 박스에 할당된 메모리 영역을 해제하고, 상기 메모리 영역에 상응하는 매핑 테이블의 엔트리를 삭제하는 것을 특징으로 하는 파라미터 공유 방법.

청구항 14

청구항 13에 있어서,

상기 공유하는 단계는

상기 파라미터 공유 장치에 기록된 메모리 주소를 참조하여 상기 메모리 박스에 저장된 파라미터 값을 읽어오는(read) 단계;

모델 알고리즘을 이용하여 상기 메모리 박스의 파라미터 값에 상응하는 파라미터 차분 값을 계산하는 단계; 및

상기 파라미터 차분 값을 이용하여 상기 메모리 박스의 파라미터 값을 수정하는 단계;

를 포함하는 것을 특징으로 하는 파라미터 공유 방법.

청구항 15

청구항 14에 있어서,

상기 파라미터 값을 수정하는 단계는

상기 메모리 박스가 합저장 기능 수행이 가능한 경우,

상기 파라미터 차분 값을 이용하여 상기 메모리 박스의 합저장 기능을 통해 상기 메모리 박스의 파라미터 값을 수정하는 것을 특징으로 하는 파라미터 공유 방법.

청구항 16

청구항 14에 있어서,

상기 파라미터 값을 수정하는 단계는

상기 메모리 박스가 합저장 기능 수행이 불가능한 경우,

상기 메모리 박스로부터 파라미터 값을 다시 읽어오고(read), 상기 파라미터 차분 값과 다시 읽어온 파라미터 값을 이용하여 산출한 파라미터 수정 값을 상기 메모리 박스에 쓰는(write) 것을 특징으로 하는 파라미터 공유

방법.

청구항 17

메모리 박스에 파라미터가 저장될 메모리 영역의 할당 관리를 수행하고, 상기 메모리 영역의 할당 관리에 따라 상기 메모리 박스에 저장된 매핑 테이블을 업데이트 하는 메모리 할당부; 및

상기 파라미터가 저장될 메모리 영역의 할당 관리를 위한 파라미터 정보를 상기 메모리 할당부에 제공하고, 상기 메모리 박스에 저장된 파라미터를 공유하는 연산 처리부;

를 포함하는 것을 특징으로 하는 파라미터 공유 장치.

청구항 18

청구항 17에 있어서,

상기 메모리 할당부는

상기 연산 처리부로부터 상기 메모리 영역의 할당 관리를 위한 상기 파라미터 정보를 수신하는 파라미터 정보 수신부;

상기 메모리 박스에 저장된 매핑 테이블의 잠금을 관리하고, 상기 매핑 테이블을 업데이트하는 매핑 테이블 관리부; 및

상기 매핑 테이블의 참조 카운트를 수정하여, 상기 메모리 영역의 할당 관리를 수행하는 공유 메모리 관리부;

를 포함하는 것을 특징으로 하는 파라미터 공유 장치.

청구항 19

청구항 18에 있어서,

상기 메모리 할당부는

상기 참조 카운트에 따라 상기 메모리 영역에서 파라미터를 공유 중인 다른 파라미터 공유 장치의 개수를 확인하는 것을 특징으로 하는 파라미터 공유 장치.

청구항 20

청구항 19에 있어서,

상기 메모리 박스는

상기 연산 처리부의 파라미터 값 수정 요청에 따라 합저장 기능을 이용하여 상기 메모리 박스에 저장된 파라미터 값을 업데이트하는 것을 특징으로 하는 파라미터 공유 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 파라미터 기술에 관한 것으로, 보다 상세하게는 분산 딥러닝을 위한 워커 장치들이 딥러닝 모델 파라미터를 공유하기 위한 기술에 관한 것이다.

배 경 기 술

[0002] 최근 이미지 인식, 음성 인식, 자연어 처리의 발전에 기여하며 주목 받고 있는 딥러닝 모델은 사람의 신경세포(Biological Neuron)를 모사하여 기계가 학습하도록 하는 인공신경망 (Artificial Neural Network) 기반의 기계 학습법이다.

[0003] 최근 딥러닝 모델들은 응용의 인식 성능을 높이기 위해 모델의 계층이 깊어지고(Deep), 특징(피쳐)이 많아지는 대규모 모델로 진화하고 있다. 딥러닝 모델의 규모가 커지고 입력 데이터의 양이 많아질수록 학습할 파라미터가 많아지고 계산량도 늘어난다. 이에 따라 많은 컴퓨터가 필요하며 분산 시스템에서 병렬적으로 연산하면 학습을 가속화 할 수 있다.

[0004] 딥러닝 학습의 분산 병렬 처리 시에 매 학습과정이 반복될 때마다 각 분산 컴퓨터 장치(워커 장치)들이 계산한 파라미터를 서로 공유할 수 있다. 많은 분산 딥러닝 시스템에서는 중앙 집중형 파라미터 서버를 이용하여 파라미터를 공유한다. 파라미터 서버는 매 학습과정 마다 각 워커 장치에서 학습된 파라미터를 수집하고 종합하여 다시 워커 장치들에게 나누어주는 역할을 한다.

[0005] 분산 학습된 파라미터를 업데이트하는 시점에 따라서 동기식 방식과 비동기식 방식으로 나눌 수 있다. 동기식 업데이트의 경우는 모든 워커 장치가 한 학습과정을 마친 시점에 파라미터 서버와 통신하여 파라미터를 업데이트 하는 방식으로써, 여러 워커 장치들 중에서 연산 속도가 가장 느린 워커 장치에 의해서 전체 학습 성능이 결정되게 된다. 비동기식 업데이트 방식은 파라미터 서버가 컴퓨터들로부터 늦거나 빨리 도착하는 파라미터들의 동기를 맞추지 않고 학습을 진행하는 방식이다. 비동기식 방식은 동기식 방식에 비해 정확성을 크게 희생시키지 않으면서 빠르게 트레이닝 할 수 있는 장점이 있어서, 최근 분산 딥러닝 학습에서는 비동기식 방식을 많이 채택하고 있다.

[0006] 분산 딥러닝 학습에서 워커 장치가 많아질수록 병렬도가 높아지므로 연산 속도는 빨라지지만 연산된 결과를 파라미터 서버와 통신하는데 걸리는 시간이 상대적으로 늘어나게 된다. 파라미터 서버와의 통신 속도가 느릴 경우에 전체 학습 성능이 저하될 수 있다. 따라서 분산 병렬 환경에서 딥러닝 모델을 학습할 때 파라미터 교환 시간이 중요한 요소라고 볼 수 있다.

[0007] 한편, 한국공개특허 제 10-2012-0140104 호"메모리의 데이터 저장 방법"는 메모리의 데이터 저장 방법에 관한 것으로서, 더욱 상세하게는 차량의 제어기 등에서 각 변수 조건에 따라 연산된 학습치를 메모리 영역에 효율적으로 저장할 수 있는 메모리의 데이터 저장 방법에 관하여 개시하고 있다.

[0008] 그러나, 한국공개특허 제 10-2012-0140104 호는 학습치(파라미터)를 메모리에 효율적으로 저장하기 위한 것으로, 메모리에 저장된 파라미터를 다수의 워커 장치들이 효과적으로 공유하는 측면에 대해서는 침묵하고 있다.

발명의 내용

해결하려는 과제

[0009] 본 발명은 분산 딥러닝 학습에서 다수의 워커 장치들 간의 파라미터 공유를 위해서 파라미터 서버를 사용하는 대신에, 공유 메모리 장치인 메모리 박스의 공유 메모리를 통해서 파라미터를 공유하도록 하여 딥러닝 학습을 가속화하는 것을 목적으로 한다.

[0010] 파라미터 서버와 워커들은 컴퓨터 간의 통신 네트워크(예를 들어, 이더넷(Ethernet))를 통해서 요청-응답(request-response) 방식으로 파라미터를 송수신한다. 다시 말해서, 워커가 파라미터 값을 필요로 할 때는, 파라미터 서버에게 파라미터 값에 대한 요청 메시지를 전송하고, 파라미터 서버는 자신의 메인 메모리로부터 파라미터 값을 읽어와서 요청에 대한 응답으로 워커에게 전송한다. 반대로, 워커가 파라미터 값을 업데이트하고자 할 때는, 파라미터 차분값 또는 파라미터 수정값을 포함한 파라미터 업데이트 요청 메시지를 파라미터 서버에게 전송하고, 파라미터 서버는 받은 값을 이용하여 메인 메모리에 저장된 파라미터의 값을 업데이트하고 워커에게 응답 메시지를 전송한다.

[0011] 분산 딥러닝 학습을 수행할 때 다수의 분산 워커 간에 대규모 파라미터 송수신이 빈번하게 발생하는데, 상기에서 기술한 방식대로 파라미터 서버를 이용하면 네트워크를 통한 통신 오버헤드가 크게 발생하고, 워커와 파라미터 서버에서 메시지 처리 시간도 크게 나타날 수 있다. 따라서, 이보다 개선된 방식이 필요하다.

[0012] 그에 반해서 메모리 박스는 독립적인(stand-alone) 컴퓨터가 아니고, 컴퓨터에 장착하여 사용 할 수 있는 하나의 장치(device)이다. 메모리 박스는 대용량의 메모리를 보유하고 PCIe와 같은 시스템 버스를 통해서 컴퓨터에 연결된다. 따라서, 파라미터 서버에 비해서 매우 빠른 속도로 데이터를 제공할 수 있다. 또한, 메모리 박스는 다수의 연결 커넥터를 보유하고 있어서, 동시에 다수의 워커와 연결되어 그들로부터 공유될 수 있다. 메모리 박스가 보유한 대규모 메모리는 다수의 워커들이 공유 메모리로 사용할 수 있다.

[0013] 파라미터 서버와 메모리 박스는 이러한 차이점으로 인해서 사용 방법이 크게 다르다. 메모리 박스는 컴퓨터 장치이므로 이를 사용할 때 워커가 주도적(active)으로 동작한다. 다시 말해서, 워커가 메모리 박스로부터 데이터를 리드(read)하여 파라미터 값을 가져 가고, 반대로 메모리 박스에 데이터를 라이트(write)하여 파라미터 값을 저장할 수 있다. 또한, 분산 워커들이 공유 메모리를 활용하여 딥러닝 파라미터를 공유하기 위해서는 새로운 파

라미터 공유 방법이 필요하다.

[0014] 이러한 특징들로 인하여 기존의 파라미터 서버를 사용하던 분산 딥러닝 프레임워크로는 메모리 박스를 이용할 수 없다. 메모리 박스를 이용하여 파라미터를 공유하면 메모리 박스의 빠른 접근 속도로 인해서 딥러닝 학습을 가속화 할 수 있다. 그러나, 메모리 박스를 이용하기 위해서는 분산 딥러닝 프레임워크가 메모리 박스를 통하여 파라미터를 공유하도록 수정되어야 한다.

[0015] 따라서, 상기와 같은 이유로 인해서 본 발명의 목적은, 분산 딥러닝 학습에서 다수의 워커 장치들이 메모리 박스의 공유 메모리를 통해서 파라미터를 공유할 수 있는, 파라미터 공유 장치 및 방법을 제공하는데 있다.

[0016] 또한, 본 발명은 분산 딥러닝 학습에서 파라미터 서버를 메모리 박스로 대체 지원함에 있어서 딥러닝 프레임워크가 가진 원래의 기능과 사용자가 사용하는 딥러닝 모델 개발 및 학습 인터페이스에 수정을 가하지 않고, 다수의 워커 장치가 투명하게 메모리 박스를 통해 파라미터를 공유하는 것을 목적으로 한다.

과제의 해결 수단

[0017] 상기한 목적을 달성하기 위한 본 발명의 일실시예에 따른 파라미터 공유 방법은 파라미터 공유 장치를 이용하는 방법에 있어서, 메모리 박스에 저장될 파라미터의 메모리 영역을 할당하기 위한 파라미터 정보를 수신하는 단계; 상기 메모리 박스의 매핑 테이블에 잠금을 걸고, 매핑 테이블을 읽어오는 단계; 상기 파라미터 정보에 기반하여 상기 매핑 테이블에서 상기 메모리 박스에 파라미터를 저장할 메모리 영역의 할당 여부를 확인하는 단계; 상기 메모리 영역의 할당 여부에 따라 매핑 정보를 수정한 매핑 테이블을 상기 메모리 박스에 쓰고, 상기 매핑 테이블의 잠금을 해제하는 단계 및 상기 메모리 영역을 할당한 메모리 주소를 고려하여 상기 파라미터를 공유하는 단계를 포함한다.

[0018] 이 때, 상기 수신하는 단계는 상기 파라미터의 파라미터 식별자 및 파라미터 크기 중 적어도 하나를 포함하는 파라미터 정보를 수신할 수 있다.

[0019] 이 때, 파라미터 크기는 파라미터를 저장하기 위해 필요한 메모리 크기일 수 있다.

[0020] 이 때, 상기 매핑 테이블은 각각의 엔트리에 파라미터의 식별자, 메모리 영역에 대한 메모리 주소 및 참조 카운트를 포함하는 매핑 정보가 기록될 수 있다.

[0021] 이 때, 상기 할당 여부를 확인하는 단계는 상기 매핑 테이블의 엔트리를 확인하여 상기 메모리 박스에 상기 파라미터의 메모리 영역이 기 할당되어 있는 경우, 상기 매핑 테이블에 상기 파라미터에 상응하는 엔트리에 참조 카운트를 증가시켜 상기 매핑 테이블을 업데이트 할 수 있다.

[0022] 이 때, 상기 할당 여부를 확인하는 단계는 상기 매핑 테이블의 엔트리를 확인하여 상기 메모리 박스에 상기 파라미터의 메모리 영역이 할당되지 않은 경우, 상기 파라미터 크기만큼 상기 메모리 박스에 메모리 영역을 할당하고, 상기 메모리 영역이 할당된 파라미터에 대한 매핑 정보를 상기 매핑 테이블의 새로운 엔트리에 추가하여 상기 매핑 테이블을 업데이트 할 수 있다.

[0023] 이 때, 상기 매핑 테이블의 잠금을 해제하는 단계는 상기 메모리 영역이 할당된 파라미터에 대한 상기 메모리 박스의 메모리 주소를 상기 파라미터 공유 장치에 기록할 수 있다.

[0024] 이 때, 상기 공유하는 단계는 상기 파라미터 공유 장치에 기록된 메모리 주소를 참조하여 상기 메모리 박스에 저장된 파라미터 값을 읽어오는(read) 단계; 모델 알고리즘을 이용하여 상기 메모리 박스의 파라미터 값에 상응하는 파라미터 차분 값을 계산하는 단계 및 상기 파라미터 차분 값을 이용하여 상기 메모리 박스의 파라미터 값을 수정하는 단계를 포함할 수 있다.

[0025] 이 때, 상기 파라미터 값을 수정하는 단계는 상기 메모리 박스가 합저장 기능 수행이 가능한 경우, 상기 파라미터 차분 값을 이용하여 상기 메모리 박스의 합저장 기능을 통해 상기 메모리 박스의 파라미터 값을 수정할 수 있다.

[0026] 이 때, 상기 파라미터 값을 수정하는 단계는 상기 메모리 박스가 합저장 기능 수행이 불가능한 경우, 상기 메모리 박스로부터 파라미터 값을 다시 읽어오고(read), 상기 파라미터 차분 값과 다시 읽어온 파라미터 값을 이용하여 산출한 파라미터 수정 값을 상기 메모리 박스에 쓸 수 있다(write).

[0027] 또한, 상기한 목적을 달성하기 위한 본 발명의 일실시예에 따른 파라미터 공유 방법은 파라미터 공유 장치를 이용하는 방법에 있어서, 메모리 박스에 파라미터가 저장된 메모리 영역을 해제하기 위한 파라미터 정보를 수신하

는 단계; 상기 메모리 박스의 매핑 테이블에 잠금을 걸고, 매핑 테이블을 읽어오는 단계; 상기 매핑 테이블에 기반하여 상기 메모리 박스에 상기 파라미터가 할당된 메모리 영역의 해제 여부를 확인하는 단계; 상기 메모리 영역의 해제 여부에 따라 매핑 정보를 수정한 매핑 테이블을 상기 메모리 박스에 쓰고, 상기 매핑 테이블의 잠금을 해제하는 단계 및 상기 메모리 영역을 해제한 메모리 주소를 고려하여 상기 파라미터를 공유하는 단계를 포함한다.

[0028] 이 때, 상기 수신하는 단계는 상기 파라미터의 파라미터 식별자 및 상기 파라미터가 저장된 메모리 영역에 대한 메모리 주소 중 적어도 하나를 포함하는 파라미터 정보를 수신할 수 있다.

[0029] 이 때, 상기 매핑 테이블은 각각의 엔트리에 파라미터의 식별자, 메모리 영역에 대한 메모리 주소 및 참조 카운트를 포함하는 매핑 정보가 기록될 수 있다.

[0030] 이 때, 상기 읽어오는 단계는 상기 매핑 테이블에 상기 파라미터에 상응하는 엔트리에 참조 카운트를 감소시켜 상기 매핑 테이블을 업데이트 하는 단계 및 상기 파라미터 공유 장치에 기록된 상기 파라미터에 상응하는 메모리 주소를 삭제하는 단계를 포함할 수 있다.

[0031] 이 때, 상기 해제 여부를 확인하는 단계는 상기 매핑 테이블의 참조 카운트가 최소값인 경우, 상기 메모리 박스에 할당된 메모리 영역을 해제하고, 상기 메모리 영역에 상응하는 매핑 테이블의 엔트리를 삭제할 수 있다.

[0032] 이 때, 상기 공유하는 단계는 상기 파라미터 공유 장치에 기록된 메모리 주소를 참조하여 상기 메모리 박스에 저장된 파라미터 값을 읽어오는(read) 단계; 모델 알고리즘을 이용하여 상기 메모리 박스의 파라미터 값에 상응하는 파라미터 차분 값을 계산하는 단계 및 상기 파라미터 차분 값을 이용하여 상기 메모리 박스의 파라미터 값을 수정하는 단계를 포함할 수 있다.

[0033] 이 때, 상기 파라미터 값을 수정하는 단계는 상기 메모리 박스가 합저장 기능 수행이 가능한 경우, 상기 파라미터 차분 값을 이용하여 상기 메모리 박스의 합저장 기능을 통해 상기 메모리 박스의 파라미터 값을 수정할 수 있다.

[0034] 이 때, 상기 파라미터 값을 수정하는 단계는 상기 메모리 박스가 합저장 기능 수행이 불가능한 경우, 상기 메모리 박스로부터 파라미터 값을 다시 읽어오고(read), 상기 파라미터 차분 값과 다시 읽어온 파라미터 값을 이용하여 산출한 파라미터 수정 값을 상기 메모리 박스에 쓸 수 있다(write).

[0035] 또한, 상기한 목적을 달성하기 위한 본 발명의 일실시예에 따른 파라미터 공유 장치는 메모리 박스에 파라미터가 저장될 메모리 영역의 할당 관리를 수행하고, 상기 메모리 영역의 할당 관리에 따라 상기 메모리 박스에 저장된 매핑 테이블을 업데이트 하는 메모리 할당부 및 상기 파라미터가 저장될 메모리 영역의 할당 관리를 위한 파라미터 정보를 상기 메모리 할당부에 제공하고, 상기 메모리 박스에 저장된 파라미터를 공유하는 연산 처리부를 포함한다.

[0036] 이 때, 상기 메모리 할당부는 상기 연산 처리부로부터 상기 메모리 영역의 할당 관리를 위한 상기 파라미터 정보를 수신하는 파라미터 정보 수신부; 상기 메모리 박스의 잠금을 관리하고, 상기 매핑 테이블을 업데이트하는 매핑 테이블 관리부 및 상기 매핑 테이블의 참조 카운트를 수정하여, 상기 메모리 영역의 할당 관리를 수행하는 공유 메모리 관리부를 포함할 수 있다.

[0037] 이 때, 상기 메모리 할당부는 상기 참조 카운트에 따라 상기 메모리 영역에서 파라미터를 공유 중인 다른 파라미터 공유 장치의 개수를 확인할 수 있다.

[0038] 이 때, 상기 메모리 박스는 상기 연산 처리부의 파라미터 값 수정 요청에 따라 합저장 기능을 이용하여 상기 메모리 박스에 저장된 파라미터 값을 업데이트할 수 있다.

발명의 효과

[0039] 본 발명은 분산 딥러닝 학습에서 다수의 워커 장치들이 파라미터를 공유하기 위해서 파라미터 서버를 사용하는 대신에, 공유 메모리 장치인 메모리 박스에서 제공하는 공유 메모리를 통해서 파라미터를 공유할 수 있다.

[0040] 또한, 본 발명은 통신 메시지 형태가 아니라 로컬 메모리 접근 방식으로 파라미터를 송수신 함으로써 통신 오버헤드 경감 및 메시지 처리 시간 감축을 통해서 딥러닝 학습을 가속화할 수 있다.

[0041] 또한, 본 발명은 분산 딥러닝 학습에서 파라미터 서버를 메모리 박스로 대체 지원함에 있어서 딥러닝 프레임워크가 가진 원래의 기능과 사용자가 사용하는 딥러닝 모델 개발 및 학습 인터페이스에 수정을 가하지 않고, 다수

의 워커 장치가 투명하게 메모리 박스를 통해 파라미터를 공유할 수 있다.

도면의 간단한 설명

[0042] 도 1은 본 발명의 일실시예에 따른 분산 딥러닝 프레임워크를 나타낸 블록도이다.

도 2는 도 1에 도시된 메모리 박스의 일 예를 세부적으로 나타낸 블록도이다.

도 3은 본 발명의 일실시예에 따른 파라미터 공유 장치를 나타낸 블록도이다.

도 4는 도 3에 도시된 메모리 박스 접근부의 일 예를 세부적으로 나타낸 블록도이다.

도 5는 도 2 및 도 4에 도시된 메모리 박스 접근부와 메모리 박스의 일 예를 세부적으로 나타낸 블록도이다.

도 6은 본 발명의 일실시예에 따른 분산 딥러닝 프레임워크에서 파라미터 공유를 나타낸 도면이다.

도 7은 본 발명의 일실시예에 따른 메모리 영역 할당을 위한 파라미터 공유 방법을 나타낸 동작흐름도이다.

도 8은 본 발명의 일실시예에 따른 메모리 영역 해제를 위한 파라미터 공유 방법을 나타낸 동작흐름도이다.

도 9는 본 발명의 일실시예에 따른 메모리 박스가 합저장 기능 수행이 가능한 경우, 합저장 기능을 이용한 파라미터 공유 방법을 나타낸 시퀀스 다이어그램이다.

도 10은 본 발명의 일실시예에 따른 메모리 박스가 합저장 기능 수행이 불가능한 경우, 파라미터 공유 방법을 나타낸 시퀀스 다이어그램이다.

발명을 실시하기 위한 구체적인 내용

[0043] 본 발명을 첨부된 도면을 참조하여 상세히 설명하면 다음과 같다. 여기서, 반복되는 설명, 본 발명의 요지를 불필요하게 흐릴 수 있는 공지 기능, 및 구성에 대한 상세한 설명은 생략한다. 본 발명의 실시형태는 당 업계에서 평균적인 지식을 가진 자에게 본 발명을 보다 완전하게 설명하기 위해서 제공되는 것이다. 따라서, 도면에서의 요소들의 형상 및 크기 등은 보다 명확한 설명을 위해 과장될 수 있다.

[0044] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성 요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0045] 이하, 본 발명에 따른 바람직한 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0046] 도 1은 본 발명의 일실시예에 따른 분산 딥러닝 프레임워크를 나타낸 블록도이다. 도 2는 도 1에 도시된 메모리 박스의 일 예를 세부적으로 나타낸 블록도이다. 도 3은 본 발명의 일실시예에 따른 파라미터 공유 장치를 나타낸 블록도이다. 도 4는 도 3에 도시된 메모리 박스 접근부의 일 예를 세부적으로 나타낸 블록도이다.

[0047] 도 1을 참조하면, 본 발명의 일실시예에 따른 분산 딥러닝 프레임워크는 복수개의 파라미터 공유 장치(10, 20, 30)들과 메모리 박스(1)로 구성된다.

[0048] 파라미터 공유 장치(10, 20, 30)들은 분산 딥러닝의 워커 장치라 불리는 독립적인 컴퓨터 장치에 상응할 수 있다.

[0049] 이 때, 파라미터 공유 장치(10, 20, 30)들은 메모리 박스(1)에 딥러닝 파라미터를 저장하고, 메모리 박스(1)에 저장된 딥러닝 파라미터를 공유하여 서로 협력적으로 학습할 수 있다.

[0050] 메모리 박스(1)는 전용 하드웨어로 구현된 공유 메모리 장치(device)에 상응할 수 있으며, 보다 빠르게 데이터를 저장하고 공유할 수 있다. 메모리 박스(1)는 다중 머신 간에 데이터를 저지연, 고속으로 공유 가능하게 하는 FPGA 통신 가속 공유 메모리 장치에 상응할 수 있다.

[0051] 메모리 박스(1)는 각 머신에 연결 가능한 PCIe 기반의 다수의 연결 커넥터를 보유하고 있어서, 각 머신에서 로컬 디바이스처럼 접근할 수 있으며 다중 머신에서 동시에 접근 할 수 있고, 대용량의 메모리를 가질 수 있다.

[0052] 또한, 메모리 박스(1)는 일반적인 네트워크보다 빠른 속도로 데이터를 리드(read)하거나 라이트(write)할 수 있다. 컴퓨터 노드에서는 DMA나 PIO 방식을 통해서 메모리 박스에 데이터를 읽거나 쓸 수 있다. 응용 프로그램은 메모리 박스의 디바이스 드라이버와 그 상위에 위치한 라이브러리를 통해서 메모리 박스(1)를 사용할 수 있다. 딥러닝 모델의 분산 병렬 학습에 있어서 메모리 박스(1)를 이용하면 워커 장치들 간에 파라미터를 저지연, 고속으로 공유할 수 있다.

[0053] 또한, 메모리 박스(1)는 데이터의 합저장(AssignAdd) 기능을 보유할 수도 있어서 딥러닝 파라미터를 효과적으로 업데이트 할 수 있다.

[0054] 도 2를 참조하면, 본 발명의 일실시예에 따른 메모리 박스(1)는 매핑 테이블 저장부(1a)와 파라미터 저장부(1b)를 포함한다.

[0055] 메모리 박스(1)는 매핑 테이블 저장부(1a)와 파라미터 저장부(1b)를 통해서 공유 메모리에 매핑 테이블과 딥러닝 파라미터를 저장할 수 있다.

[0056] 매핑 테이블은 각각의 엔트리에 파라미터의 식별자, 메모리 영역에 대한 메모리 주소 및 참조 카운트를 포함하는 매핑 정보가 기록될 수 있다.

[0057] 이 때, 매핑 테이블 저장부(1a)는 메모리 박스(1)와 파라미터를 공유하는 파라미터 공유 장치들(10, 20, 30)에 상응하는 각각의 매핑 정보를 기록할 수도 있다.

[0058] 도 3을 참조하면, 본 발명의 일실시예에 따른 파라미터 공유 장치 1(10)는 딥러닝 모델 복제부(11), 분산 딥러닝 학습 엔진부(12), CPU 장치 접근부(13), GPU 장치 접근부(14) 및 메모리 박스 접근부(100)를 포함할 수 있다.

[0059] 딥러닝 모델 복제부(11)는 파라미터를 학습할 수 있다.

[0060] 분산 딥러닝 학습 엔진부(12)는 딥러닝 모델 복제부(11)를 실행시키는 하부 엔진으로서, 메모리 박스(1)를 로컬에 위치한 독립적인 장치로 인식하여 메모리 박스 접근부(100)를 통해 메모리 박스(1)에 파라미터를 저장하거나, 메모리 박스(1)로부터 파라미터를 읽어와서 학습을 진행할 수 있다.

[0061] CPU 장치 접근부(13)는 CPU 에 접근할 수 있다.

[0062] GPU 장치 접근부(14)는 GPU 에 접근할 수 있다.

[0063] 메모리 박스 접근부(100)는 메모리 박스(1)에 접근할 수 있다.

[0064] CPU와 GPU는 계산 연산을 실행하는데 반해서 메모리 박스(1)는 계산 연산이 아니라 파라미터를 저장하기 위한 용도로 사용될 수 있다.

[0065] 메모리 박스 접근부(100)는 메모리 박스 장치 드라이버 또는 장치 드라이버의 상위에 위치한 메모리 박스 장치 라이브러리에서 제공하는 인터페이스를 통해서 메모리 박스(1)가 제공하는 기능을 이용할 수 있다.

[0066] 다시 말해서, 메모리 박스 접근부(100)는 메모리 박스(1)에서 제공하는 공유 메모리에 읽기, 쓰기, 잠금 걸기, 잠금 해제, 합저장 등의 기능을 이용할 수 있다.

[0067] CPU, GPU는 각 워커 장치들이 포함하는 구성이지만, 메모리 박스(1)는 여러 워커 장치들이 공유할 수 있다. 이 때, 메모리 박스 접근부(100)는 워커 장치들이 동일한 파라미터에 접근하는 경우, 동일한 메모리 주소에 접근하도록 하여 파라미터를 공유할 수 있다.

[0068] 이 때, 메모리 박스 접근부(100)는 모듈화되어 메모리 박스(1)와의 파라미터 공유를 위하여 기존의 워커 장치에 연결시켜 사용될 수 있다.

[0069] 즉, 메모리 박스 접근부(100)는 워커 장치에 연결시켜 메모리 박스(1)기반 분산형 딥러닝 파라미터를 공유하기 위한 파라미터 공유 장치 1(10)에 상응할 수도 있다.

[0070] 도 4를 참조하면, 본 발명의 일실시예에 따른 메모리 박스 접근부(100)는 메모리 할당부(110) 및 연산 처리부(120)를 포함할 수 있다.

[0071] 메모리 할당부(110)는 메모리 박스(1)에 파라미터가 저장될 메모리 영역의 할당 관리를 수행할 수 있고, 메모리 영역의 할당 관리에 따라 상기 메모리 박스에 저장된 매핑 테이블을 업데이트 할 수 있다.

[0072] 이 때, 메모리 할당부(113)는 참조 카운트에 따라 상기 메모리 영역에서 파라미터를 공유 중인 파라미터 공유 장치의 개수를 확인할 수 있고, 상기 참조 카운트가 최소값이 되는 경우, 상기 연산 처리부의 메모리 영역에 대한 메모리 주소를 삭제하고 상기 메모리 박스에서 상기 메모리 영역을 해제할 수 있다.

[0073] 연산 처리부(120)는 파라미터가 저장될 메모리 영역의 할당 관리를 위한 파라미터 정보를 메모리 할당부(110)에 제공하고, 메모리 박스(1)에 저장된 파라미터를 공유할 수 있다.

[0074] 이 때, 연산 처리부(120)는 메모리 박스(1)의 합저장 기능을 이용하여 메모리 박스(1)에 저장된 메모리 박스의 파라미터 값을 업데이트 할 수 있다.

[0075] 도 5는 도 2 및 도 4에 도시된 메모리 박스 접근부와 메모리 박스의 일 예를 세부적으로 나타낸 블록도이다.

[0076] 도 5를 참조하면, 메모리 할당부(110)는 파라미터 정보 수신부(111), 매핑 테이블 관리부(112) 및 공유 메모리 관리부(113)를 포함할 수 있다.

[0077] 연산 처리부(120)는 파라미터 객체부(121) 및 파라미터 처리부(122)를 포함할 수 있다.

[0078] 파라미터 정보 수신부(111)는 메모리 박스(1)에 파라미터를 저장할 메모리를 할당하거나 할당된 메모리를 해제하기 위해 필요한 파라미터에 관한 정보를 파라미터 객체부(121)로부터 수신할 수 있다.

[0079] 이 때, 파라미터 정보 수신부(111)는 파라미터의 파라미터 식별자, 파라미터 크기 및 파라미터가 저장된 메모리 영역에 대한 메모리 주소 중 적어도 하나를 포함하는 파라미터 정보를 파라미터 객체부(121)로부터 수신할 수 있다.

[0080] 매핑 테이블 관리부(112)는 메모리 박스(1)의 공유 메모리로부터 매핑 테이블을 읽어올 수 있다. 매핑 테이블은 파라미터와 파라미터가 저장된 공유 메모리 주소의 매핑 정보를 관리하는 테이블에 상응할 수 있다. 매핑 테이블의 각 엔트리에는 파라미터 식별자, 공유 메모리 주소, 그리고 참조 카운트를 포함할 수 있다.

[0081] 공유 메모리 관리부(113)는 메모리 박스(1)에 파라미터의 메모리 영역을 할당할 수 있다.

[0082] 공유 메모리 관리부(113)는 매핑 테이블 관리부(112)에서 읽어온 매핑 테이블을 검색하여 메모리 박스(1)에 상기 파라미터를 저장할 메모리 영역의 할당 여부를 판단할 수 있다.

[0083] 이 때, 공유 메모리 관리부(113)는 매핑 테이블에서 파라미터를 검색하여 파라미터의 메모리 영역이 할당 되었는지 여부를 확인하고, 메모리가 할당되어 있는 경우, 매핑 테이블의 엔트리에 참조 카운트를 증가시키고 메모리 주소를 파라미터 객체부(121)에 기록할 수 있다.

[0084] 이 때, 공유 메모리 관리부(113)는 메모리가 아직 할당되지 않은 경우, 메모리 박스(1)에 상기 파라미터를 저장할 메모리 영역을 할당하고 매핑 테이블에 새로운 엔트리를 추가하며, 할당한 메모리 주소를 파라미터 객체부(121)에 기록할 수 있다. 매핑 테이블 관리부(112)는 수정된 매핑 테이블을 메모리 박스(1)에 쓸 수 있다(write). 이 때, 매핑 테이블 관리부(112)는 메모리 박스(1)의 메모리 영역에 대해 잠금을 걸거나 잠금을 해제할 수 있다.

[0085] 이 때, 매핑 테이블 관리부(112)는 매핑 테이블을 읽어오기 전에 메모리 박스(1)의 매핑 테이블이 저장된 메모리 영역에 잠금을 걸고, 메모리 박스(1)에 수정된 매핑 테이블을 쓴 후에 상기 잠금을 해제할 수 있다.

[0086] 메모리 주소는 메모리 박스(1)의 메모리 영역 전체에서 특정한 메모리 위치를 지정한 것에 상응할 수 있다. 메모리 주소는 메모리 박스의 디바이스 드라이버 및 접근 라이브러리에서 제공하는 방식에 따르며, 디바이스 메모리 주소 또는 이와 매핑된 가상 주소 또는 디바이스 메모리 주소에 매핑된 식별자 등에 상응할 수 있다.

[0087] 또한, 공유 메모리 관리부(113)는 메모리 박스(1)에 할당된 파라미터의 메모리 영역을 해제할 수 있다.

[0088] 공유 메모리 관리부(113)는 매핑 테이블 관리부(112)가 읽어온 매핑 테이블을 검색하여 메모리 박스(1)에서 파라미터가 할당된 메모리 영역의 해제 여부를 판단할 수 있다.

[0089] 이 때, 공유 메모리 관리부(113)는 매핑 테이블에서 파라미터 식별자 또는 메모리 주소를 이용하여 메모리 영역을 해제할 파라미터에 관한 엔트리를 검색한 후, 해당 엔트리의 참조 카운트를 감소시킬 수 있다.

[0090] 이 때, 공유 메모리 관리부(113)는 참조 카운트값에 따라서 메모리 영역의 해제 여부를 판단할 수 있다. 이 때, 공유 메모리 관리부(113)는 참조 카운트가 최소값(예를 들어, 0)이면 메모리 영역을 해제할 수 있고, 참조 카운트가 최소값이 아닌 경우, 메모리 영역의 해제를 수행하지 않을 수 있다.

[0091] 이 때, 공유 메모리 관리부(113)는 메모리 영역을 해제하도록 결정한 경우, 메모리 박스(1)에서 파라미터의 메모리 영역을 해제하고, 매핑 테이블에서 해당 엔트리를 삭제할 수 있다. 매핑 테이블 관리부(112)는 수정된 매핑 테이블을 메모리 박스에 쓸 수 있다(write). 매핑 테이블 관리부(112)는 메모리 박스(1)의 메모리 영역에 대해 잠금을 걸거나 잠금을 해제할 수 있다. 매핑 테이블 관리부(112)는 매핑 테이블을 읽어오기 전에 메모리 박스(1)의 매핑 테이블이 저장된 메모리 영역에 잠금을 걸고, 메모리 박스(1)에 매핑 테이블을 쓴 후에 잠금을 해

제할 수 있다.

[0092] 파라미터 객체부(121)는 파라미터의 메모리 영역에 대한 메모리 주소를 파라미터 정보 수신부(111)에 파라미터 정보로 제공할 수 있다.

[0093] 파라미터 처리부(122)는 파라미터 객체(121)에 기록된 메모리 주소를 통해서 메모리 박스(1)에 파라미터 값을 쓰기(write), 읽기(read) 및 메모리 박스(1)에서 제공하는 합저장(AssignAdd) 기능을 이용하여 학습한 파라미터를 수정(업데이트)할 수 있다.

[0094] 이 때, 파라미터 처리부(122)는 메모리 박스(1)에 저장된 메모리 박스의 파라미터 값을 읽고(read), 모델 알고리즘을 이용하여 상기 메모리 박스의 파라미터 값에 상응하는 파라미터 차분 값을 계산할 수 있다.

[0095] 이 때, 모델 알고리즘은 확률적 경사 하강 법(Stochastic Gradient Descent) 알고리즘이 사용될 수 있으며, 이외에도 파라미터 차분 값을 계산하기 위한 다양한 알고리즘이 사용될 수 있다.

[0096] 이 때, 파라미터 처리부(122)는 메모리 박스(1)의 합저장 기능 수행 가능 여부에 따라서, 메모리 박스(1)가 합저장 기능 수행이 가능한 경우, 상기 합저장 기능을 통해 상기 파라미터 차분 값을 이용하여 상기 메모리 박스의 파라미터 값을 수정할 수 있다.

[0097] 이 때, 파라미터 처리부(122)는 메모리 박스(1)의 합저장 기능 수행 가능 여부에 따라서, 메모리 박스(1)가 합저장 기능 수행이 불가능한 경우, 상기 파라미터 차분 값과 상기 메모리 박스의 파라미터 값에 대한 파라미터 수정 값을 산출하여 상기 메모리 박스에 쓸 수 있다(write).

[0098] 도 6은 본 발명의 일실시예에 따른 분산 딥러닝 프레임워크에서 파라미터 공유를 나타낸 도면이다.

[0099] 도 6을 참조하면, 본 발명의 일실시예에 따른 분산 딥러닝 프레임워크에서 파라미터 공유 기법은 파라미터 공유 장치들(10, 20, 30)은 메모리 박스 접근부(100, 200, 300)을 통해 각자 독립적으로 복수개의 파라미터들을 학습할 수 있다.

[0100] 이 때, 파라미터 공유 장치들(10, 20, 30)은 메모리 박스 접근부(100, 200, 300)를 통해서 메모리 박스(1)에 복수개의 파라미터들을 저장하고 학습을 수행할 수 있다. 메모리 박스 접근부(100, 200, 300)는 파라미터 공유 장치들(10, 20, 30)이 메모리 박스(1)의 동일한 파라미터에 접근하는 경우에는 동일한 메모리 주소에 접근하여 동일한 파라미터를 공유할 수 있다.

[0101] 이 때, 파라미터 공유 장치들(10, 20, 30)은 동일한 메모리 주소에 파라미터를 업데이트하고, 업데이트한 파라미터를 읽어갈 수 있다. 따라서, 파라미터 공유 장치들(10, 20, 30)는 복수개의 파라미터들을 서로 협력적으로 학습할 수 있다.

[0102] 본 발명의 일실시예에 따른 파라미터 공유 방법은 메모리 영역 할당을 위한 파라미터 공유 방법과 할당된 메모리 영역의 해제를 위한 파라미터 공유 방법을 나눠서 설명한다.

[0103] 도 7은 본 발명의 일실시예에 따른 메모리 영역 할당을 위한 파라미터 공유 방법을 나타낸 동작흐름도이다.

[0104] 도 7을 참조하면, 본 발명의 일실시예에 따른 메모리 영역 할당을 위한 파라미터 공유 방법은 먼저 파라미터 정보를 수신할 수 있다(S210).

[0105] 즉, 단계(S210)는 메모리 박스(1)에 파라미터의 메모리 영역을 할당하기 위한 파라미터 정보를 수신할 수 있다.

[0106] 이 때, 단계(S210)는 파라미터의 파라미터 식별자 및 파라미터 크기 중 적어도 하나를 포함하는 파라미터 정보를 수신할 수 있다.

[0107] 이 때, 파라미터 크기는 파라미터를 저장하기 위해 필요한 메모리 크기일 수 있다.

[0108] 이 때, 매핑 테이블은 각각의 엔트리에 파라미터의 식별자, 메모리 영역에 대한 메모리 주소 및 참조 카운트를 포함하는 매핑 정보가 기록될 수 있다.

[0109] 이 때, 매핑 테이블은 메모리 박스(1)와 파라미터를 공유하는 파라미터 공유 장치들(10, 20, 30)에 상응하는 각각의 매핑 정보가 기록될 수도 있다.

[0110] 즉, 매핑 테이블은 메모리 박스(1)와 파라미터를 공유하는 파라미터 공유 장치들(10, 20, 30)에 관한 정보를 사전에 더 포함할 수도 있다.

[0111] 또한, 본 발명의 일실시예에 따른 메모리 영역 할당을 위한 파라미터 공유 방법은 매핑 테이블의 잠금 및 읽기를 수행할 수 있다(S220).

[0112] 즉, 단계(S220)는 메모리 박스(1)의 매핑 테이블에 잠금을 걸고, 매핑 테이블을 읽어올 수 있다.

[0113] 또한, 본 발명의 일실시예에 따른 메모리 영역 할당을 위한 파라미터 공유 방법은 메모리 영역의 할당 여부를 확인할 수 있다(S230).

[0114] 즉, 단계(S230)는 매핑 테이블의 엔트리를 확인하여 메모리 박스에 파라미터의 메모리 영역이 할당되지 않은 경우, 메모리 영역을 할당할 수 있다(S240).

[0115] 이 때, 단계(S240)는 파라미터 크기만큼 상기 메모리 박스에 메모리 영역을 할당할 수 있다.

[0116] 또한, 본 발명의 일실시예에 따른 메모리 영역 할당을 위한 파라미터 공유 방법은 매핑 테이블에 매핑 정보를 추가할 수 있다(S250).

[0117] 즉, 단계(S250)는 매핑 테이블의 엔트리를 확인하여 메모리 박스에 파라미터의 메모리 영역이 할당되지 않은 경우, 메모리 영역이 할당된 파라미터에 대한 매핑 정보를 매핑 테이블의 새로운 엔트리에 추가하여 매핑 테이블을 업데이트 할 수 있다.

[0118] 또한, 단계(S230)는 매핑 테이블의 엔트리를 확인하여 메모리 박스에 파라미터의 메모리 영역이 기 할당되어 있는 경우, 매핑 테이블의 참조 카운트를 증가시킬 수 있다(S260).

[0119] 즉, 단계(S260)는 매핑 테이블에 파라미터에 상응하는 엔트리에 참조 카운트를 증가시켜 매핑 테이블을 업데이트 할 수 있다.

[0120] 이 때, 단계(S260)는 참조 카운트를 '1'씩 증가 시킬 수 있다.

[0121] 또한, 본 발명의 일실시예에 따른 메모리 영역 할당을 위한 파라미터 공유 방법은 매핑 테이블의 쓰기 및 잠금 해제를 수행할 수 있다(S270).

[0122] 즉, 단계(S270)는 메모리 영역의 할당 여부에 따라 매핑 정보를 수정한 매핑 테이블을 메모리 박스(1)에 쓰고, 매핑 테이블의 잠금을 해제할 수 있다.

[0123] 이 때, 단계(S270)는 메모리 영역이 할당된 파라미터에 대한 메모리 박스의 메모리 주소를 파라미터 공유 장치에 기록할 수 있다.

[0124] 또한, 본 발명의 일실시예에 따른 메모리 영역 할당을 위한 파라미터 공유 방법은 파라미터를 공유할 수 있다(S280).

[0125] 즉, 단계(S280)는 메모리 영역을 할당한 메모리 주소를 고려하여 파라미터를 공유할 수 있다.

[0126] 이 때, 단계(S280)는 파라미터 공유 장치들(10, 20, 30)에 메모리 영역을 할당한 메모리 주소가 추가된 것으로 기록된 메모리 영역의 메모리 주소를 참조하여 메모리 박스(1)에 저장된 파라미터를 공유할 수 있다.

[0127] 이 때, 단계(S280)는 메모리 박스(1)의 매핑 테이블에 기록된 파라미터 공유 장치들(10, 20, 30)에 상응하는 메모리 주소를 참조하여 메모리 박스(1)에 저장된 파라미터를 공유할 수도 있다.

[0128] 이 때, 단계(S280)는 파라미터 공유 장치들(10, 20, 30)이 메모리 박스(1)에 저장된 파라미터 값을 읽어 올 수 있다(read).

[0129] 이 때, 단계(S280)는 모델 알고리즘을 이용하여 메모리 박스(1)의 파라미터 값에 상응하는 파라미터 차분 값을 계산할 수 있다.

[0130] 이 때, 모델 알고리즘은 확률적 경사 하강 법(Stochastic Gradient Descent) 알고리즘이 사용될 수 있으며, 이외에도 파라미터 차분 값을 계산하기 위한 다양한 알고리즘이 사용될 수 있다.

[0131] 이 때, 단계(S280)는 파라미터 차분 값을 이용하여 메모리 박스(1)의 파라미터 값을 수정할 수 있다.

[0132] 이 때, 단계(S280)는 메모리 박스(1)가 합저장 기능 수행이 가능한 경우, 메모리 박스(1)의 합저장 기능을 통해 상기 파라미터 차분 값을 이용하여 상기 메모리 박스(1)의 파라미터 값을 수정할 수 있다.

[0133] 이 때, 단계(S280)는 파라미터 공유 장치들(10, 20, 30)이 메모리 박스(1)의 합저장 기능 수행 가능 여부를 미

리 확인해둘 수 있다.

[0134] 또한, 단계(S280)는 메모리 박스(1)가 합저장 기능 수행이 불가능한 경우, 메모리 박스(1)로부터 파라미터 값을 다시 읽어오고(read), 파라미터 차분 값과 다시 읽어온 파라미터 값을 이용하여 산출한 파라미터 수정 값을 메모리 박스(1)에 쓸 수 있다(write).

[0135] 이 때, 단계(S280)는 한 번 또는 그 이상 반복적으로 실행될 수 있다.

[0136] 즉, 단계(S280)는 메모리 박스(1)가 공유하는 파라미터를 이용하여 학습을 진행할 수 있다.

[0137] 이 때, 단계(S280)는 파라미터 공유 장치들(10, 20, 30)이 동일한 파라미터를 접근하는 경우에는 동일한 메모리 주소에 접근하도록 하여 동일한 파라미터가 공유될 수 있다.

[0138] 이 때, 단계(S280)는 파라미터 공유 장치들(10, 20, 30)이 동일한 메모리 주소에서 파라미터를 읽어가고, 업데이트를 하고, 업데이트된 파라미터를 다시 읽어가서 파라미터 공유 장치들(10, 20, 30)이 서로 협력적으로 학습할 수 있다.

[0139] 나아가, 단계(S280)에서 파라미터를 공유하는 과정은 도 9 및 도 10에 대한 설명을 일 예로 하여 아래에서 상세하게 설명한다.

[0140] 도 8은 본 발명의 일실시예에 따른 메모리 영역 해제를 위한 파라미터 공유 방법을 나타낸 동작흐름도이다.

[0141] 본 발명의 일실시예에 따른 메모리 영역 해제를 위한 파라미터 공유 방법은 먼저 파라미터 정보를 수신할 수 있다(S310).

[0142] 즉, 단계(S310)는 메모리 박스(1)에 파라미터의 메모리 영역을 해제하기 위한 파라미터 정보를 수신할 수 있다.

[0143] 이 때, 단계(S310)는 파라미터의 파라미터 식별자 및 파라미터가 저장된 메모리 영역에 대한 메모리 주소 중 적어도 하나를 포함하는 파라미터 정보를 수신할 수 있다.

[0144] 이 때, 매핑 테이블은 각각의 엔트리에 파라미터의 식별자, 메모리 영역에 대한 메모리 주소 및 참조 카운트를 포함하는 매핑 정보가 기록될 수 있다.

[0145] 이 때, 매핑 테이블은 메모리 박스(1)와 파라미터를 공유하는 파라미터 공유 장치들(10, 20, 30)에 상응하는 각각의 매핑 정보가 기록될 수도 있다.

[0146] 즉, 매핑 테이블은 메모리 박스(1)와 파라미터를 공유하는 파라미터 공유 장치들(10, 20, 30)에 관한 정보를 사전에 더 포함할 수도 있다.

[0147] 또한, 본 발명의 일실시예에 따른 메모리 영역 해제를 위한 파라미터 공유 방법은 매핑 테이블의 잠금 및 읽기를 수행할 수 있다(S320).

[0148] 즉, 단계(S320)는 메모리 박스(1)의 매핑 테이블에 잠금을 걸고, 매핑 테이블을 읽어올 수 있다.

[0149] 또한, 본 발명의 일실시예에 따른 메모리 영역 해제를 위한 파라미터 공유 방법은 참조 카운트를 감소시킬 수 있다(S330).

[0150] 즉, 단계(S330)는 매핑 테이블에 파라미터에 상응하는 엔트리에 참조 카운트를 감소시켜 매핑 테이블을 업데이트 할 수 있다.

[0151] 이 때, 단계(S330)는 참조 카운트를 '1'씩 감소 시킬 수 있다.

[0152] 또한, 본 발명의 일실시예에 따른 메모리 영역 해제를 위한 파라미터 공유 방법은 메모리 주소를 삭제할 수 있다(S340).

[0153] 즉, 단계(S340)는 파라미터 공유 장치에 기록된 파라미터에 상응하는 메모리 주소를 삭제할 수 있다.

[0154] 또한, 본 발명의 일실시예에 따른 메모리 영역 해제를 위한 파라미터 공유 방법은 참조 카운트가 최소값(예를 들어, '0')인지 여부를 확인할 수 있다(S350).

[0155] 즉, 단계(S350)는 매핑 테이블의 참조 카운트가 최소값인 경우, 메모리 박스(1)에 할당된 메모리 영역을 해제하고(S360), 매핑 테이블의 매핑 정보를 삭제할 수 있다(S370).

[0156] 즉, 단계(S370)는 메모리 영역에 상응하는 매핑 테이블의 엔트리를 삭제할 수 있다.

[0157] 또한, 단계(S350)는 매핑 테이블의 참조 카운트가 최소값이 아닌 경우, 메모리 영역을 해제 하지 않고, 참조 카운트가 수정된 매핑 테이블을 업데이트 할 수 있다.

[0158] 또한, 본 발명의 일실시예에 따른 메모리 영역 해제를 위한 파라미터 공유 방법은 매핑 테이블 쓰기 및 잠금 해제를 수행할 수 있다(S380).

[0159] 즉, 단계(S380)는 메모리 영역의 해제 여부에 따라 매핑 정보를 수정한 매핑 테이블을 메모리 박스(1)에 쓰고, 매핑 테이블의 잠금을 해제할 수 있다.

[0160] 또한, 본 발명의 일실시예에 따른 메모리 영역 해제를 위한 파라미터 공유 방법은 파라미터를 공유할 수 있다(S390).

[0161] 즉, 단계(S390)는 메모리 영역을 해제한 메모리 주소를 고려하여 파라미터를 공유할 수 있다.

[0162] 이 때, 단계(S390)는 파라미터 공유 장치들(10, 20, 30)에 메모리 영역을 해제한 메모리 주소가 삭제된 것으로 기록된 나머지 메모리 영역에 대한 메모리 주소를 참조하여 메모리 박스(1)에 저장된 파라미터를 공유할 수 있다.

[0163] 이 때, 단계(S390)는 메모리 박스(1)의 매핑 테이블에 기록된 파라미터 공유 장치들(10, 20, 30)에 상응하는 메모리 주소를 참조하여 메모리 박스(1)에 저장된 파라미터를 공유할 수도 있다.

[0164] 이 때, 단계(S390)는 파라미터 공유 장치에 기록된 메모리 주소를 참조하여 메모리 박스(1)에 저장된 파라미터 값을 읽어 올 수 있다(read).

[0165] 이 때, 단계(S390)는 모델 알고리즘을 이용하여 메모리 박스의 파라미터 값에 상응하는 파라미터 차분 값을 계산할 수 있다.

[0166] 이 때, 모델 알고리즘은 확률적 경사 하강 법(Stochastic Gradient Descent) 알고리즘이 사용될 수 있으며, 이외에도 파라미터 차분 값을 계산하기 위한 다양한 알고리즘이 사용될 수 있다.

[0167] 이 때, 단계(S390)는 파라미터 차분 값을 이용하여 상기 메모리 박스의 파라미터 값을 수정할 수 있다.

[0168] 이 때, 단계(S390)는 상기 메모리 박스가 합저장 기능 수행이 가능한 경우, 상기 파라미터 차분 값을 이용하여 상기 메모리 박스의 합저장 기능을 통해 상기 메모리 박스의 파라미터 값을 수정할 수 있다.

[0169] 이 때, 단계(S390)는 파라미터 공유 장치들(10, 20, 30)이 메모리 박스(1)의 합저장 기능 수행 가능 여부를 미리 확인해둘 수 있다.

[0170] 이 때, 단계(S390)는 상기 메모리 박스가 합저장 기능 수행이 불가능한 경우, 상기 메모리 박스로부터 파라미터 값을 다시 읽어오고(read), 상기 파라미터 차분 값과 다시 읽어온 파라미터 값을 이용하여 산출한 파라미터 수정 값을 상기 메모리 박스에 쓸 수 있다(write).

[0171] 이 때, 단계(S390)는 한 번 또는 그 이상 반복적으로 실행될 수 있다.

[0172] 즉, 단계(S390)는 메모리 박스(1)가 공유하는 파라미터를 이용하여 학습을 진행할 수 있다.

[0173] 이 때, 단계(S390)는 파라미터 공유 장치들(10, 20, 30)이 동일한 파라미터를 접근하는 경우에는 동일한 메모리 주소에 접근하도록 하여 동일한 파라미터가 공유될 수 있다.

[0174] 이 때, 단계(S390)는 파라미터 공유 장치들(10, 20, 30)이 동일한 메모리 주소에서 파라미터를 읽어가고, 업데이트를 하고, 업데이트된 파라미터를 다시 읽어가서 파라미터 공유 장치들(10, 20, 30)이 서로 협력적으로 학습할 수 있다.

[0175] 나아가, 단계(S390)에서 파라미터를 공유하는 과정은 도 9 및 도 10에 대한 설명을 일 예로 하여 아래에서 상세하게 설명한다.

[0176] 도 9는 본 발명의 일실시예에 따른 메모리 박스가 합저장 기능 수행이 가능한 경우, 합저장 기능을 이용한 파라미터 공유 방법을 나타낸 시퀀스 다이어그램이다.

[0177] 도 9를 참조하면, 본 발명의 일실시예에 따른 메모리 박스가 합저장 기능 수행이 가능한 경우, 합저장 기능을 이용한 파라미터 공유 방법은 먼저 파라미터 공유 장치 1(10)이 파라미터를 읽어올 수 있다(S410).

[0178] 즉, 단계(S410)는 파라미터 공유 장치 1(10)이 메모리 박스(1)로부터 제1 메모리 박스의 파라미터 값을 읽어올

수 있다.

[0179] 또한, 본 발명의 일실시예에 따른 합저장 기능을 이용한 파라미터 공유 방법은 파라미터 차분 값을 계산할 수 있다(S420).

[0180] 즉, 단계(S420)는 파라미터 공유 장치 1(10)이 모델 알고리즘을 이용하여 제1 파라미터 차분 값을 계산할 수 있다.

[0181] 이 때, 모델 알고리즘은 확률적 경사 하강 법(Stochastic Gradient Descent) 알고리즘이 사용될 수 있으며, 이 외에도 파라미터 차분 값을 계산하기 위한 다양한 알고리즘이 사용될 수 있다.

[0182] 또한, 본 발명의 일실시예에 따른 합저장 기능을 이용한 파라미터 공유 방법은 합저장 기능을 이용하여 파라미터를 수정할 수 있다(S430).

[0183] 즉, 단계(S430)는 메모리 박스(1)의 합저장(AssignAdd) 기능을 통해 제1 파라미터 차분 값을 이용하여 제1 메모리 박스의 파라미터 값을 수정(업데이트)하여 제2 메모리 박스의 파라미터 값을 생성할 수 있다.

수학식 1

[0184] $$W_{t+1} = W_t + \triangle W_t$$

[0185] 예를 들어, 수학식 1은 합저장 기능의 일 예를 수학식으로 나타낸 것을 알 수 있다.

[0186] 이 때, 단계(S430)는 수학식 1과 같이, 제1 메모리 박스의 파라미터 값(W_t)에서 제1 파라미터 차분값(ΔW_t)을 합산하여 제2 메모리 박스의 파라미터 값(W_{t+1})을 생성할 수 있다.

[0187] 또한, 본 발명의 일실시예에 따른 합저장 기능을 이용한 파라미터 공유 방법은 파라미터 공유 장치 2(20)가 파라미터를 읽어올 수 있다(S440).

[0188] 즉, 단계(S440)는 파라미터 공유 장치 2(20)가 메모리 박스(1)로부터 제2 메모리 박스의 파라미터 값을 읽어올 수 있다.

[0189] 이 때, 단계(S440)는 파라미터 공유 장치 1(10)과 파라미터 공유 장치(2)가 비동기적으로 파라미터 공유를 수행하게 되므로, 도 9에 도시된 바와 같이 반드시 단계(S430) 이후에 수행되는 것이 아니라, 파라미터 공유 장치(1)의 파라미터 업데이트 과정과 무관하게 단계(S410) 내지 단계(S430) 중 어느 단계에서도 함께 수행될 수도 있다.

[0190] 따라서, 단계(S440)는 합저장 기능을 통해 제1 파라미터 차분 값이 업데이트 되지 않은 제1 메모리 박스의 파라미터 값을 읽어 올 수도 있다.

[0191] 그러나, 이하에서는 업데이트가 완료된 제2 메모리 박스의 파라미터 값을 읽어 오는 것으로 설명한다.

[0192] 이러한 비동기적 파라미터 공유 방법에서, 파라미터 값의 업데이트 과정 중에는 메모리 박스의 파라미터 값의 업데이트가 정확한 계산 값으로 반영되지 않을 수 있지만, 파라미터 공유가 완료되는 시점에서는 결과적으로 높은 속도로 파라미터 공유를 완료할 수 있다.

[0193] 또한, 본 발명의 일실시예에 따른 합저장 기능을 이용한 파라미터 공유 방법은 파라미터 차분 값을 계산할 수 있다(S450).

[0194] 즉, 단계(S450)는 파라미터 공유 장치 2(20)가 모델 알고리즘을 이용하여 제2 파라미터 차분 값을 계산할 수 있다.

[0195] 이 때, 모델 알고리즘은 확률적 경사 하강 법(Stochastic Gradient Descent) 알고리즘이 사용될 수 있으며, 이 외에도 파라미터 차분 값을 계산하기 위한 다양한 알고리즘이 사용될 수 있다.

[0196] 또한, 본 발명의 일실시예에 따른 합저장 기능을 이용한 파라미터 공유 방법은 합저장 기능을 이용하여 파라미터를 수정할 수 있다(S460).

[0197] 즉, 단계(S460)는 메모리 박스(1)의 합저장(AssignAdd) 기능을 통해 제2 파라미터 차분 값을 이용하여 제2 메모

리 박스의 파라미터 값을 수정(업데이트)하여 제3 메모리 박스의 파라미터 값을 생성할 수 있다.

[0198] 이 때, 단계(S460)는 수학식 1과 같이, 제2 메모리 박스의 파라미터 값(W_t)에서 제2 파라미터 차분값(ΔW_t)을 합산하여 제3 메모리 박스의 파라미터 값(W_{t+1})을 생성할 수 있다.

[0199] 이러한 과정을 통해, 복수개의 파라미터 공유 장치들(워커)이 비동기적으로 메모리 박스(1)로부터 파라미터 값을 읽어(read) 나가면서 합저장(AssignAdd) 기능을 이용하여 메모리 박스(1)의 파라미터 값을 업데이트 할 수 있다.

[0200] 도 10는 본 발명의 일실시예에 따른 메모리 박스가 합저장 기능 수행이 불가능한 경우, 파라미터 공유 방법을 나타낸 시퀀스 다이어그램이다.

[0201] 도 10을 참조하면, 본 발명의 일실시예에 따른 메모리 박스가 합저장 기능 수행이 불가능한 경우, 파라미터 공유 방법은 먼저 파라미터 공유 장치 1(10)이 파라미터를 읽어올 수 있다(S510).

[0202] 즉, 단계(S510)는 파라미터 공유 장치 1(10)이 메모리 박스(1)로부터 제1 메모리 박스의 파라미터 값을 읽어올 수 있다(read).

[0203] 또한, 본 발명의 일실시예에 따른 메모리 박스가 합저장 기능 수행이 불가능한 경우, 파라미터 공유 방법은 파라미터 차분 값을 계산할 수 있다(S520).

[0204] 즉, 단계(S520)는 파라미터 공유 장치 1(10)이 모델 알고리즘을 이용하여 제1 파라미터 차분 값을 계산할 수 있다.

[0205] 이 때, 모델 알고리즘은 확률적 경사 하강 법(Stochastic Gradient Descent) 알고리즘이 사용될 수 있으며, 이외에도 파라미터 차분 값을 계산하기 위한 다양한 알고리즘이 사용될 수 있다.

[0206] 또한, 본 발명의 일실시예에 따른 메모리 박스가 합저장 기능 수행이 불가능한 경우, 파라미터 공유 방법은 파라미터 값을 읽어올 수 있다(S530).

[0207] 즉, 단계(S530)는 단계(S520)의 파라미터 차분 값 계산 과정 동안 다른 파라미터 공유 장치에 의하여 메모리 박스(1)의 파라미터 값이 수정될 수 있으므로, 메모리 박스(1)의 파라미터 값을 다시 읽어올 수 있다.

[0208] 이 때, 단계(S530)는 단계(S520)의 파라미터 차분 값의 계산이 기설정된 시간을 초과할 때까지 계산되지 않은 경우에만, 제1 메모리 박스의 파라미터 값을 다시 읽어 올 수 있다.

[0209] 또한, 본 발명의 일실시예에 따른 메모리 박스가 합저장 기능 수행이 불가능한 경우, 파라미터 공유 방법은 파라미터 수정 값을 산출할 수 있다(S540).

[0210] 즉, 단계(S540)는 계산된 제1 파라미터 차분 값과 메모리 박스(1)에서 읽어온 제1 메모리 박스의 파라미터 값을 이용하여 제1 파라미터 수정 값을 산출 할 수 있다.

[0211] 이 때, 단계(S540)는 단계(S520)에서 기설정된 시간 이내에 파라미터 차분 값이 계산된 경우, 단계(S510)에서 읽어온 제1 메모리 박스의 파라미터 값을 이용하여 제1 파라미터 수정 값을 산출할 수 있다.

[0212] 또한, 단계(S540)는 단계(S520)에서 기설정된 시간을 초과하여 파라미터 차분 값이 계산된 경우, 단계(S530)에서 다시 읽어온 제1 메모리 박스의 파라미터 값을 이용하여 제1 파라미터 수정 값을 산출할 수 있다.

수학식 2

[0213] $$W_{t+1} = W_t' + \triangle W_t$$

[0214] 예를 들어, 수학식 2는 파라미터 값을 업데이트하는 일 예를 수학식으로 나타낸 것을 알 수 있다.

[0215] 이 때, 단계(S540)는 상기 수학식 2와 같이, 제1 메모리 박스의 파라미터 값(W_t')에서 제1 파라미터 차분값(ΔW_t)을 합산하여 제1 파라미터 수정 값(W_{t+1})을 생성할 수 있다.

[0216] 또한, 본 발명의 일실시예에 따른 메모리 박스가 합저장 기능 수행이 불가능한 경우, 파라미터 공유 방법은 파라미터 수정 값을 쓸 수 있다(S550).

[0217] 즉, 단계(S550)는 산출된 제1 파라미터 수정 값을 메모리 박스(1)에 쓰는(write) 것으로 메모리 박스(1)의 파라미터 값을 수정(업데이트)할 수 있다.

[0218] 이 때, 단계(S550)는 제1 메모리 박스의 파라미터 값에 제1 파라미터 수정 값을 쓰는 것으로, 제2 메모리 박스의 파라미터 값을 생성할 수 있다.

[0219] 또한, 본 발명의 일실시예에 따른 메모리 박스가 합저장 기능 수행이 불가능한 경우, 파라미터 공유 방법은 파라미터 공유 장치 2(20)가 파라미터를 읽어올 수 있다(S560).

[0220] 즉, 단계(S560)는 파라미터 공유 장치 2(10)가 메모리 박스(1)로부터 제2 메모리 박스의 파라미터 값을 읽어올 수 있다(read).

[0221] 이 때, 단계(S560)는 파라미터 공유 장치 1(10)과 파라미터 공유 장치(2)가 비동기적으로 파라미터 공유를 수행하게 되므로, 도 10에 도시된 바와 같이 반드시 단계(S530) 이후에 수행되는 것이 아니라, 파라미터 공유 장치(1)의 파라미터 업데이트 과정과 무관하게 단계(S510) 내지 단계(S550) 중 어느 단계에서도 수행될 수도 있다.

[0222] 따라서, 단계(S560)는 도 10에 도시된 바와 같이 제1 파라미터 수정 값이 업데이트 되지 않은 제1 메모리 박스의 파라미터 값을 읽어 올 수도 있다.

[0223] 그러나, 이하에서는 업데이트가 완료된 제2 메모리 박스의 파라미터 값을 읽어 오는 것으로 설명한다.

[0224] 이러한 비동기적 파라미터 공유 방법에서, 파라미터 값의 업데이트 과정 중에는 메모리 박스의 파라미터 값의 업데이트가 정확한 계산값으로 반영되지 않을 수 있지만, 파라미터 공유가 완료되는 시점에서는 결과적으로 높은 속도로 파라미터 공유를 완료할 수 있다.

[0225] 또한, 본 발명의 일실시예에 따른 메모리 박스가 합저장 기능 수행이 불가능한 경우, 파라미터 공유 방법은 파라미터 차분 값을 계산할 수 있다(S570).

[0226] 즉, 단계(S570)는 파라미터 공유 장치 2(20)가 모델 알고리즘을 이용하여 제2 파라미터 차분 값을 계산할 수 있다.

[0227] 이 때, 모델 알고리즘은 확률적 경사 하강 법(Stochastic Gradient Descent) 알고리즘이 사용될 수 있으며, 이외에도 파라미터 차분 값을 계산하기 위한 다양한 알고리즘이 사용될 수 있다.

[0228] 또한, 본 발명의 일실시예에 따른 메모리 박스가 합저장 기능 수행이 불가능한 경우, 파라미터 공유 방법은 파라미터 값을 읽어올 수 있다(S580).

[0229] 즉, 단계(S580)는 단계(S570)의 파라미터 차분 값 계산 과정 동안 다른 파라미터 공유 장치에 의하여 메모리 박스(1)의 파라미터 값이 수정될 수 있으므로, 메모리 박스(1)의 파라미터 값을 다시 읽어올 수 있다.

[0230] 이 때, 단계(S580)는 단계(S570)의 파라미터 차분 값이 계산이 기설정된 시간을 초과할 때까지 계산되지 않은 경우에만, 제2 메모리 박스의 파라미터 값을 다시 읽어 올 수 있다.

[0231] 또한, 본 발명의 일실시예에 따른 메모리 박스가 합저장 기능 수행이 불가능한 경우, 파라미터 공유 방법은 파라미터 수정 값을 산출할 수 있다(S590).

[0232] 즉, 단계(S590)는 계산된 제2 파라미터 차분 값과 메모리 박스(1)에서 읽어온 제2 메모리 박스의 파라미터 값을 이용하여 제2 파라미터 수정 값을 산출 할 수 있다.

[0233] 이 때, 단계(S590)는 단계(S570)에서 기설정된 시간 이내에 파라미터 차분 값이 계산된 경우, 단계(S560)에서 읽어온 제2 메모리 박스의 파라미터 값을 이용하여 제2 파라미터 수정 값을 산출할 수 있다.

[0234] 또한, 단계(S590)는 단계(S570)에서 기설정된 시간을 초과하여 파라미터 차분 값이 계산된 경우, 단계(S580)에서 다시 읽어온 제2 메모리 박스의 파라미터 값을 이용하여 제2 파라미터 수정 값을 산출할 수 있다.

[0235] 이 때, 단계(S590)는 상기 수학식 2와 같이, 제2 메모리 박스의 파라미터 값(W_t')에서 제2 파라미터 차분값(ΔW_t)을 합산하여 제2 파라미터 수정 값(W_{t+1})을 생성할 수 있다.

[0236] 또한, 본 발명의 일실시예에 따른 메모리 박스가 합저장 기능 수행이 불가능한 경우, 파라미터 공유 방법은 파라미터 수정 값을 쓸 수 있다(S600).

[0237] 즉, 단계(S600)는 산출된 제2 파라미터 값을 메모리 박스(1)에 쓰는(write) 것으로 메모리 박스(1)의 파라미터

값을 수정(업데이트)할 수 있다.

[0238] 이 때, 단계(S600)는 제2 메모리 박스의 파라미터 값에 제2 파라미터 수정 값을 쓰는 것으로, 제3 메모리 박스의 파라미터 값을 생성할 수 있다.

[0239] 이러한 과정을 통해, 메모리 박스(1)가 합저장 기능 수행이 불가능한 경우에도, 복수개의 파라미터 공유 장치들(워커)이 비동기적으로 메모리 박스(1)로부터 파라미터 값을 읽어(read) 나가면서 산출된 파라미터 수정 값을 쓰는(write) 것으로, 메모리 박스(1)의 파라미터 값을 업데이트 할 수 있다.

[0240] 이상에서와 같이 본 발명에 따른 파라미터 공유 장치 및 방법은 상기한 바와 같이 설명된 실시예들의 구성과 방법이 한정되게 적용될 수 있는 것이 아니라, 상기 실시예들은 다양한 변형이 이루어질 수 있도록 각 실시예들의 전부 또는 일부가 선택적으로 조합되어 구성될 수도 있다.

부호의 설명

[0241] 1: 메모리 박스 1a: 매핑 테이블 저장부

1b: 파라미터 저장부 10, 20, 30: 파라미터 공유 장치

11: 딥러닝 모델 복제부 12: 분산 딥러닝 학습 엔진부

13: CPU 장치 접근부 14: GPU 장치 접근부

100,200, 300: 메모리 박스 접근부 110: 메모리 할당부

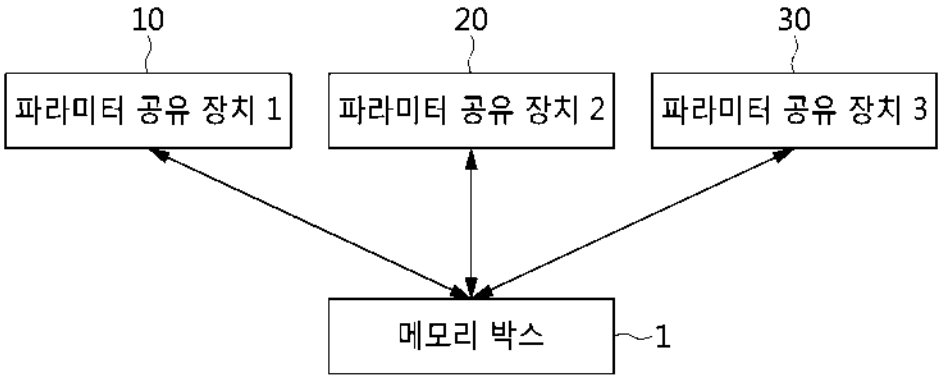
111: 파라미터 정보 수신부 112: 매핑 테이블 관리부

113: 공유 메모리 관리부 120: 연산 처리부

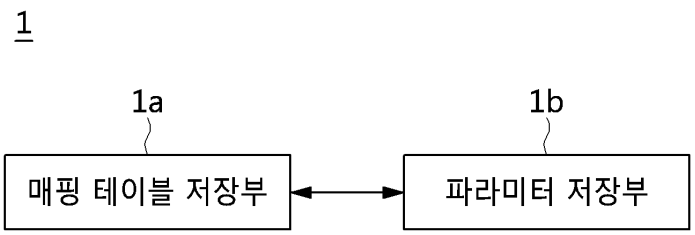
121: 파라미터 객체부 122: 파라미터 처리부

도면

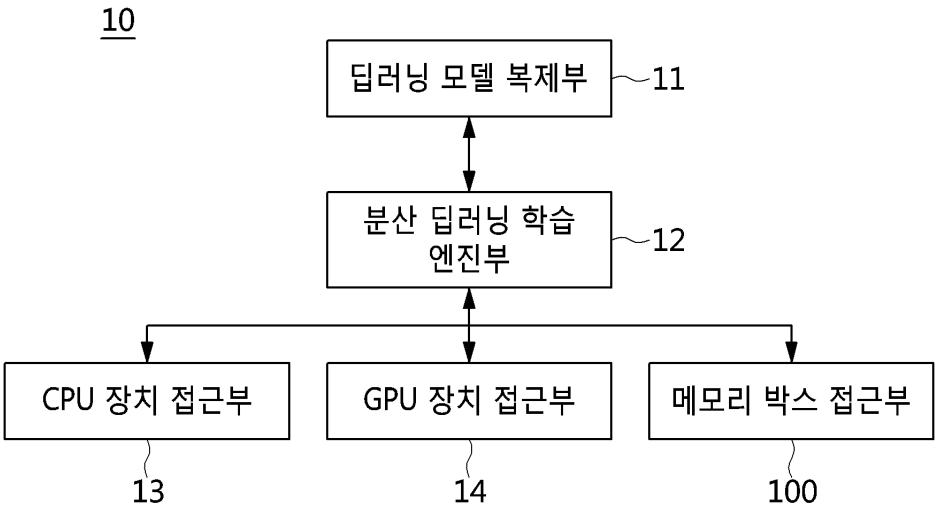
도면1



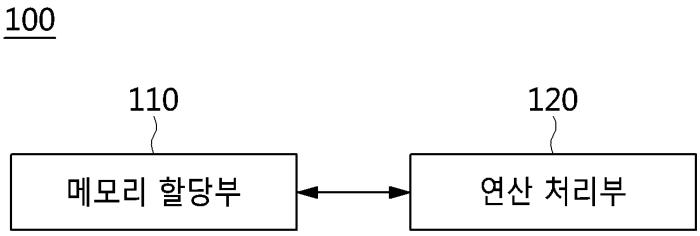
도면2



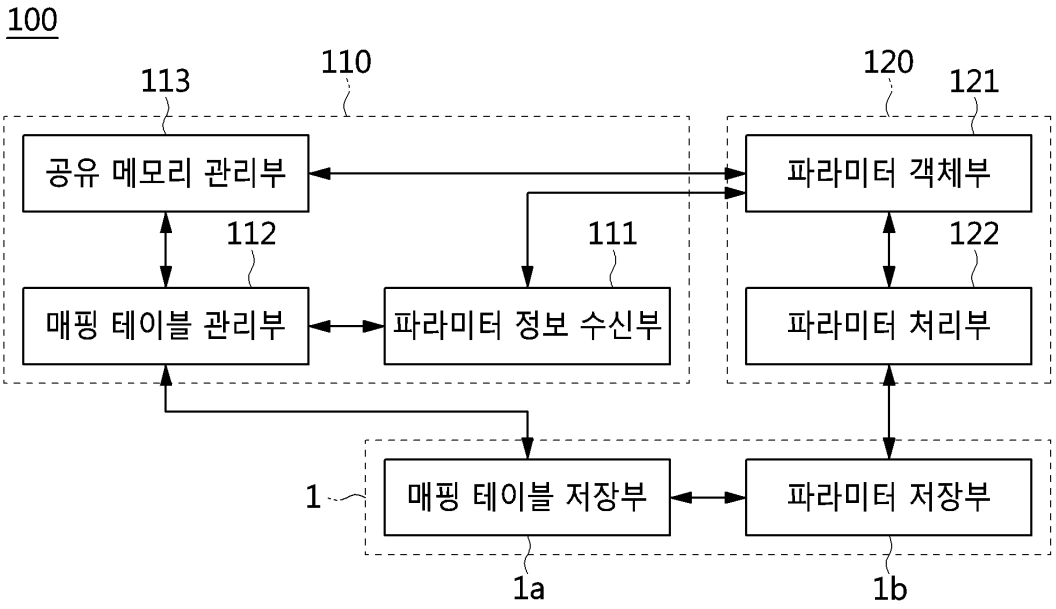
도면3


10
딥러닝 모델 복제부
11
분산 딥러닝 학습 엔진부
12
CPU 장치 접근부
GPU 장치 접근부
메모리 박스 접근부
13
14
100

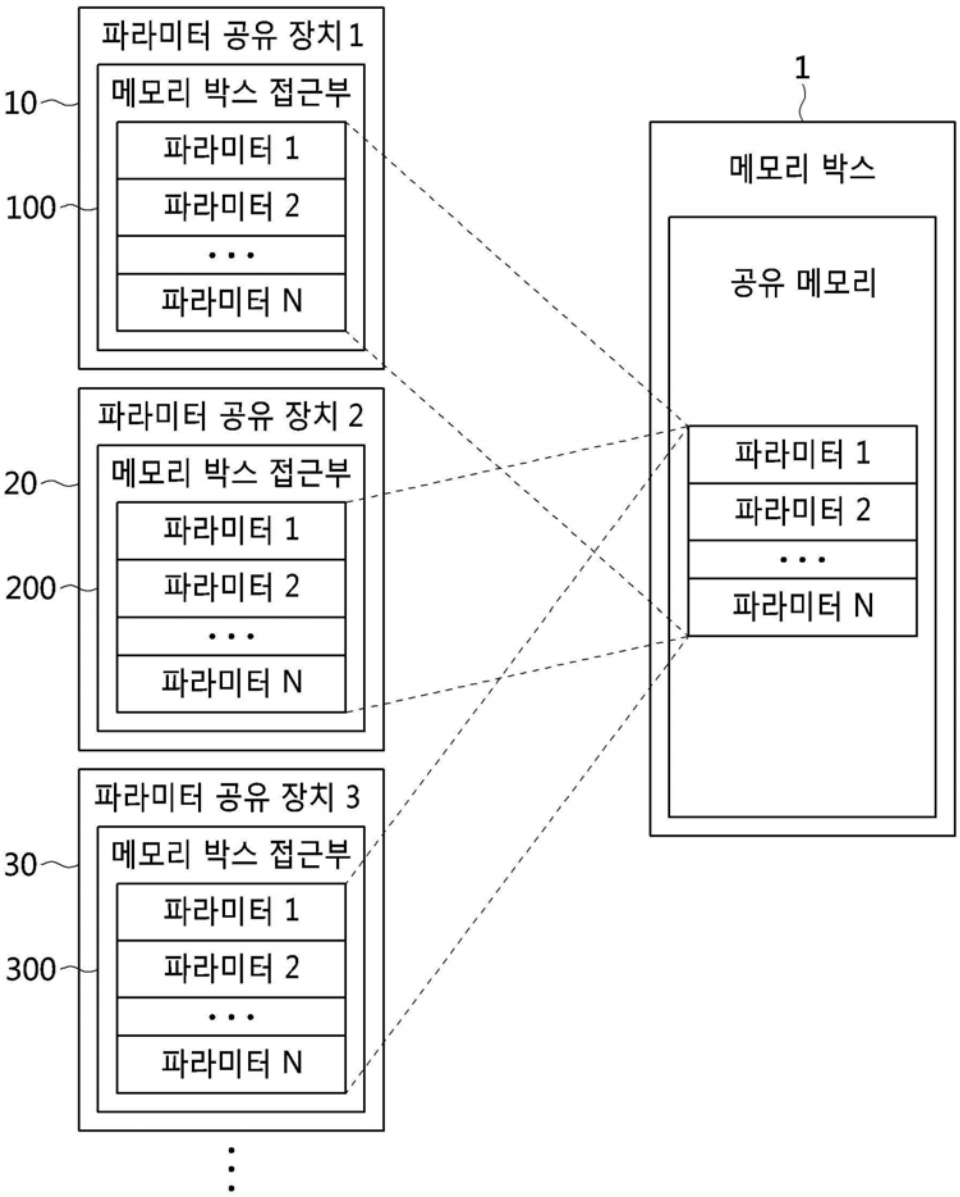
도면4


100
110
120
메모리 할당부
연산 처리부

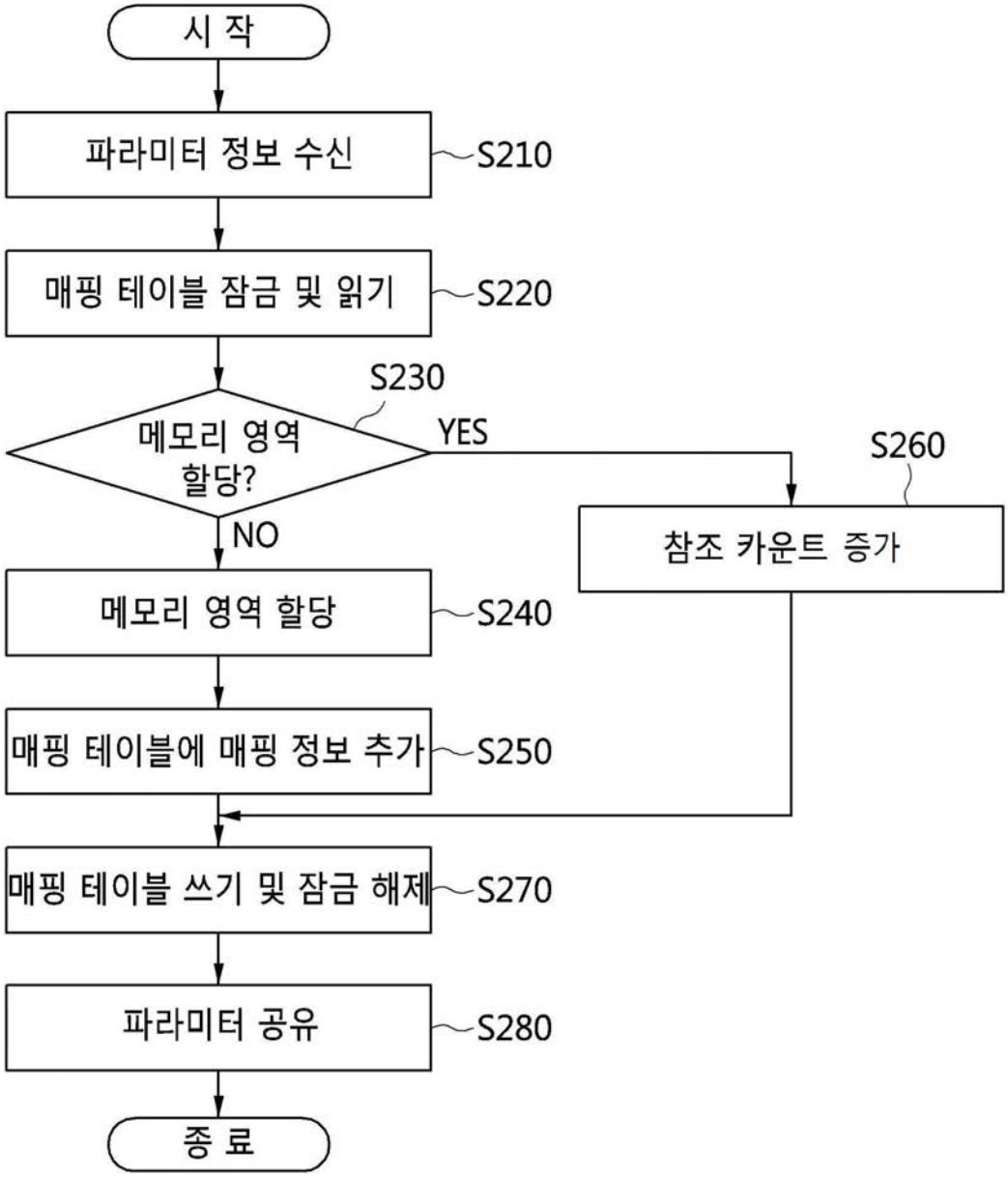
도면5


100
113
110
120
121
공유 메모리 관리부
파라미터 객체부
112
111
122
매핑 테이블 관리부
파라미터 정보 수신부
파라미터 처리부
1
매핑 테이블 저장부
파라미터 저장부
1a
1b

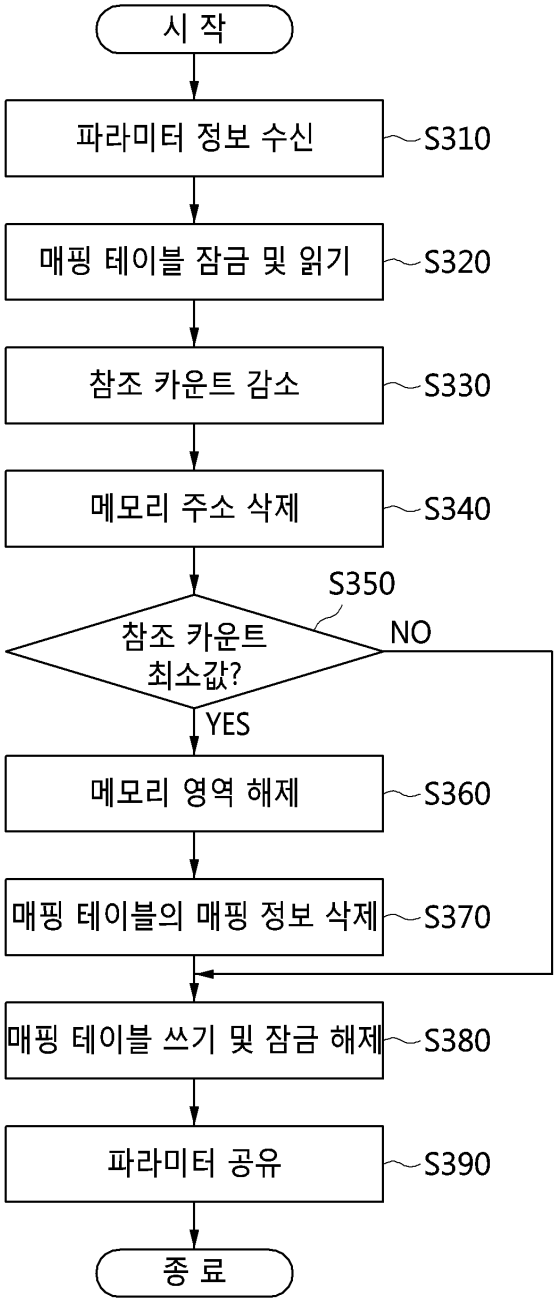
도면6


파라미터 공유 장치 1
10
메모리 박스 접근부
100
파라미터 1
파라미터 2
• • •
파라미터 N
파라미터 공유 장치 2
20
메모리 박스 접근부
200
파라미터 1
파라미터 2
• • •
파라미터 N
파라미터 공유 장치 3
30
메모리 박스 접근부
300
파라미터 1
파라미터 2
• • •
파라미터 N
1
메모리 박스
공유 메모리
파라미터 1
파라미터 2
• • •
파라미터 N

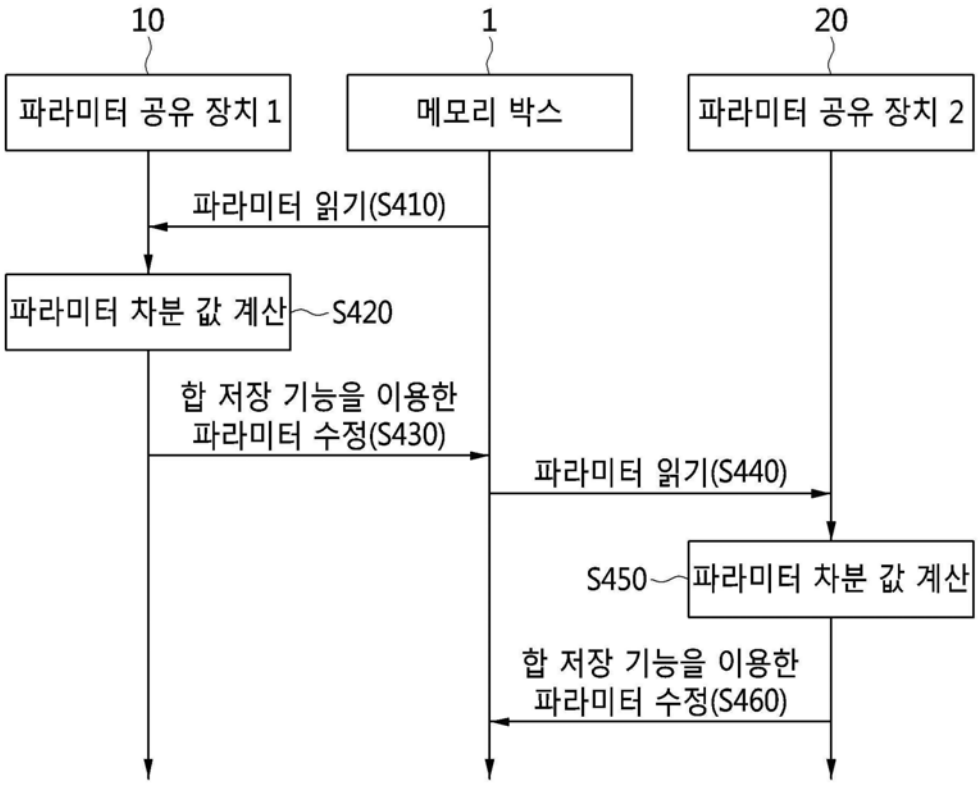
도면7


시 작
파라미터 정보 수신
S210
매핑 테이블 잠금 및 읽기
S220
S230
메모리 영역 할당?
YES
NO
S260
참조 카운트 증가
메모리 영역 할당
S240
매핑 테이블에 매핑 정보 추가
S250
매핑 테이블 쓰기 및 잠금 해제
S270
파라미터 공유
S280
종 료

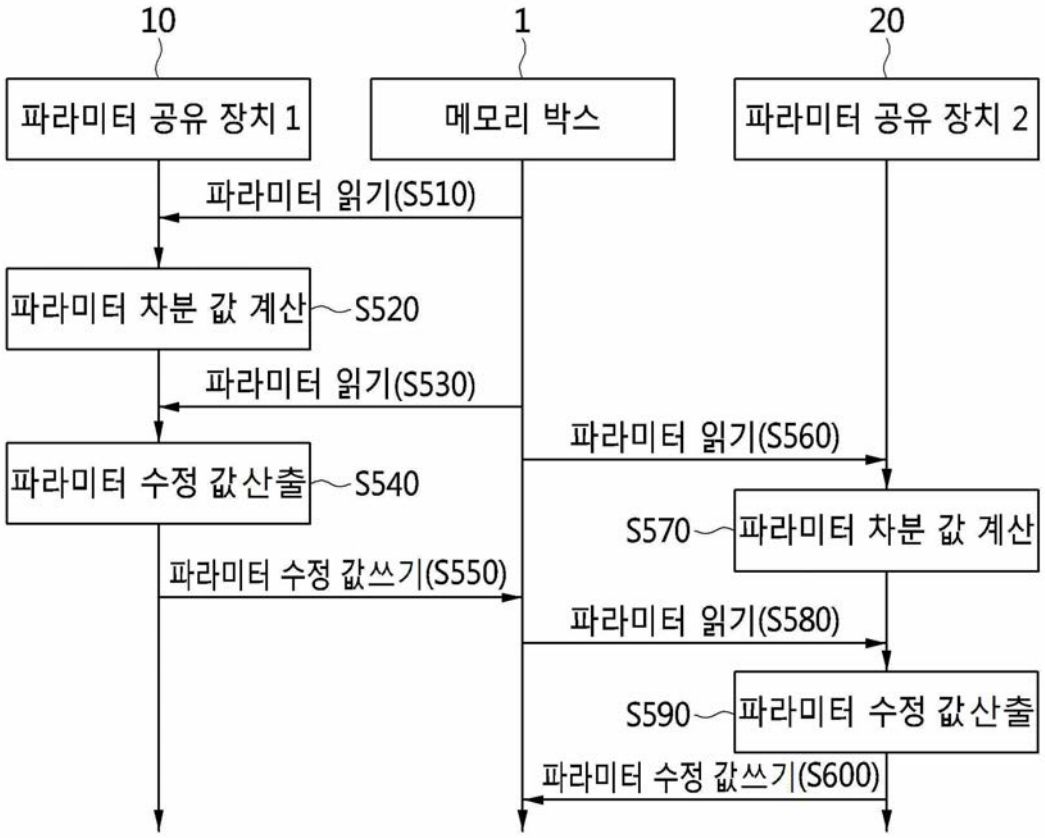
도면8


시 작
파라미터 정보 수신
S310
매핑 테이블 잠금 및 읽기
S320
참조 카운트 감소
S330
메모리 주소 삭제
S340
S350
참조 카운트
최소값?
NO
YES
메모리 영역 해제
S360
매핑 테이블의 매핑 정보 삭제
S370
매핑 테이블 쓰기 및 잠금 해제
S380
파라미터 공유
S390
종 료

도면9


10
1
20
파라미터 공유 장치 1
메모리 박스
파라미터 공유 장치 2
파라미터 읽기(S410)
파라미터 차분 값 계산
S420
합 저장 기능을 이용한
파라미터 수정(S430)
파라미터 읽기(S440)
S450
파라미터 차분 값 계산
합 저장 기능을 이용한
파라미터 수정(S460)

도면10


10
1
20
파라미터 공유 장치 1
메모리 박스
파라미터 공유 장치 2
파라미터 읽기(S510)
파라미터 차분 값 계산
S520
파라미터 읽기(S530)
파라미터 읽기(S560)
파라미터 수정 값산출
S540
S570
파라미터 차분 값 계산
파라미터 수정 값쓰기(S550)
파라미터 읽기(S580)
S590
파라미터 수정 값산출
파라미터 수정 값쓰기(S600)

(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년12월31일
(11) 등록번호 10-2197247
(24) 등록일자 2020년12월24일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) *G06N 3/04* (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2017-0068445
(22) 출원일자 2017년06월01일
심사청구일자 2019년03월14일
(65) 공개번호 10-2018-0131836
(43) 공개일자 2018년12월11일
(56) 선행기술조사문헌
KR1020180051987 A
(뒷면에 계속)
전체 청구항 수 : 총 16 항

(73) 특허권자
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
안신영
대전광역시 서구 둔산북로 160, 5동 701호
임은지
대전광역시 유성구 노은동로 187, 602동 1801호
(뒷면에 계속)
(74) 대리인
한양특허법인

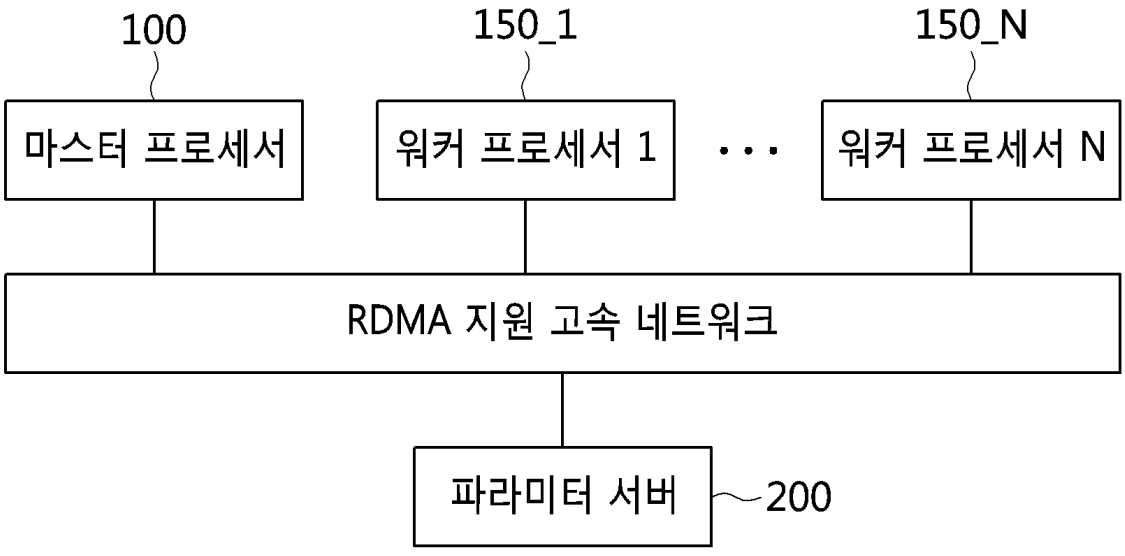
심사관 : 송근배

(54) 발명의 명칭 **파라미터 서버 및 그것에 의해 수행되는 분산 딥러닝 파라미터 공유 방법**

(57) 요 약

파라미터 서버 및 그것에 의해 수행되는 분산 딥러닝 파라미터 공유 방법이 개시된다. 본 발명에 따른 파라미터 서버에 의해 수행되는 분산 딥러닝 파라미터 공유 방법은, 마스터 프로세스의 초기화 요청에 상응하도록 전역 가중치 파라미터를 초기화하는 단계, 로컬 가중치 파라미터를 상기 전역 가중치 파라미터로 업데이트한 후 딥러닝 트레이닝을 수행한 상기 워커 프로세스로부터, 학습된 로컬 그래디언트 파라미터를 입력받아 업데이트하는 단계, 상기 마스터 프로세스의 요청에 따라, 그래디언트 파라미터 누적을 연산하는 단계, 그리고 상기 하나 이상의 워커 프로세스의 상기 그래디언트 파라미터 누적을 이용하여 전역 가중치 파라미터를 계산한 상기 마스터 프로세스로부터, 상기 전역 가중치 파라미터를 입력받아 업데이트하는 단계를 포함한다.

대 표 도 - 도1



(72) 발명자

최용석

전광역시 유성구 지족북로 60, 207동 303호

우영춘

대전광역시 유성구 어은로 57, 113동 404호

최완

대전광역시 서구 관저북로 52, 108동 306호

(56) 선행기술조사문헌

KR1020180125734 A

US20150324690 A1*

US20160103901 A1

US20170098171 A1

US20180218257 A1

Deep Image, Scaling up Image Recognition. Ren Wu et al. Baidu. 2015.02.06.*

*는 심사관에 의하여 인용된 문헌

이 발명을 지원한 국가연구개발사업

과제고유번호	R7117-16-0235
부처명	미래창조과학부
과제관리(전문)기관명	정보통신기술진흥센터(IITP)
연구사업명	정보통신방송기술개발사업(SW컴퓨팅 산업원천기술개발사업)
연구과제명	대규모 딥러닝 고속 처리를 위한 HPC 시스템 개발
기 여 율	1/1
과제수행기관명	한국전자통신연구원
연구기간	2016.04.01 ~ 2016.12.31

명 세 서

청구범위

청구항 1

파라미터 서버에 의해 수행되는 분산 딥러닝 파라미터 공유 방법에 있어서,

마스터 프로세스 및 적어도 하나의 워커 프로세스를 포함하는 분산 딥러닝 프로세스의 요청에 상응하도록, 원격 공유 메모리를 생성 및 할당하는 단계,

상기 원격 공유 메모리의 마스터 가중치 파라미터 영역을 초기화하는 단계,

상기 분산 딥러닝 프로세스들이 상기 원격 공유 메모리를 통해 공유한 분산 딥러닝 파라미터를 이용하여 분산 딥러닝 트레이닝을 수행하는 단계, 그리고

상기 분산 딥러닝 트레이닝의 수행이 완료된 후, 사용이 완료된 상기 원격 공유 메모리를 해제 및 삭제하는 단계를 포함하고,

상기 원격 공유 메모리를 생성 및 할당하는 단계는,

상기 마스터 프로세스의 요청에 따라 원격 공유 메모리를 생성하고, 상기 적어도 하나의 워커 프로세스에게 상기 원격 공유 메모리에 접근하기 위한 접근 정보를 전달하고,

상기 마스터 프로세스와 상기 적어도 하나의 워커 프로세스가, 상기 원격 공유 메모리에 상응하는 각각의 로컬 물리 메모리를 할당하고, 상기 각각의 로컬 물리 메모리를 분산 딥러닝 트레이닝 엔진의 가상 주소 공간에 맵핑하고,

상기 분산 딥러닝 트레이닝을 수행하는 단계는

상기 파라미터 서버, 상기 마스터 프로세스 및 상기 적어도 하나의 워커 프로세스가, 상기 각각의 로컬 물리 메모리와 상기 원격 공유 메모리의 명시적 동기화를 통해 상기 분산 딥러닝 파라미터를 공유하여 상기 분산 딥러닝 트레이닝을 수행하고,

상기 원격 공유 메모리는

상기 마스터 프로세스에 의해 생성되고, 마스터 가중치 파라미터와 마스터 그래디언트 파라미터를 저장하는 마스터 영역; 및

상기 적어도 하나의 워커 프로세스의 개수에 상응하도록 생성되고, 적어도 하나의 워커 그래디언트 파라미터를 각각 저장하는 적어도 하나의 워커 영역;

을 포함하고,

상기 분산 딥러닝 트레이닝을 수행하는 단계는

상기 마스터 프로세스가, 상기 마스터 영역에 상기 마스터 가중치 파라미터와 상기 마스터 그래디언트 파라미터를 업데이트하고, 상기 적어도 하나의 워커 프로세스에게 상기 마스터 영역에 접근하기 위한 접근 정보를 전송하고,

상기 적어도 하나의 워커 프로세스가, 상기 접근 정보를 이용하여 상기 마스터 영역에 접근하여 상기 마스터 가중치 파라미터를 자신의 워커 가중치 파라미터로 업데이트하고,

상기 원격 공유 메모리의 적어도 하나의 워커 영역 중 자신이 생성한 워커 영역에 상기 분산 딥러닝 트레이닝을 수행한 결과로부터 학습된 워커 그래디언트 파라미터를 업데이트하고,

상기 파라미터 서버가, 상기 적어도 하나의 워커 프로세스로부터 상기 적어도 하나의 워커 영역에 상기 적어도 하나의 워커 그래디언트 파라미터가 업데이트되었음을 알림받으면, 상기 적어도 하나의 워커 그래디언트 파라미터를 상기 마스터 그래디언트 파라미터에 업데이트하고,

상기 마스터 프로세스가, 업데이트된 상기 마스터 그래디언트 파라미터를 이용하여 상기 마스터 가중치 파라미

터를 업데이트하는 것을 특징으로 하는 분산 딥러닝 파라미터 공유 방법.

청구항 2

제1항에 있어서,

상기 원격 공유 메모리를 생성 및 할당하는 단계는,

상기 마스터 프로세스로부터 상기 분산 딥러닝 파라미터를 저장하기 위한 상기 원격 공유 메모리의 생성 요청을 수신하는 단계,

상기 원격 공유 메모리의 생성 요청에 상응하도록 상기 원격 공유 메모리를 생성하는 단계,

생성된 상기 원격 공유 메모리에 상응하는 원격 공유 메모리 생성키 및 상기 원격 공유 메모리에 접근하기 위한 접근 정보를 상기 마스터 프로세스로 전송하는 단계,

상기 마스터 프로세스로부터 상기 원격 공유 메모리에 할당된 원격 공유 메모리 영역에서의 상기 분산 딥러닝 파라미터의 업데이트 및 상기 분산 딥러닝 파라미터의 연산 완료 여부 중 적어도 하나와 관련된 이벤트를 설정하는 이벤트 설정 요청을 수신하여, 상기 원격 공유 메모리에 상기 이벤트를 설정하는 단계,

상기 마스터 프로세스로부터 상기 원격 공유 메모리 생성키를 전달받은 상기 워커 프로세스로부터, 상기 원격 공유 메모리의 할당 요청을 수신하는 단계, 그리고

상기 원격 공유 메모리의 할당 요청에 상응하도록 상기 원격 공유 메모리에 상기 분산 딥러닝 파리미터를 공유하기 위한 원격 공유 메모리 영역을 할당하며, 할당된 상기 원격 공유 메모리의 원격 공유 메모리 영역에 접근하기 위한 접근 정보를 상기 워커 프로세스로 전송하는 단계를 포함하는 것을 특징으로 하는 분산 딥러닝 파라미터 공유 방법.

청구항 3

제1항에 있어서,

상기 원격 공유 메모리를 해제 및 삭제하는 단계는,

상기 워커 프로세스로부터 원격 공유 메모리 해제 요청을 수신하여, 상기 원격 공유 메모리를 해제하는 단계,

상기 원격 공유 메모리의 해제 완료 시 상기 마스터 프로세스로부터, 원격 공유 메모리 삭제 요청을 수신하는 단계, 그리고

상기 원격 공유 메모리 삭제 요청에 상응하도록 상기 원격 공유 메모리를 삭제하는 단계를 더 포함하는 것을 특징으로 하는 분산 딥러닝 파라미터 공유 방법.

청구항 4

제1항에 있어서,

상기 분산 딥러닝 트레이닝을 수행하도록 하는 단계는,

상기 분산 딥러닝 프로세스들이 상기 원격 공유 메모리를 통하여 동기식 또는 비동기식으로 업데이트된 딥러닝 파라미터를 공유하는 것을 특징으로 하는 분산 딥러닝 파라미터 공유 방법.

청구항 5

제4항에 있어서,

상기 동기식으로 업데이트된 상기 딥러닝 파라미터를 공유하여, 상기 분산 딥러닝 트레이닝을 수행하도록 하는 단계는,

상기 분산 딥러닝 프로세스들의 워커 로컬 가중치 파라미터 영역을 상기 원격 공유 메모리의 상기 마스터 가중치 파라미터의 값으로 업데이트하는 단계,

동기 방식으로 분산 딥러닝 트레이닝을 수행한 상기 워커 프로세스로부터, 학습된 워커 로컬 그래디언트 파라미터를 입력받아 그래디언트 파라미터 누적을 연산하는 단계,

상기 하나 이상의 워커 프로세스의 상기 그래디언트 파라미터 누적을 이용하여 마스터 가중치 파라미터를 계산한 상기 마스터 프로세스로부터, 상기 마스터 가중치 파라미터를 입력받아 상기 마스터 가중치 파라미터 영역을 업데이트하는 단계, 그리고

상기 마스터 가중치 파라미터 영역의 업데이트를 적어도 어느 하나의 상기 워커 프로세스로 알리는 단계를 포함하는 것을 특징으로 하는 분산 딥러닝 파라미터 공유 방법.

청구항 6

제5항에 있어서,

상기 그래디언트 파라미터 누적을 연산하는 단계는,

상기 분산 딥러닝 트레이닝을 수행한 상기 워커 프로세스들이 학습한 워커 로컬 그래디언트 파라미터를 상기 원격 공유 메모리의 워커 그래디언트 파라미터 영역에 저장하는 단계,

상기 워커 프로세스들로부터 그래디언트 파라미터 누적 연산을 요청받는 단계,

요청에 상응하는 상기 원격 공유 메모리의 상기 워커 그래디언트 파라미터를 마스터 그래디언트 파라미터에 누적 연산하는 단계, 그리고

상기 누적 연산의 완료를 상기 마스터 프로세스로 알리는 단계

를 포함하는 것을 특징으로 하는 분산 딥러닝 파라미터 공유 방법.

청구항 7

제4항에 있어서,

상기 비동기식으로 업데이트된 상기 딥러닝 파라미터를 공유하여, 상기 분산 딥러닝 트레이닝을 수행하도록 하는 단계는,

하나 이상의 상기 워커 프로세스의 워커 로컬 가중치 파라미터 영역을 상기 원격 공유 메모리의 상기 마스터 가중치 파라미터의 값으로 업데이트 하는 단계,

상기 분산 딥러닝 트레이닝을 수행한 상기 하나 이상의 워커 프로세스들이 원격 공유 메모리 상의 워커 그래디언트 파라미터를 업데이트하는 단계,

상기 하나 이상의 워커 프로세스로부터 수신한 마스터 가중치 파라미터의 업데이트 요청에 상응하도록, 상기 마스터 가중치 파라미터 영역을 업데이트하는 단계, 그리고

상기 분산 딥러닝 트레이닝의 수행이 완료된 후, 사용이 완료된 상기 원격 공유 메모리를 해제 및 삭제하는 단계를 포함하는 것을 특징으로 하는 분산 딥러닝 파라미터 공유 방법.

청구항 8

마스터 프로세스 및 적어도 하나의 워커 프로세스를 포함하는 분산 딥러닝 프로세스의 요청과 관련된 메시지를 송수신하는 통신 처리부,

상기 분산 딥러닝 프로세스의 요청에 상응하도록, 분산 딥러닝 파라미터를 저장하기 위한 원격 공유 메모리를 생성, 할당 및 해제하는 원격 공유 메모리 관리부, 그리고

상기 분산 딥러닝 프로세스가 상기 원격 공유 메모리를 통해 공유한 분산 딥러닝 파라미터를 이용하여 분산 딥러닝 트레이닝을 수행하는 파라미터 연산부를 포함하고,

상기 원격 공유 메모리 관리부는

상기 마스터 프로세스의 요청에 따라 상기 원격 공유 메모리를 생성하고, 상기 적어도 하나의 워커 프로세스에게 상기 원격 공유 메모리에 접근하기 위한 접근 정보를 전달하고,

상기 마스터 프로세스와 상기 적어도 하나의 워커 프로세스는

상기 원격 공유 메모리에 상응하는 각각의 로컬 물리 메모리를 할당하고, 상기 각각의 로컬 물리 메모리를 분산

딥러닝 트레이닝 엔진의 가상 주소 공간에 맵핑하고,

상기 파라미터 연산부는

상기 마스터 프로세스 및 상기 적어도 하나의 워커 프로세스와 함께, 상기 각각의 로컬 물리 메모리와 상기 원격 공유 메모리의 명시적 동기화를 통해 상기 분산 딥러닝 파라미터를 공유하여 상기 분산 딥러닝 트레이닝을 수행하고,

상기 원격 공유 메모리는

상기 마스터 프로세스에 의해 생성되고, 마스터 가중치 파라미터와 마스터 그래디언트 파라미터를 저장하는 마스터 영역; 및

상기 적어도 하나의 워커 프로세스의 개수에 상응하도록 생성되고, 적어도 하나의 워커 그래디언트 파라미터를 각각 저장하는 적어도 하나의 워커 영역;

을 포함하고,

상기 마스터 프로세스는

상기 마스터 영역에 상기 마스터 가중치 파라미터와 상기 마스터 그래디언트 파라미터를 업데이트하고,

상기 적어도 하나의 워커 프로세스에게 상기 마스터 영역에 접근하기 위한 접근 정보를 전송하고,

상기 적어도 하나의 워커 프로세스는

상기 접근 정보를 이용하여 상기 마스터 영역에 접근하여 상기 마스터 가중치 파라미터를 자신의 워커 가중치 파라미터로 업데이트하고,

상기 원격 공유 메모리의 적어도 하나의 워커 영역 중 자신이 생성한 워커 영역에 상기 분산 딥러닝 트레이닝을 수행한 결과로부터 학습된 워커 그래디언트 파라미터를 업데이트하고,

상기 파라미터 연산부는

상기 적어도 하나의 워커 프로세스로부터 상기 적어도 하나의 워커 영역에 상기 적어도 하나의 워커 그래디언트 파라미터가 업데이트되었음을 알림받으면, 상기 적어도 하나의 워커 그래디언트 파라미터를 상기 마스터 그래디언트 파라미터에 누적 연산하여 업데이트하고,

상기 마스터 프로세스는

업데이트된 상기 마스터 그래디언트 파라미터를 이용하여 상기 마스터 가중치 파라미터를 업데이트하는 것을 특징으로 하는 파라미터 서버.

청구항 9

제8항에 있어서,

상기 파라미터 연산부는,

두 개의 원격 공유 메모리 영역에 대한 벡터 연산을 수행하는 것을 특징으로 하는 파라미터 서버.

청구항 10

제9항에 있어서,

상기 파라미터 연산부는,

제1 벡터에 제1 상수를 곱하는 연산, 상기 제1 상수를 곱한 제1 벡터와 제2 벡터를 합하는 연산 및 상기 제1 상수를 곱한 상기 제1 벡터와 제2 상수를 곱한 상기 제2 벡터를 합하는 연산 중 적어도 어느 하나의 상기 벡터 연산을 수행하는 것을 특징으로 하는 파라미터 서버.

청구항 11

제8항에 있어서,

상기 파라미터 연산부는,

가중치 파라미터 및 그래디언트 파라미터 중 적어도 어느 하나를 포함하는 상기 분산 딥러닝 파라미터를 연산하는 것을 특징으로 하는 파라미터 서버.

청구항 12

제11항에 있어서,

상기 마스터 프로세스는,

상기 마스터 프로세스가 할당한 모든 상기 원격 공유 메모리의 영역에 접근 가능하고,

상기 워커 프로세스는,

마스터 파라미터 영역 및 상기 워커 프로세스가 딥러닝 트레이닝을 수행한 결과를 저장하는 워커 파라미터 영역만 접근 가능한 것을 특징으로 하는 파라미터 서버.

청구항 13

제12항에 있어서,

상기 파라미터 연산부는,

동기식으로 상기 분산 딥러닝 파라미터를 공유하는 경우, 상기 그래디언트 파라미터 누적을 연산하는 것을 특징으로 하는 파라미터 서버.

청구항 14

제12항에 있어서,

상기 파라미터 연산부는,

비동기식으로 상기 분산 딥러닝 파라미터를 공유하는 경우, 상기 워커 프로세스로부터 수신한 워커 그래디언트 파라미터를 마스터 가중치 파라미터 영역에 업데이트하는 것을 특징으로 하는 파라미터 서버.

청구항 15

삭제

청구항 16

삭제

청구항 17

제8항에 있어서,

상기 원격 공유 메모리 관리부는,

상기 워커 프로세스로부터 수신한 원격 공유 메모리 해제 요청에 상응하도록 상기 원격 공유 메모리를 해제하고, 상기 마스터 프로세스로부터 수신한 원격 공유 메모리 삭제 요청에 상응하도록 상기 원격 공유 메모리를 삭제하는 것을 특징으로 하는 파라미터 서버.

청구항 18

제8항에 있어서,

상기 마스터 프로세스 및 워커 프로세스는,

원격 직접 메모리 접근(RDMA)을 지원하는 고속 네트워크를 통하여, 상기 파라미터 서버에 저장한 상기 분산 딥러닝 파라미터를 직접 읽어오거나 쓰는 방식으로 상기 분산 딥러닝 파라미터를 공유하는 것을 특징으로 하는 파라미터 서버.

발명의 설명

기 술 분 야

[0001] 본 발명은 분산 딥러닝 프레임워크에서 분산 트레이닝되는 파라미터를 공유하는 기술에 관한 것으로, 특히 분산 딥러닝 프로세스들이 파라미터 서버의 물리 메모리를 공유 메모리 형태로 접근할 수 있도록 하여, 딥러닝 프로세스들 간 파라미터 공유를 가속화하는 기술에 관한 것이다.

배 경 기 술

[0002] 딥러닝이란 사람의 신경세포(Biological Neuron)를 모사하여 기계가 학습하도록 하는 인공신경망(Artificial Neural Network) 기반의 기계 학습법을 의미한다. 최근, 딥러닝 기술은 이미지 인식, 음성 인식, 자연어 처리의 발전에 기여하며 주목 받고 있다. 그리고 오늘날의 딥러닝 모델들은 응용의 인식 성능을 높이기 위해 모델의 계층이 깊어지고(Deep), 특징(Feature)이 많아지는(Wide) 대규모 모델로 진화하고 있다.

[0003] 그러나 대형화되는 딥러닝 모델과 대규모의 학습 데이터를 단일 머신에서 처리하기에는 한계가 있다. 이에, 대규모 분산 컴퓨팅 자원을 활용하려는 노력의 일환으로 딥러닝 분산 플랫폼 기술이 개발되고 있다.

[0004] 딥러닝 분산 플랫폼에서는 분산 병렬 처리를 통하여 딥러닝 트레이닝 가속을 시도하는데, 분산 병렬 처리 방법으로 데이터 병렬 처리(Data Parallelism)와 모델 병렬 처리(Model Parallelism) 방법이 있다. 데이터 병렬 처리란 학습해야 하는 입력 데이터 집합을 다수의 컴퓨터들이 나누어 트레이닝하는 방법이고, 모델 병렬 처리란 딥러닝 모델을 나누어 다수의 컴퓨터들이 트레이닝하는 방법이다.

[0005] 딥러닝 트레이닝 분산 병렬 처리 시에는 트레이닝의 대상이 되는 가중치와 특징값 등의 파라미터들을 모든 컴퓨터가 공유해야 한다. 파라미터를 공유하는 방법에는 각 컴퓨터들이 다른 모든 컴퓨터들에게 직접 파라미터를 전달하는 풀 메시(full mesh) 토폴로지 기반 공유 방법과 모든 분산 컴퓨터들이 공유 장소를 이용하여 파라미터를 읽고 쓰는 스타(star) 토폴로지 기반의 공유 방법이 있다. 그리고 대부분의 분산 플랫폼은 일반적으로 중앙 집중형 파라미터 공유 저장소(파라미터 서버)를 통해 파라미터를 교환하는 두 번째 방식을 선택하고 있다.

[0006] 파라미터 공유 방법에서는 분산된 컴퓨터들이 각각 파라미터를 중앙 집중형으로 업데이트 하기 때문에 가중치를 갱신해야 하는 주기(일정 트레이닝 반복)마다 분산 트레이닝 중인 컴퓨터 간 파라미터 동기화가 필요하다. 동기식 업데이트의 경우는 딥러닝을 분산 처리하는 컴퓨터들의 일정 트레이닝 반복마다 파라미터를 파라미터 서버로 전송하여 분산 트레이닝된 파라미터를 통합한다.

[0007] 반면, 비동기식 업데이트 방식은 파라미터 서버가 분산 컴퓨터들로부터 늦거나 빨리 도착하는 파라미터들의 동기를 맞추지 않고 트레이닝을 진행하는 방법이다. 비동기 방식은 동기식에 비해 정확성을 크게 희생시키지 않으면서 빠르게 트레이닝 할 수 있는 장점이 있다. 대부분의 분산 프레임워크들에서는 동기식 방법과 비동기식 방법을 모두 또는 선택적으로 제공하고 있다.

[0008] 각 딥러닝 분산 플랫폼에서 파라미터 서버를 구현하는 방법으로는 마스터 역할을 하는 프로세스가 자신의 메모리에 마스터 파라미터를 저장하는 영역을 할당한다. 그리고 분산 트레이닝을 수행하는 워커(또는 슬레이브) 프로세스들이 통신 메시지 형태로 주기적으로 보내주는 파라미터로 마스터 파라미터를 업데이트 한 후 다시 워커 프로세스들에 업데이트된 마스터 파라미터를 배포하는 방식이 있다. Petuum, CNTK와 같은 분산 플랫폼은 파라미터 서버 전용 목적으로 개발된 분산 키-밸류 저장소를 이용하기도 한다.

[0009] 종래 기술에 따르면, 파라미터 서버와 분산 컴퓨터간에 메시지 송수신 형태의 통신방법을 통해 파라미터를 교환한다. 그러나, 메시지 송수신 형태의 통신 방법으로 파라미터를 교환할 경우, 통신 오버헤드가 높고, CPU, GPU 등이 대기하는 시간이 길어지며, 이는 자원 사용률의 저하로 이어진다.

[0010] 따라서, 통신 프로토콜로 대규모 파라미터를 송수신하는 기술의 한계를 극복하여, 추가적인 메모리 복사 및 프로토콜 처리 등의 통신 오버헤드를 대폭 경감하고, 통신 성능을 개선할 수 있는 파라미터 공유 기술의 개발이 필요하다.

선행기술문헌

특허문헌

[0011] (특허문헌 0001) 한국 등록 특허 제10-1559089호, 2015년 10월 02일 공개(명칭: 장치의 컴포넌트들 간에 메모리

자원들을 공유하기 위한 통신 프로토콜)

발명의 내용

해결하려는 과제

[0012] 본 발명의 목적은 분산 딥러닝 플랫폼에서 분산 트레이닝을 수행하는 프로세스들이 대규모 파라미터를 교환할 수 있도록 하는 것이다.

[0013] 또한, 본 발명의 목적은 파라미터 서버와 분산 컴퓨터가 메시지 송수신 형태의 통신 방법으로 파라미터를 교환할 경우 발생하는 추가적인 메모리 복사 및 통신 오버헤드를 대폭 경감하는 것이다.

[0014] 또한, 본 발명의 목적은 메시지 송수신 형태의 통신 방법으로 파라미터를 교환하는 방법에 비하여, 통신 성능을 개선하고, 파라미터 송수신 시 유휴 상태인 계산 자원의 이용률을 극대화하는 것이다.

과제의 해결 수단

[0015] 상기한 목적을 달성하기 위한 본 발명에 따른 파라미터 서버에 의해 수행되는 분산 딥러닝 파라미터 공유 방법은 마스터 프로세스 및 워커 프로세스 중 적어도 어느 하나를 포함하는 분산 딥러닝 프로세스의 요청에 상응하도록, 공유 메모리를 생성 및 할당하는 단계, 상기 공유 메모리의 마스터 가중치 파라미터 영역을 초기화하는 단계, 상기 분산 딥러닝 프로세스들이 상기 공유 메모리를 통해 공유한 딥러닝 파라미터를 이용하여, 분산 딥러닝 트레이닝을 수행하도록 하는 단계, 그리고 상기 분산 딥러닝 트레이닝의 수행이 완료된 후, 사용이 완료된 상기 공유 메모리를 해제 및 삭제하는 단계를 포함한다.

[0016] 이때, 상기 공유 메모리를 생성 및 할당하는 단계는, 상기 마스터 프로세스로부터 파라미터용 원격 공유 메모리 생성 요청을 수신하는 단계, 상기 파라미터용 원격 공유 메모리 생성 요청에 상응하도록 공유 메모리를 생성하는 단계, 생성된 상기 공유 메모리에 상응하는 공유 메모리 생성키 및 접근 정보를 상기 마스터 프로세스로 전송하는 단계, 상기 마스터 프로세스로부터 이벤트 설정 요청을 수신하여, 상기 공유 메모리의 이벤트를 설정하는 단계, 상기 마스터 프로세스로부터 상기 공유 메모리 생성키를 전달받은 상기 워커 프로세스로부터, 공유 메모리 할당 요청을 수신하는 단계, 그리고 상기 공유 메모리를 할당하고, 할당된 상기 공유 메모리의 접근 정보를 상기 워커 프로세스로 전송하는 단계를 더 포함할 수 있다.

[0017] 이때, 상기 공유 메모리를 해제 및 삭제하는 단계는, 상기 워커 프로세스로부터 공유 메모리 해제 요청을 수신하여, 상기 공유 메모리를 해제하는 단계, 상기 공유 메모리의 해제 완료 시 상기 마스터 프로세스로부터, 공유 메모리 삭제 요청을 수신하는 단계, 그리고 상기 공유 메모리 삭제 요청에 상응하도록 상기 공유 메모리를 삭제하는 단계를 더 포함할 수 있다.

[0018] 이때, 상기 분산 딥러닝 트레이닝을 수행하도록 하는 단계는, 상기 분산 딥러닝 프로세스들이 상기 공유 메모리를 통하여 동기식 또는 비동기식으로 업데이트된 가중치 파라미터를 공유할 수 있다.

[0019] 이때, 상기 분산 딥러닝 프로세스들이 상기 공유 메모리를 통해 공유한 딥러닝 파라미터를 이용하여, 동기식 분산 딥러닝 트레이닝을 수행하도록 하는 단계는, 상기 분산 딥러닝 프로세스들의 워커 로컬 가중치 파라미터 영역을 상기 공유 메모리의 상기 마스터 가중치 파라미터의 값으로 업데이트하는 단계, 동기 방식으로 분산 딥러닝 트레이닝을 수행한 상기 워커 프로세스로부터, 학습된 워커 로컬 그래디언트 파라미터를 입력받아 그래디언트 파라미터 누적을 연산하는 단계, 상기 하나 이상의 워커 프로세스의 상기 그래디언트 파라미터 누적을 이용하여 마스터 가중치 파라미터를 계산한 상기 마스터 프로세스로부터, 상기 마스터 가중치 파라미터를 입력받아 상기 마스터 가중치 파라미터 영역을 업데이트하는 단계, 그리고 상기 마스터 가중치 파라미터 영역의 업데이트를 적어도 어느 하나의 상기 워커 프로세스로 알리는 단계를 포함할 수 있다.

[0020] 이때, 상기 그래디언트 파라미터 누적을 연산하는 단계는, 상기 분산 딥러닝 트레이닝을 수행한 상기 워커 프로세스들이 학습한 워커 로컬 그래디언트 파라미터를 상기 공유 메모리의 워커 그래디언트 파라미터 영역에 저장하는 단계, 상기 워커 프로세스들로부터 그래디언트 파라미터 누적 연산을 요청받는 단계, 요청에 상응하는 상기 공유 메모리의 상기 워커 그래디언트 파라미터를 마스터 그래디언트 파라미터에 누적 연산하는 단계, 그리고 상기 누적 연산의 완료를 상기 마스터 프로세스로 알리는 단계를 포함한다.

[0021] 이때, 상기 분산 딥러닝 프로세스들이 상기 공유 메모리를 통해 공유한 딥러닝 파라미터를 이용하여, 비동기식

분산 딥러닝 트레이닝을 수행하도록 하는 단계는, 하나 이상의 상기 워커 프로세스의 워커 로컬 가중치 파라미터 영역을 상기 공유 메모리의 상기 마스터 가중치 파라미터의 값으로 업데이트 하는 단계, 상기 분산 딥러닝 트레이닝을 수행한 상기 하나 이상의 워커 프로세스들이 공유 메모리 상의 워커 그래디언트 파라미터를 업데이트하는 단계, 상기 하나 이상의 워커 프로세스로부터 수신한 마스터 가중치 파라미터의 업데이트 요청에 상응하도록, 상기 마스터 가중치 파라미터 영역을 업데이트하는 단계, 그리고 상기 분산 딥러닝 트레이닝의 수행이 완료된 후, 사용이 완료된 상기 공유 메모리를 해제 및 삭제하는 단계를 포함한다.

[0022] 또한, 본 발명의 일실시예에 따른 파라미터 서버는 마스터 프로세스 및 워커 프로세스 중 적어도 어느 하나와 메시지를 송수신하고, 원격 직접 메모리 접근(RDMA) 방식의 읽기 및 쓰기를 지원하는 통신 처리부, 공유 메모리의 할당 및 해제를 관리하는 공유 메모리 관리부, 분산 딥러닝 파라미터를 계산하는 파라미터 연산부, 그리고 상기 공유 메모리에 대한 이벤트 발생 시, 상기 공유 메모리에 상응하는 상기 마스터 프로세스 및 하나 이상의 상기 워커 프로세스 중 적어도 어느 하나로 상기 이벤트의 발생을 알리는 이벤트 처리부를 포함한다.

[0023] 이때, 상기 파라미터 연산부는, 두 개의 공유 메모리 영역에 대한 벡터/매트릭스 연산을 수행할 수 있다.

[0024] 이때, 상기 파라미터 연산부는, 제1 벡터에 제1 상수를 곱하는 연산, 상기 제1 상수를 곱한 제1 벡터와 제2 벡터를 합하는 연산 및 상기 제1 상수를 곱한 상기 제1 벡터와 제2 상수를 곱한 상기 제2 벡터를 합하는 연산 중 적어도 어느 하나의 상기 벡터 연산을 수행할 수 있다.

[0025] 이때, 상기 파라미터 연산부는, 가중치 파라미터 및 그래디언트 파라미터 중 적어도 어느 하나를 포함하는 상기 분산 딥러닝 파라미터를 연산할 수 있다.

[0026] 이때, 상기 마스터 프로세스는, 상기 마스터 프로세스가 할당한 모든 상기 공유 메모리의 영역에 접근 가능하고, 상기 워커 프로세스는, 마스터 파라미터 영역 및 상기 워커 프로세스가 딥러닝 트레이닝을 수행한 결과를 저장하는 워커 파라미터 영역만 접근 가능할 수 있다.

[0027] 이때, 상기 파라미터 연산부는, 동기식으로 상기 분산 딥러닝 파라미터를 공유하는 경우, 상기 그래디언트 파라미터 누적을 연산할 수 있다.

[0028] 이때, 상기 파라미터 연산부는, 비동기식으로 상기 분산 딥러닝 파라미터를 공유하는 경우, 상기 워커 프로세스로부터 수신한 워커 그래디언트 파라미터를 마스터 가중치 파라미터 영역에 업데이트할 수 있다.

[0029] 이때, 상기 공유 메모리 관리부는, 상기 마스터 프로세스로부터 수신한 파라미터용 원격 공유 메모리 생성 요청에 상응하도록 공유 메모리를 생성하고, 상기 공유 메모리의 공유 메모리 생성키 및 접근 정보를 상기 마스터 프로세스로 전송할 수 있다.

[0030] 이때, 상기 공유 메모리 관리부는, 상기 마스터 프로세스로부터 상기 공유 메모리 생성키를 전달받은 상기 워커 프로세스로부터 공유 메모리 할당 요청을 수신하고, 상기 공유 메모리 할당 요청에 상응하도록 상기 공유 메모리를 할당하며, 할당된 상기 공유 메모리의 접근 정보를 상기 워커 프로세스로 전송할 수 있다.

[0031] 이때, 상기 공유 메모리 관리부는, 상기 워커 프로세스로부터 수신한 공유 메모리 해제 요청에 상응하도록 상기 공유 메모리를 해제하고, 상기 마스터 프로세스로부터 수신한 공유 메모리 삭제 요청에 상응하도록 상기 공유 메모리를 삭제할 수 있다.

[0032] 이때, 상기 마스터 프로세스 및 워커 프로세스는, 상기 원격 직접 메모리 접근(RDMA)을 지원하는 고속 네트워크를 통하여, 상기 파라미터 서버에 저장한 상기 분산 딥러닝 파라미터를 직접 읽어오거나 쓰는 방식으로 상기 분산 딥러닝 파라미터를 공유할 수 있다.

발명의 효과

[0033] 본 발명에 따르면, 분산 딥러닝 플랫폼에서 분산 트레이닝을 수행하는 프로세스들이 대규모 파라미터를 교환할 수 있다.

[0034] 또한 본 발명에 따르면, 파라미터 서버와 분산 컴퓨터가 메시지 송수신 형태의 통신 방법으로 파라미터를 교환할 경우 발생하는 추가적인 메모리 복사 및 통신 오버헤드를 대폭 경감할 수 있다.

[0035] 또한 본 발명에 따르면, 메시지 송수신 형태의 통신 방법으로 파라미터를 교환하는 방법에 비하여, 통신 성능을 개선하고, 파라미터 송수신 시 유휴 상태인 계산 자원의 이용률을 극대화할 수 있다.

도면의 간단한 설명

도 1은 본 발명의 일실시예에 따른 파라미터 서버가 적용되는 분산 딥러닝 프레임워크 환경을 개략적으로 나타낸 도면이다.

도 2는 본 발명의 일실시예에 따른 파라미터 서버의 구성을 나타낸 블록도이다.

도 3은 본 발명의 일실시예에 따른 파라미터 공유를 위한 원격 공유 메모리 가상 매핑 매커니즘을 나타낸 예시도이다.

도 4는 본 발명의 일실시예에 따른 분산 딥러닝 프레임워크의 기능을 설명하기 위한 구조도이다.

도 5는 본 발명의 일실시예에 따른 프로세스별 원격 공유 메모리 할당의 일 예를 나타낸 예시도이다.

도 6은 본 발명의 일실시예에 따른 분산 딥러닝 파라미터 공유 방법을 나타낸 순서도이다.

도 7은 본 발명의 일실시예에 따른 원격 공유 메모리의 생성 및 할당 과정을 나타낸 순서도이다.

도 8은 본 발명의 일실시예에 따른 원격 공유 메모리의 삭제 및 해제 과정을 나타낸 순서도이다.

도 9는 본 발명의 일실시예에 따른 동기식 파라미터 공유 방법을 설명하기 위한 순서도이다.

도 10은 본 발명의 일실시예에 따른 비동기식 파라미터 공유 방법을 설명하기 위한 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0037] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시 예들을 도면에 예시하고 상세하게 설명하고자 한다.

[0038] 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0039] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0040] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가진 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0041] 이하, 첨부한 도면들을 참조하여, 본 발명의 바람직한 실시예를 보다 상세하게 설명하고자 한다. 본 발명을 설명함에 있어 전체적인 이해를 용이하게 하기 위하여 도면상의 동일한 구성요소에 대해서는 동일한 참조부호를 사용하고 동일한 구성요소에 대해서 중복된 설명은 생략한다.

[0043] 도 1은 본 발명의 일실시예에 따른 파라미터 서버가 적용되는 분산 딥러닝 프레임워크 환경을 개략적으로 나타낸 도면이다.

[0044] 도 1에 도시한 바와 같이, 딥러닝 트레이닝을 수행하는 분산된 계산 노드에서 실행되는 분산 딥러닝 프로세스는 마스터 프로세스(100) 및 하나 이상의 워커 프로세스(150)를 포함하고, 마스터 프로세스(100), 워커 프로세스(150) 및 파라미터 서버(200)는 원격 직접 메모리 접근(RDMA, Remote Direct Memory Access)을 지원하는 고속 네트워크로 연결된다.

[0045] 마스터 프로세스(100)는 파라미터 서버(200)에 원격 공유 메모리를 생성하고, 분산 딥러닝 프레임워크의 전체적인 제어를 담당한다. 그리고 마스터 프로세스(100)는 워커 프로세스들(150)에 원격 공유 메모리 정보를 전달하여, 워커 프로세스들(150)이 파라미터 서버(200) 상의 동일 메모리 영역에 접근할 수 있도록 한다. 반면, 워크 프로세스들(150)은 분산 트레이닝을 수행하고, 트레이닝한 결과를 저장한다.

[0046] 파라미터 서버(200)는 가중치(Weight) 파라미터 및 그래디언트(Gradient) 파라미터 중 적어도 어느 하나를 포함하는 분산 딥러닝 파라미터를 공유하기 위한 공유 메모리를 제공한다. 그리고 파라미터 서버(200)는 분산 딥러닝 프로세스들(100, 150)이 공유 메모리를 통해 공유한 딥러닝 파라미터를 이용하여, 분산 딥러닝 트레이닝을 수행하도록 한다.

[0048] 이하에서는 도 2를 통하여, 본 발명의 일실시예에 따른 파라미터 서버의 구성 및 기능에 대하여 더욱 상세하게 설명한다.

[0049] 도 2는 본 발명의 일실시예에 따른 파라미터 서버의 구성을 나타낸 블록도이다.

[0050] 도 2에 도시한 바와 같이, 파라미터 서버(200)는 통신 처리부(210), 공유 메모리 관리부(220), 파라미터 연산부(230) 및 이벤트 처리부(240)를 포함한다.

[0051] 먼저, 통신 처리부(210)는 마스터 프로세스 및 하나 이상의 워커 프로세스 중 적어도 어느 하나의 분산 딥러닝 트레이닝 엔진과 메시지를 송수신한다. 그리고 통신 처리부(210)는 마스터 프로세스 및 워커 프로세스 중 적어도 어느 하나의 원격 직접 메모리 접근(RDMA) 방식의 읽기 및 쓰기를 지원한다.

[0052] 그리고 공유 메모리 관리부(220)는 공유 메모리의 생성/할당 및 삭제/해제를 관리한다.

[0053] 공유 메모리 관리부(220)는 분산된 마스터 프로세스 또는 워커 프로세스로부터 수신한 파라미터용 원격 공유 메모리 생성 요청에 상응하도록 공유 메모리를 생성하고, 공유 메모리의 공유 메모리 생성키 및 접근 정보를 마스터 프로세스로 전송할 수 있다. 또한, 공유 메모리 관리부(220)는 워커 프로세스로부터 공유 메모리 할당 요청을 수신하고, 공유 메모리 할당 요청에 상응하도록 공유 메모리를 할당한다. 그리고 할당된 공유 메모리의 접근 정보를 워커 프로세스로 전송할 수 있다.

[0054] 그리고 공유 메모리 관리부(220)는 워커 프로세스로부터 공유 메모리 해제 요청을 수신하여 공유 메모리를 해제하고, 마스터 프로세스로부터 공유 메모리 삭제 요청을 수신하여 공유 메모리를 삭제할 수 있다.

[0055] 다음으로 파라미터 연산부(230)는 분산 딥러닝 파라미터를 계산한다. 이때, 분산 딥러닝 파라미터는 가중치 파라미터 및 그래디언트 파라미터를 포함할 수 있다.

[0056] 그리고 파라미터 연산부(230)는 두 개의 공유 메모리 영역에 대한 벡터/매트릭스 연산을 수행할 수 있으며, 벡터 연산은 제1 벡터(X)에 제1 상수(a)를 곱하는 scal 연산(X=aX), 제1 상수(a)를 곱한 제1 벡터(X)와 제2 벡터(Y)를 합하는 axpy 연산(Y=aX+Y), 제1 상수(a)를 곱한 제1 벡터(X)와 제2 상수(b)를 곱한 제2 벡터(Y)를 합하는 axpby 연산(Y=aX+bY) 등을 의미할 수 있다.

[0057] 또한, 파라미터 연산부(230)는 동기식으로 분산 딥러닝 파라미터를 공유하는 경우, 그래디언트 파라미터 누적을 연산하고, 마스터 프로세스의 마스터 가중치 파라미터를 입력받아 마스터 가중치 파라미터 영역을 업데이트할 수 있다.

그리고 비동기식으로 분산 딥러닝 파라미터를 공유하는 경우, 파라미터 연산부(230)는 공유 메모리의 마스터 가중치 파라미터 값으로 워커 프로세스의 워커 로컬 가중치 파라미터 영역을 업데이트하고, 분산 딥러닝을 수행한 워커 프로세스로부터 수신한 워커 그래디언트 파라미터로 마스터 가중치 파라미터 영역을 업데이트할 수 있다.

[0058] 마지막으로 이벤트 처리부(240)는 공유 메모리에 대한 이벤트 발생 시, 공유 메모리에 상응하는 마스터 프로세스 및 워커 프로세스 중 적어도 어느 하나로 이벤트의 발생을 알릴 수 있다. 이벤트 처리부(240)는 특정 공유 메모리 영역에 대한 이벤트를 해당 공유 메모리를 공유하는 지정된 분산 마스터 프로세스 또는 워커 프로세스에 알리는 통보 메시지를 전송할 수 있다.

[0059] 예를 들어, 특정 공유 메모리 영역이 업데이트 되었거나, 특정 공유 메모리 영역에 대해 설정된 연산이 완료된 경우, 이벤트 처리부(240)는 지정된 분산 딥러닝 트레이닝 엔진에 통보 메시지를 전송할 수 있다.

[0061] 이하에서는 도 3 내지 도 5를 통하여 본 발명의 일실시예에 따른 분산 딥러닝 프레임워크의 동작 및 기능에 대하여 더욱 상세하게 설명한다.

[0062] 도 3은 본 발명의 일실시예에 따른 파라미터 공유를 위한 원격 공유 메모리 가상 매핑 매커니즘을 나타낸 예시도이다.

[0063] 도 3과 같이, 분산 딥러닝 트레이닝 엔진을 포함하는 마스터 프로세스(310) 및 워커 프로세스(320)는 파라미터를 공유하기 위하여, 파라미터 서버(330)에 원격 공유 메모리를 생성 및 할당하고, 로컬 물리 메모리(파라미터

의 임시 저장을 위한 호스트 물리 메모리 또는 GPU 등의 가속기 물리 메모리)를 할당하며, 가상 주소 공간에 맵핑한다.

[0064] 마스터 프로세스(310) 및 워커 프로세스(320) 각각은 분산 딥러닝 트레이닝 엔진 및 파라미터 서버 접근 지원부로 구성될 수 있으며, 분산 딥러닝 트레이닝 엔진은 딥러닝 모델 복사본(Model replica)을 이용하여 트레이닝을 수행할 수 있다. 이때, 분산 딥러닝 트레이닝 엔진의 역할은 마스터 프로세스(310)인지, 워커 프로세스(320)인지 여부에 따라 상이할 수 있다.

[0065] 마스터 프로세스(310)의 분산 딥러닝 트레이닝 엔진은 파라미터 서버(330)에 원격 공유 메모리를 생성하고, 하나 이상의 워커 프로세스(320)의 분산 딥러닝 트레이닝 엔진에 원격 공유 메모리의 정보를 전달하여, 워커 프로세스(320)들이 파라미터 서버(330) 상의 동일 메모리 영역에 접근할 수 있도록 한다. 이때, 원격 공유 메모리의 정보는 공유 메모리 생성키, 공유 메모리 크기 등을 포함할 수 있다.

[0066] 그리고 마스터 프로세스(310) 또는 워커 프로세스(320)의 분산 딥러닝 트레이닝 엔진은 파라미터 서버 접근 지원부를 통하여, 원격 계산 노드에서 실행된 파라미터 서버(330)를 이용할 수 있다. 이때, 파라미터 서버(330)가 원격 공유 메모리를 할당하면, 파라미터 서버 접근 지원부는 원격 공유 메모리와 동일한 크기의 로컬 물리 메모리를 할당받고, 할당받은 로컬 물리 메모리를 분산 딥러닝 트레이닝 엔진의 가상 주소 공간에 맵핑한다.

[0067] 마스터 프로세스(310) 또는 워커 프로세스(320)의 분산 딥러닝 트레이닝 엔진은 자신의 로컬 물리 메모리에 트레이닝된 파라미터를 저장하고, 명시적으로 파라미터 서버 접근 지원부가 제공하는 API를 이용하여 동기화(쓰기) 요청을 하면, 로컬 물리 메모리의 계산된 파라미터 데이터가 파라미터 서버(330)의 원격 공유 메모리로 복사된다. 또한, 마스터 프로세스(310) 또는 워커 프로세스(320)는 원격 공유 메모리의 업데이트된 파라미터를 읽어오는 동기화(읽기) 요청도 수행할 수 있다.

[0068] 설명의 편의를 위하여, 파라미터 서버(330)가 제공하는 메모리를 원격 공유 메모리로 명명하였으나, 이는 접근 방식이 공유 메모리 형태의 접근 방식인 것을 의미하며, 공유 메모리를 할당받은 프로세스간 자동 동기화 기능은 제공하지 않고, 원격 공유 메모리는 일종의 통신 버퍼로 활용됨을 의미할 수 있다.

[0070] 도 4는 본 발명의 일실시예에 따른 분산 딥러닝 프레임워크의 기능을 설명하기 위한 구조도이다.

[0071] 도 4에 도시한 바와 같이, 분산 딥러닝 프레임워크는 분산 프로세스(410) 및 파라미터 서버(420)를 포함하며, 분산 프로세스(410)는 분산 딥러닝 트레이닝 엔진(411) 및 파라미터 서버 접근 지원부(415)를 포함할 수 있다.

[0072] 분산 딥러닝 트레이닝 엔진(411)의 관점에서, 파라미터 서버 접근 지원부(415)는 분산 프로세스(계산 노드)(410)에 함께 링크되어 라이브러리 형태로 제공될 수 있으며, 파라미터 서버 접근 지원부(415)의 기능은 모두 유저 레벨 라이브러리 형태로 구현될 수 있다. 또한, 파라미터 서버 접근 API만 라이브러리 형태로 구현되고, 이외의 기능은 장치 드라이버 형태로 구현될 수 있다.

[0073] 분산 딥러닝 트레이닝 엔진(411)은 분산 프로세스(410)에서 실행되며, 파라미터 서버 접근 지원부(415)에서 제공하는 파라미터 서버 접근 API를 이용하여 다른 분산 프로세스의 분산 딥러닝 트레이닝 엔진(411) 간 파라미터를 공유할 수 있다.

[0074] 파라미터 서버(420)는 별도의 프로세스에서 실행되며, 분산 프로세스(410)의 파라미터 서버 접근 지원부(415)와 Infiniband 등의 고속 네트워크 채널을 통해 메시지를 송수신하고, 원격 직접 메모리 접근(RDMA) 방식으로 원격 공유 메모리의 읽기/쓰기를 수행할 수 있다.

[0075] 분산 프로세스(410)의 분산 딥러닝 트레이닝 엔진(411)은 파라미터 서버 접근 지원부(415)의 파라미터 서버 접근 API(응용 프로그램 인터페이스)를 이용하여, 공유 메모리의 할당 및 해제, 명시적 공유 메모리 동기화(읽기/쓰기), 파라미터 연산 요청 등을 수행할 수 있다.

[0076] 그리고 파라미터 서버 접근 지원부(415)는 파라미터 서버 접근 API와 원격 공유 메모리 할당 요청 모듈, 공유 메모리 동기화 모듈, 공유 메모리 파라미터 연산 요청 모듈, 공유 메모리 이벤트 요청 모듈, 메시지 송수신 처리 모듈 및 고속 네트워크 통신 처리 모듈을 포함할 수 있다.

[0077] 그리고 파라미터 서버 접근 지원부(415)는 파라미터 서버 접근 API를 통해 분산 딥러닝 트레이닝 엔진(411)의 요청을 수신하면, 구성 모듈을 이용하여 요청에 상응하는 처리를 수행할 수 있다.

[0078] 예를 들어, 공유 메모리의 할당/해제 요청을 수신한 경우 파라미터 서버 접근 지원부(415)는 원격 공유 메모리 할당 요청 모듈을 이용하여 요청을 처리하고, 공유 메모리 동기화 요청을 수신하면 공유 메모리 동기화 모듈이

원격 메모리의 읽기 및 쓰기를 수행할 수 있다.

[0079] 그리고 파라미터 계산 요청을 수신한 경우 파라미터 서버 접근 지원부(415)는 공유 메모리 파라미터 연산 요청 모듈이 파라미터 서버(420)로 특정 공유 메모리 영역들 간의 연산을 요청할 수 있다. 또한, 이벤트 메시지의 송수신 요청을 수신하면, 파라미터 서버 접근 지원부(415)는 공유 메모리 이벤트 요청 모듈을 통하여 파라미터 서버로 이벤트 메시지의 전송을 요청할 수 있다.

[0080] 파라미터 서버(420)는 분산 프로세스(410)의 파라미터 서버 접근 지원부(415)의 요청을 처리하며, 원격 공유 메모리 할당 관리 모듈, 공유 메모리 파라미터 연산 모듈, 공유 메모리 이벤트 처리 모듈, 메시지 송수신 처리 모듈 및 고속 네트워크 통신 처리 모듈을 포함할 수 있다.

[0081] 원격 공유 메모리 할당 관리 모듈은 공유 메모리의 생성, 삭제, 할당 및 해제 요청을 처리하고, 공유 메모리 파라미터 연산 모듈은 2 개의 공유 메모리 영역에 대한 벡터/매트릭스 연산을 수행할 수 있다. 또한, 공유 메모리 이벤트 처리 모듈은 특정 공유 메모리 영역에 대한 이벤트를 해당 공유 메모리 영역을 생성 및 할당받은 분산 프로세스(410)의 분산 딥러닝 트레이닝 엔진(411)에 알리는 통보 메시지를 전송할 수 있다.

[0082] 설명의 편의를 위하여, 분산 프로세스(410)를 하나만 도시하였으나, 분산 딥러닝 프레임워크는 하나 이상의 분산 프로세스(410)를 포함할 수 있고, 분산 프로세스(410)는 분산 딥러닝 트레이닝 엔진(411)의 기능에 따라 마스터 프로세스 및 워커 프로세스로 구분될 수 있다.

[0084] 도 5는 본 발명의 일실시예에 따른 프로세스별 원격 공유 메모리 할당의 일 예를 나타낸 예시도이다.

[0085] 도 5와 같이, 마스터 프로세스(510)는 마스터 파라미터를 위한 원격 공유 메모리 생성을 담당한다. 그리고 마스터 프로세스(510)는 파라미터 서버(530)에 원격 공유 메모리를 생성하므로, 생성한 모든 원격 공유 메모리 영역에 접근할 수 있으며, 공유 메모리 생성 정보를 워커 프로세스(520)로 전송하여, 워커 프로세스(520)들이 마스터 영역에 접근할 수 있도록 한다.

[0086] 반면, 워커 프로세스(520)는 자신이 트레이닝한 결과를 저장하는 워커 그래디언트 파라미터 영역을 생성하고, 생성한 워커 그래디언트 파라미터 영역에 접근할 수 있다. 즉, 워커 프로세스(520)는 다른 워커 프로세스의 메모리 영역에 접근할 수 없으며, 마스터 파라미터 영역 및 해당 워커 프로세스(520)가 트레이닝한 결과를 저장하는 워커 파라미터 영역에만 접근 가능하다. 다시 말해, 제x 워커 프로세스(520_x)는 마스터 파라미터 영역 및 제x 워커 파라미터 영역에 접근 가능하다.

[0087] 설명의 편의를 위하여, 워커 프로세스(520)가 하나의 워커 파라미터 영역의 공유 메모리를 할당 받는 것으로 도시하였다. 그러나 실제 딥러닝 레이어별로 파라미터가 존재하므로, 워커 프로세스들(520)은 딥러닝 계층별로 하나의 마스터 파라미터와 워커 파라미터에 접근할 수 있으며, 도 5의 마스터 파라미터 영역 및 워커 파라미터 영역들은 다수의 공유 메모리 집합을 의미할 수 있다.

[0089] 이하에서는 도 6 내지 도 10을 통하여 본 발명의 일실시예에 따른 분산 딥러닝 파라미터 공유 방법에 대하여 더욱 상세하게 설명한다.

도 6은 본 발명의 일실시예에 따른 분산 딥러닝 파라미터 공유 방법을 나타낸 순서도이다.

먼저, 파라미터 서버(200)는 분산 딥러닝 프로세스의 요청에 상응하도록, 공유 메모리를 생성 및 할당한다(S110).

파라미터 서버(200)는 마스터 프로세스의 파라미터용 원격 공유 메모리 생성 요청에 상응하도록 공유 메모리를 생성하고, 워커 프로세스의 공유 메모리 할당 요청에 상응하도록 공유 메모리를 할당할 수 있다. 공유 메모리를 생성 및 할당하는 과정에 대해서는 후술할 도 7을 통하여 더욱 상세하게 설명한다.

그리고 파라미터 서버(200)는 공유 메모리의 마스터 가중치 파라미터 영역을 초기화하고(S120), 분산 딥러닝 프로세스들이 공유 메모리를 통해 공유한 딥러닝 파라미터를 이용하여, 분산 딥러닝 트레이닝을 수행하도록 한다(S130).

이때, 파라미터 서버(200)는 동기식 또는 비동기식으로 분산 딥러닝 파라미터를 공유하여, 분산 딥러닝 트레이닝을 수행하도록 할 수 있다. 파라미터 서버(200)가 동기식으로 분산 딥러닝 파라미터를 공유하는 과정에 대해서는 후술할 도 9를 통하여 더욱 상세하게 설명하고, 비동기식으로 분산 딥러닝 파라미터를 공유하는 과정에 대해서는 후술할 도 10을 통하여 더욱 상세하게 설명한다.

분산 딥러닝 트레이닝의 수행이 완료되면, 파라미터 서버(200)는 사용이 완료된 공유 메모리를 해제 및 삭제한다(S140).

파라미터 서버(200)는 워커 프로세스의 공유 메모리 해제 요청에 따라 공유 메모리를 해제하고, 마스터 프로세스로부터 공유 메모리 삭제 요청을 수신하면 공유 메모리를 삭제한다. 공유 메모리를 해제 및 삭제하는 과정에 대해서는 후술할 도 8을 통하여 더욱 상세하게 설명한다.

[0090] 도 7은 본 발명의 일실시예에 따른 원격 공유 메모리의 생성 및 할당 과정을 나타낸 순서도이다.

[0091] 먼저, 마스터 프로세스(100)는 파라미터 서버(200)로 원격 공유 메모리 생성 요청을 전송한다(S610).

[0092] 그리고 파라미터 서버(200)는 수신된 원격 공유 메모리 생성 요청에 상응하도록 공유 메모리를 생성하고(S620), 마스터 프로세스(100)로 공유 메모리 생성키 및 접근 정보를 전송한다(S630).

[0093] 이때, 파라미터 서버(200)는 생성된 공유 메모리에 접근하고자 하는 경우 필요한 정보인 공유 메모리 주소, 원격 메모리 접근키 등을 공유 메모리 생성키와 함께 마스터 프로세스(100)로 전송할 수 있다.

[0094] 다음으로 마스터 프로세스(100)는 파라미터 서버(200)로 공유 메모리 이벤트 설정 요청을 전송한다(S640).

[0095] 마스터 프로세스(100)는 업데이트 통지 이벤트, 누적 완료 이벤트 등의 공유 이벤트에 대한 이벤트 설정 요청을 파라미터 서버(200)로 전송할 수 있다. 여기서, 업데이트 통지 이벤트는 마스터 프로세스(100)가 특정 공유 메모리를 업데이트한 경우, 해당 공유 메모리를 공유하고 있는 모든 워커 프로세스들(150)로 알리는 메시지를 전송하도록 하는 이벤트를 의미한다.

[0096] 그리고 누적 완료 이벤트는 워커 프로세스들(150)이 특정 공유 메모리에 누적 수행을 완료한 경우, 마스터 프로세스(100)로 누적 완료를 알리는 메시지를 전송하는 이벤트를 의미한다.

[0097] 또한, 마스터 프로세스(100)는 공유 메모리 생성키를 하나 이상의 워커 프로세스(150)로 배포한다(S650).

[0098] 설명의 편의를 위하여 도 7에는 마스터 프로세스(100)가 하나의 워크 프로세스(150_1)로 공유 메모리 생성키를 배포하는 것으로 도시하였으나 이에 한정하지 않고, 마스터 프로세스(100)는 분산 딥러닝 프레임워크에 포함된 복수 개의 워커 프로세스(150)로 공유 메모리 생성키를 배포할 수 있다. 이때, 마스터 프로세스(100)는 마스터 프로세스(100)와 워커 프로세스(150)간 별도의 통신 채널을 이용하여 공유 메모리 생성키를 배포할 수 있다.

[0099] 그리고 공유 메모리 생성키를 수신한 제1 워커 프로세스(150_1)는 파라미터 서버(200)로 공유 메모리 할당 요청을 전송하고(S660), 파라미터 서버(200)는 공유 메모리를 할당한다(S670).

[0100] 마스터 프로세스(100)로부터 공유 메모리 생성키를 수신한 워커 프로세스(150)는 공유 메모리 생성키를 이용하여 파라미터 서버(200)로 공유 메모리 할당을 요청할 수 있다. 그리고 파라미터 서버(200)는 공유 메모리 생성키로 기 생성된 공유 메모리에 대한 할당을 수행할 수 있다.

[0101] 또한, 파라미터 서버(200)는 제1 워커 프로세스(150_1)로 할당된 공유 메모리 접근 정보를 전송한다(S680).

[0102] 파라미터 서버(200)는 공유 메모리 접근에 필요한 정보인 공유 메모리 주소, 원격 메모리 접근키 등의 공유 메모리 접근 정보를 워커 프로세스(150)로 전송한다. 그리고 공유 메모리 접근 정보를 수신한 워커 프로세스(150)는 공유 메모리 접근 정보를 이용하여 할당받은 원격 공유 메모리 주소에 RDMA 직접 읽기 또는 쓰기를 수행할 수 있다.

[0103] 그리고 분산 딥러닝 프레임워크에 포함된 모든 워커 프로세스들(150)이 S670 단계를 수행하여 공유 메모리 접근 정보를 수신하면, 마스터 프로세스(100)는 딥러닝 트레이닝을 수행할 수 있다.

[0104] 도 7과 같은 공유 메모리 할당 이외에, 워커 프로세스(150)가 주도적으로 공유 메모리를 할당하여, 다른 워커 프로세스들과 공유할 수 있으며, 마스터 프로세스(100) 및 워커 프로세스들(150)이 포함하는 딥러닝 트레이닝 엔진들 간의 공유 메모리 할당이 완료되면, 딥러닝 트레이닝 엔진들은 트레이닝을 시작할 수 있다. 그리고 딥러닝 트레이닝 중에는 마스터 프로세스(100)와 워커 프로세스(150)간에 다양한 형태로 딥러닝 파라미터가 공유될 수 있다.

[0106] 도 8은 본 발명의 일실시예에 따른 원격 공유 메모리의 삭제 및 해제 과정을 나타낸 순서도이다.

[0107] 제1 워커 프로세스(150_1)는 파라미터 서버(200)로 공유 메모리 해제 요청을 전송한다(S710).

[0108] 딥러닝 트레이닝이 완료된 후, 워커 프로세스(150) 각각은 자신이 할당받은 원격 공유 메모리의 해제를 파라미터 서버(200)에 요청할 수 있다.

[0109] 그리고 공유 메모리 해제 요청을 수신한 파라미터 서버(200)는 공유 메모리를 해제하고(S720), 제1 워커 프로세스(150_1)로 공유 메모리 해제를 통보한다(S730).

[0110] 여기서, 공유 메모리의 해제는 파라미터 서버(200)가 공유 메모리에 대한 공유 정보를 삭제하는 것을 의미할 수 있다.

[0111] 또한, 마스터 프로세스(100)는 파라미터 서버(200)로 원격 공유 메모리 삭제 요청을 전송하고(S740), 공유 메모리 삭제 요청을 수신한 파라미터 서버(200)는 공유 메모리를 삭제하며(S750), 마스터 프로세스(100)로 공유 메모리 삭제 완료를 통보한다(S760).

[0113] 이하에서는 도 9 및 도 10을 통하여 본 발명의 일실시예에 따른 분산 딥러닝 프레임워크 환경에서 동기식 및 비동기식으로 파라미터를 공유하는 방법에 대하여 더욱 상세하게 설명한다.

[0114] 파라미터 서버(200)에 원격 공유 메모리가 생성 및 할당된 후, 파라미터 서버(200)는 분산 딥러닝 프로세스들(100, 150)이 공유 메모리를 통해 딥러닝 파라미터를 공유하여, 분산 딥러닝 트레이닝을 수행하도록 할 수 있다. 즉, 마스터 프로세스(100) 및 하나 이상의 워커 프로세스(150)는 파라미터 서버(200)를 기반으로 딥러닝 파라미터를 공유하여, 딥러닝 트레이닝 과정을 반복 수행할 수 있다.

[0115] 여기서, 파라미터 서버(200)에 생성되는 파라미터는 마스터 가중치 파라미터(W_{Master}), 마스터 그래디언트 파라미터(G_{Master}) 및 워커x 그래디언트 파라미터(G_{worker_x})로 구분될 수 있다.

[0116] 그리고 딥러닝 트레이닝 과정에서 분산 딥러닝 파라미터는 도 9 또는 도 10의 과정을 통하여 동기식 또는 비동기식으로 공유될 수 있다. 이때, 도 9 및 도 10의 분산 딥러닝 파라미터를 공유하는 과정은 딥러닝 알고리즘에 따라 일부 순서의 변경 및 수정이 가능하다.

[0117] 또한, 도 9 및 도 10의 파라미터 공유 과정 각각은 도 7의 공유 메모리 생성 및 할당 과정을 수행한 후 수행될 수 있으며, 도 9 및 도 10의 과정을 수행한 후 도 8의 공유 메모리 삭제 및 해제 과정이 수행될 수 있다.

[0119] 도 9는 본 발명의 일실시예에 따른 동기식 파라미터 공유 방법을 설명하기 위한 순서도이다.

[0120] 먼저, 마스터 프로세스(100)는 파라미터 서버(200)의 마스터 가중치 파라미터 영역(W_{Master}) 및 마스터 그래디언트 파라미터 영역(G_{Master})을 초기화한다(S810).

[0121] 마스터 프로세스(100)는 마스터 프로세스(100)의 로컬 메모리에 초기화된 가중치 파라미터 값을 마스터 가중치 파라미터 영역에 쓰기하여, 마스터 가중치 파라미터 영역(W_{Master})을 초기화할 수 있다. 그리고 마스터 프로세스(100)는 모든 값을 0으로 설정하여 마스터 그래디언트 파라미터 영역(G_{Master})을 리셋할 수 있다.

[0122] 그리고 파라미터 서버(200)는 제1 워커 프로세스(150_1)로 마스터 가중치 파라미터(W_{Master})가 업데이트 되었음을 알린다(S820).

[0123] 파라미터 서버(200)는 마스터 가중치 파라미터(W_{Master})영역을 공유하는 하나 이상의 워커 프로세스들(150)로, 마스터 가중치 파라미터(W_{Master}) 영역이 업데이트 되었음을 알릴 수 있다.

[0124] 제1 워커 프로세스(150_1)는 마스터 가중치 파라미터(W_{Master})를 읽어와 워커 로컬 가중치 파라미터를 업데이트하고(S830), 딥러닝 트레이닝을 수행한다(S840).

[0125] 제1 워커 프로세스(150_1)는 워커 로컬 가중치 파라미터 영역을 공유 메모리의 마스터 가중치 파라미터의 값으로 업데이트할 수 있다. 즉, 워커 프로세스들(150)은 파라미터 서버(200)의 마스터 가중치 파라미터 영역을 RDMA 방식으로 읽어, 워커 로컬 가중치 파라미터(W_{Worker}) 영역으로 복사한다($W_{Worker} = W_{Master}$). 여기서, X는 워커 프로세스의 일련 번호를 의미하며, 제1 워커 프로세스(150_1)는 로컬 가중치 파라미터($W_{Worker1}$)을 업데이트할 수 있다.

[0126] 그리고 S840 단계에서 워커 프로세스들(150) 각각은 지정된 반복 트레이닝 횟수만큼 반복하여 딥러닝 트레이닝을 수행한다. 이때, 워커 프로세스들(150)은 가중치 파라미터는 업데이트하지 않고, 그래디언트 파라미터(G_{Worker})만 연산할 수도 있다.

[0127] 또한, 제1 워커 프로세스(150_1)는 파라미터 서버(200)에 워커 로컬 그래디언트 파라미터를 저장한다(S850).

[0128] 딥러닝 트레이닝을 수행한 워커 프로세스들(150)은 학습된 워커 로컬 그래디언트 파라미터(G_{Worker})를 공유 메모리의 워커 그래디언트 파라미터 영역에 RDMA 쓰기 한다. 즉, 제1 워커 프로세스(150_1)는 제1 워커 로컬 그래디언트 파라미터($G_{Worker1}$)를 제1 워커 파라미터 영역에 RDMA 쓰기 할 수 있다.

[0129] 그리고 제1 워커 프로세스(150_1)는 파라미터 서버(200)로 그래디언트 파라미터 누적 연산을 요청하고(S860), 파라미터 서버(200)는 요청된 그래디언트 파라미터 영역들 간의 그래디언트 파라미터 누적 연산을 수행한다(S870).

[0130] 제1 워커 프로세스(150_1)는 공유 메모리의 제1 워커 파라미터 영역에 저장된 워커 로컬 그래디언트 파라미터(G_{Worker})를 마스터 그래디언트 파라미터(G_{Master})에 누적하도록, 파라미터 서버(200)에 요청한다. 그리고 파라미터 서버(200)는 요청된 그래디언트 파라미터 영역들 간의 파라미터를 누적하는 연산인 $G_{Master}' = G_{Master} + G_{Worker}$ 연산을 수행할 수 있다.

[0131] 모든 워커 프로세스(150)들의 그래디언트 파라미터 누적 연산이 완료되면 파라미터 서버(200)는 마스터 프로세스(100)로 그래디언트 파라미터(G_{Master}) 누적 연산의 완료를 통보한다(S880).

[0132] 마스터 프로세스(100)는 분산 딥러닝 프레임워크에 포함된 모든 워커 프로세스들(150)의 그래디언트 파라미터 누적이 완료될 때까지 대기한 후, 파라미터 서버(200)로부터 누적 완료된 마스터 그래디언트 파라미터(G_{Master}) 영역을 읽어온다(S890).

[0133] 이때, 마스터 프로세스(100)는 모든 워커 프로세스들(150)의 그래디언트 파라미터가 누적된 마스터 그래디언트 파라미터 영역(G_{Master})을 RDMA 방식으로 읽어올 수 있다.

[0134] 그리고 마스터 프로세스(100)는 마스터 가중치 파라미터(W_{Master}')를 연산하며(S900), 마스터 가중치 파라미터(W_{Master}')를 파라미터 서버(200)에 업데이트한다(S910).

[0135] 마스터 프로세스(100)는 S890 단계에서 읽어온 그래디언트 누적 값(G_{Master})의 평균을 이용하여, 마스터 가중치 파라미터(W_{Master}')를 연산할 수 있다. 또한, 마스터 프로세스(100)는 새로 업데이트된 마스터 가중치 파라미터(W_{Master}')를 파라미터 서버(200)의 마스터 가중치 파라미터 영역에 저장할 수 있다.

[0136] 마스터 프로세스(100) 및 워커 프로세스들(150)은 지정된 반복 트레이닝 횟수만큼 S820 단계 내지 S910 단계의 수행을 반복할 수 있다.

[0138] 도 10은 본 발명의 일실시예에 따른 비동기식 파라미터 공유 방법을 설명하기 위한 순서도이다.

[0139] 먼저, 마스터 프로세스(100)는 파라미터 서버(200)의 마스터 가중치 파라미터(W_{Master}) 영역을 초기화한다(S1010). 그리고 파라미터 서버(200)는 제1 워커 프로세스(150_1)로 마스터 가중치 파라미터(W_{Master})의 업데이트를 통보한다(S1020).

[0140] 설명의 편의를 위하여, 파라미터 서버(200)가 제1 워커 프로세스(150_1)로 마스터 가중치 파라미터의 업데이트를 통보하는 것으로 설명하였으나 이에 한정하지 않고, 파라미터 서버(200)는 분산 딥러닝 프레임워크에 포함된 하나 이상의 워커 프로세스들(150)로 마스터 가중치 파라미터(W_{Master})가 업데이트 되었음을 알릴 수 있다.

[0141] 다음으로, 제1 워커 프로세스(150_1)는 공유 메모리의 마스터 가중치 파라미터(W_{Master})를 읽어와, 워커 로컬 가중치 파라미터(W_{Worker}) 영역을 업데이트하고(S1030), 딥러닝 트레이닝을 수행한다(S1040).

[0142] 제1 워커 프로세스(150_1)는 RDMA 방식으로 마스터 가중치 파라미터(W_{Master})를 읽어올 수 있으며, 읽어온 마스터 파라미터(W_{Master})를 워커 로컬 가중치 파라미터(W_{Worker})로 복사($W_{Worker} = W_{Master}$)하여, 워커 로컬 가중치 파라미터

(W_{Worker})를 업데이트할 수 있다. 그리고 제1 워커 프로세스(150_1)는 지정된 반복 횟수만큼 딥러닝 트레이닝을 수행하여, 워커 로컬 그래디언트 파라미터(G_{Worker})를 계산할 수 있다.

[0143] 딥러닝 트레이닝을 수행한 제1 워커 프로세스(150_1)는 새로 학습된 워커 그래디언트 파라미터(G_{Worker})를 공유 메모리에 RDMA 쓰기하여 업데이트한다(S1050). 그리고 제1 워커 프로세스(150_1)는 마스터 파라미터 서버(200)에 마스터 가중치 파라미터(W_{Master})의 업데이트를 요청한다(S1060).

[0144] 파라미터 서버(200)는 마스터 가중치 파라미터(W_{Master})의 업데이트를 수행하고(S1070), 업데이트를 요청한 제1 워커 프로세스(150_1)로 업데이트의 완료를 통보한다(S1080).

[0145] 이때, 파라미터 서버(200)는 복수의 워커 프로세스들(150)로부터 수신된 마스터 가중치 파라미터의 업데이트 수행 요청을 동시에 수행하지 않고, 순차적으로 처리할 수 있다.

[0146] 그리고 파라미터 서버(200)는 마스터 파라미터 영역의 업데이트가 완료되었음을 하나 이상의 워커 프로세스(150)에 통보할 수 있다. 이때, 딥러닝 트레이닝이 종료되지 않은 경우, S1030 단계 내지 S1080 단계의 과정을 반복하여 수행할 수 있다.

[0147] 도 9 및 도 10에는 도시하지 않았으나, 딥러닝 트레이닝 종료 시, 마스터 가중치 파라미터를 저장하는 과정을 수행한 후, 딥러닝 트레이닝을 종료할 수 있다.

[0149] 이상에서와 같이 본 발명에 따른 파라미터 서버 및 그것에 의해 수행되는 분산 딥러닝 파라미터 공유 방법은 상기한 바와 같이 설명된 실시예들의 구성과 방법이 한정되게 적용될 수 있는 것이 아니라, 상기 실시예들은 다양한 변형이 이루어질 수 있도록 각 실시예들의 전부 또는 일부가 선택적으로 조합되어 구성될 수도 있다.

부호의 설명

[0150] 100: 마스터 프로세스 150: 워커 프로세스

200: 파라미터 서버 210: 통신 처리부

220: 공유 메모리 관리부 230: 파라미터 연산부

240: 이벤트 처리부 310: 마스터 프로세스

320: 워커 프로세스 330: 파라미터 서버

410: 분산 프로세스 411: 분산 딥러닝 트레이닝 엔진

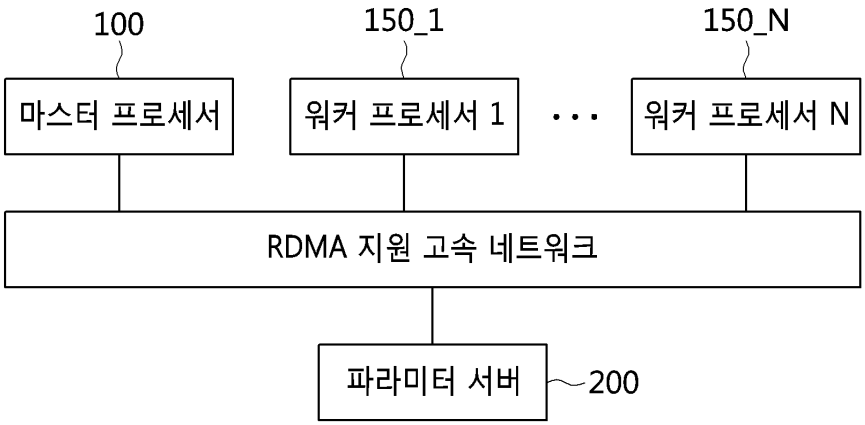
415: 파라미터 서버 접근 지원부 420: 파라미터 서버

510: 마스터 프로세스 520: 워커 프로세스

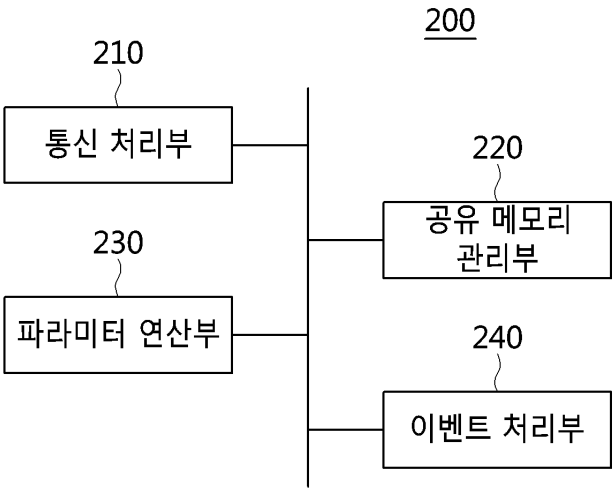
530: 파라미터 서버

도면

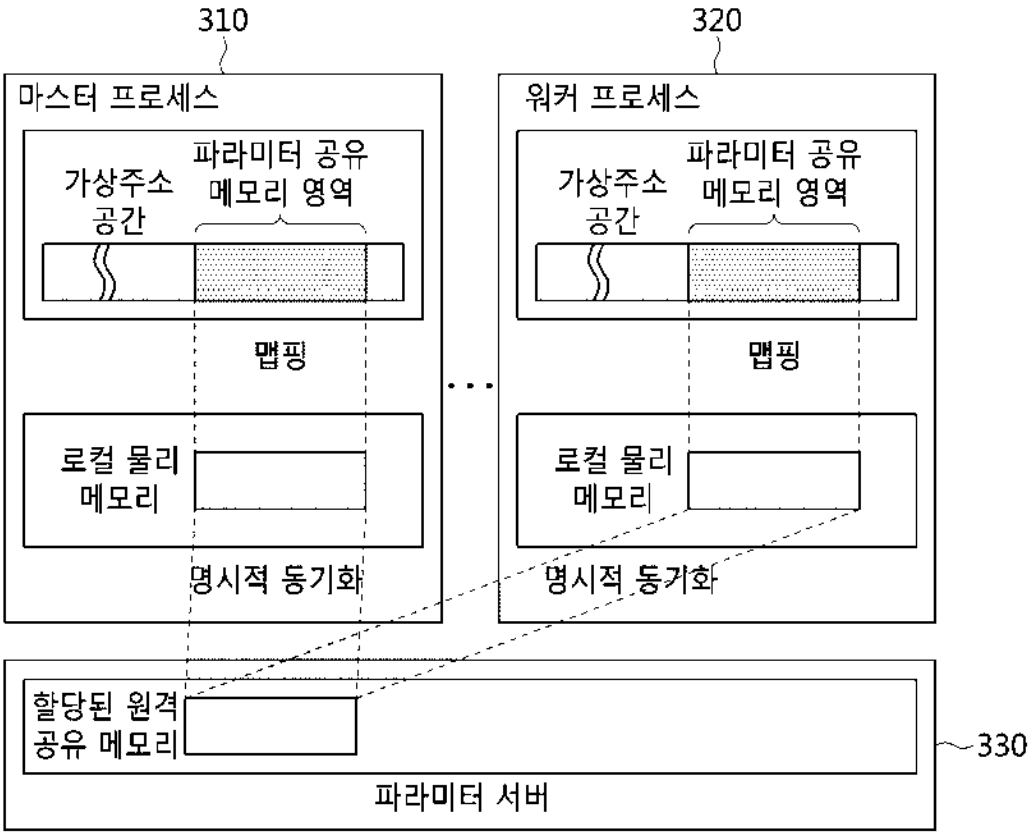
도면1


100
150_1
150_N
마스터 프로세서
워커 프로세서 1
• • •
워커 프로세서 N
RDMA 지원 고속 네트워크
파라미터 서버
200

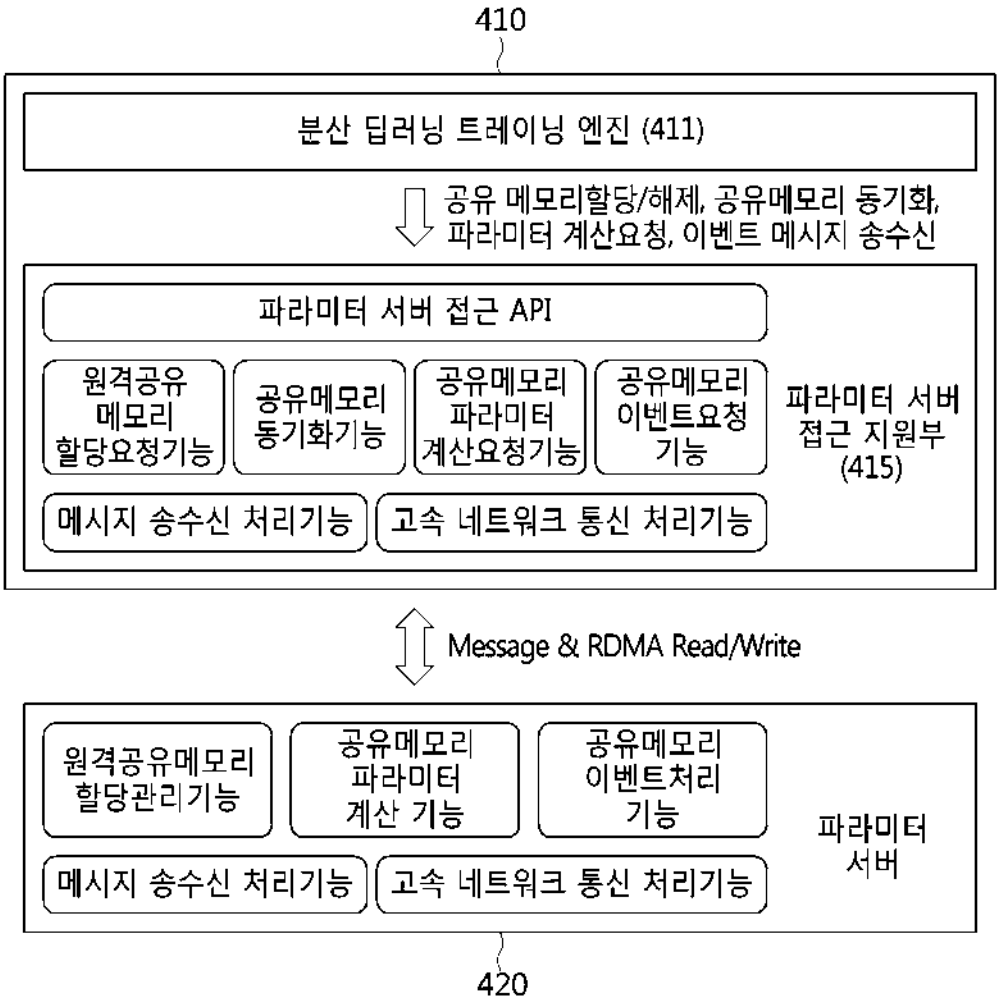
도면2


200
210
통신 처리부
220
공유 메모리 관리부
230
파라미터 연산부
240
이벤트 처리부

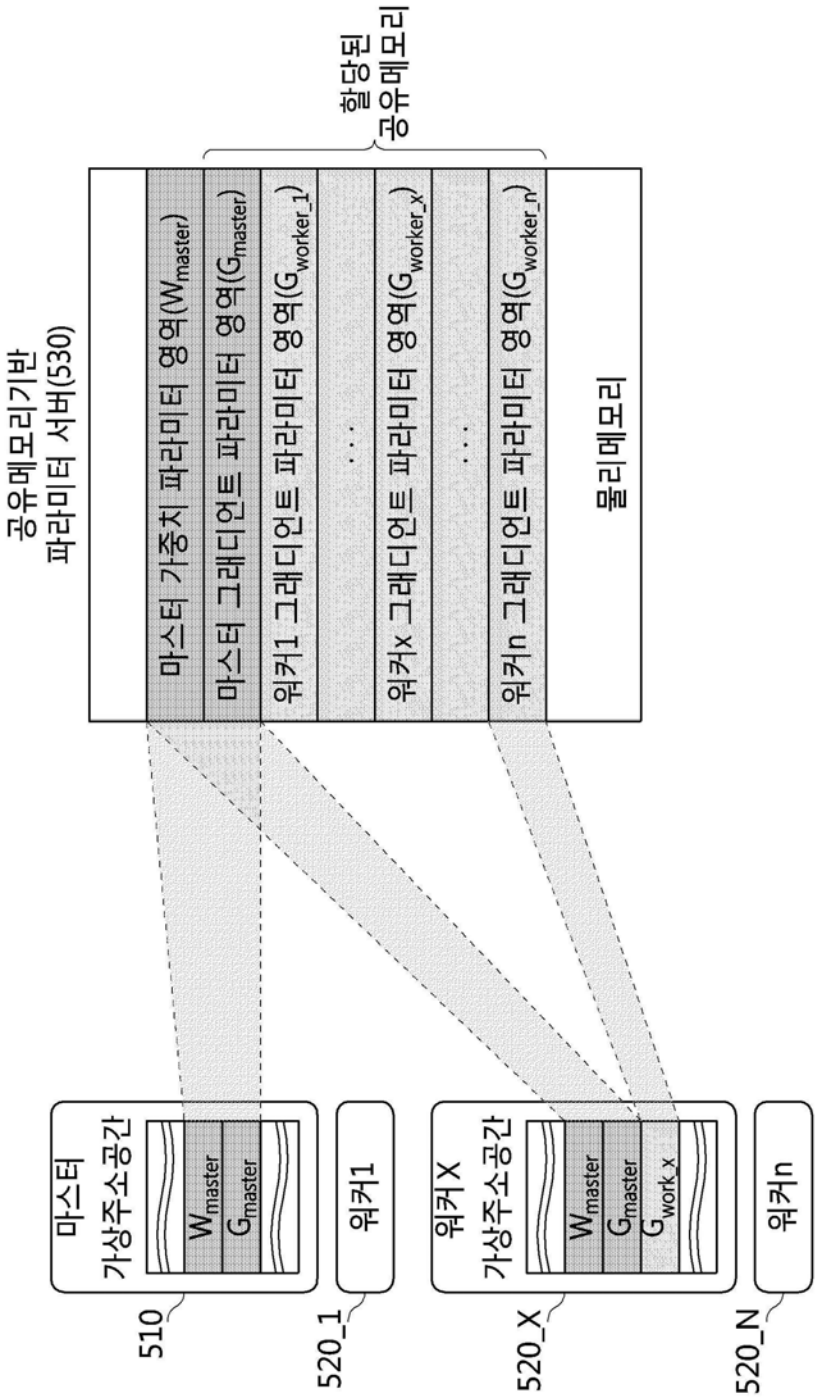
도면3


310
320
마스터 프로세스
워커 프로세스
가상주소 공간
파라미터 공유 메모리 영역
맵핑
로컬 물리 메모리
명시적 동기화
할당된 원격 공유 메모리
330
파라미터 서버

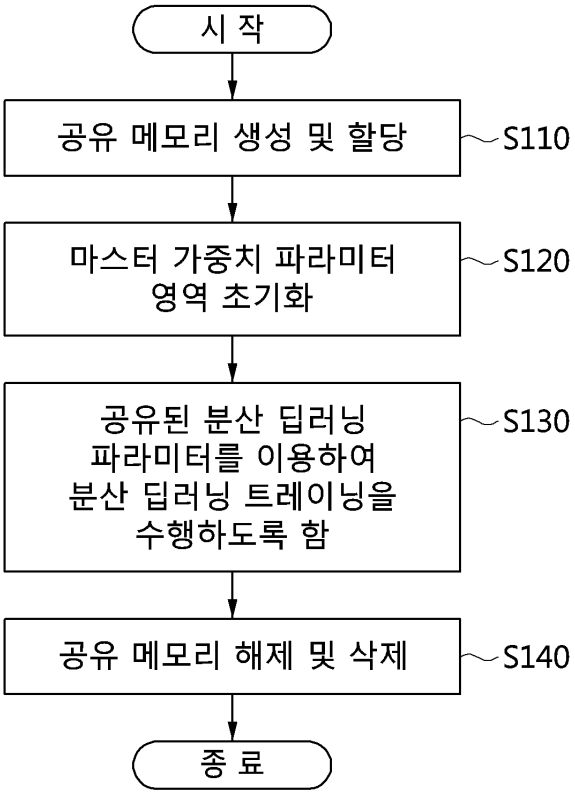
도면4


410
분산 딥러닝 트레이닝 엔진 (411)
공유 메모리할당/해제, 공유메모리 동기화,
파라미터 계산요청, 이벤트 메시지 송수신
파라미터 서버 접근 API
원격공유 메모리 할당요청기능
공유메모리 동기화기능
공유메모리 파라미터 계산요청기능
공유메모리 이벤트요청 기능
파라미터 서버 접근 지원부 (415)
메시지 송수신 처리기능
고속 네트워크 통신 처리기능
Message & RDMA Read/Write
원격공유메모리 할당관리기능
공유메모리 파라미터 계산 기능
공유메모리 이벤트처리 기능
파라미터 서버
메시지 송수신 처리기능
고속 네트워크 통신 처리기능
420

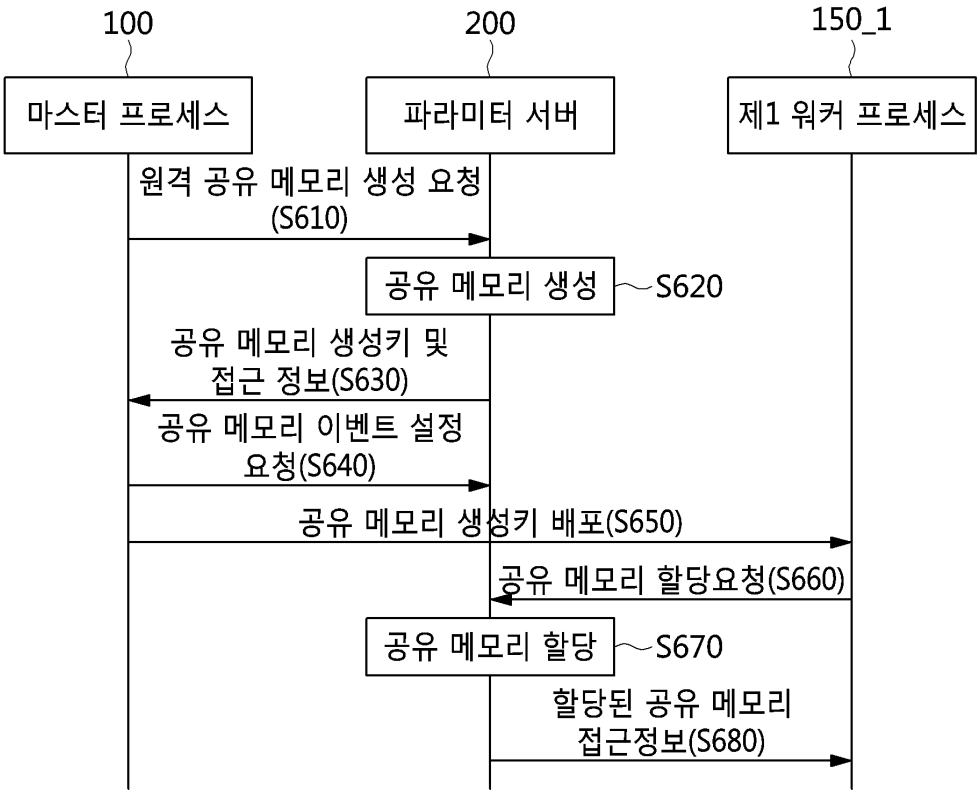
도면5


할당된
공유메모리
공유메모리기반
파라미터 서버(530)
마스터 가중치 파라미터 영역(W_{master})
마스터 그래디언트 파라미터 영역(G_{master})
워커1 그래디언트 파라미터 영역(G_{worker_1})
...
워커x 그래디언트 파라미터 영역(G_{worker_x})
...
워커n 그래디언트 파라미터 영역(G_{worker_n})
물리메모리
마스터
가상주소공간
W_{master}
G_{master}
510
520_1
워커1
워커 X
가상주소공간
W_{master}
G_{master}
G_{work_x}
520_X
520_N
워커n

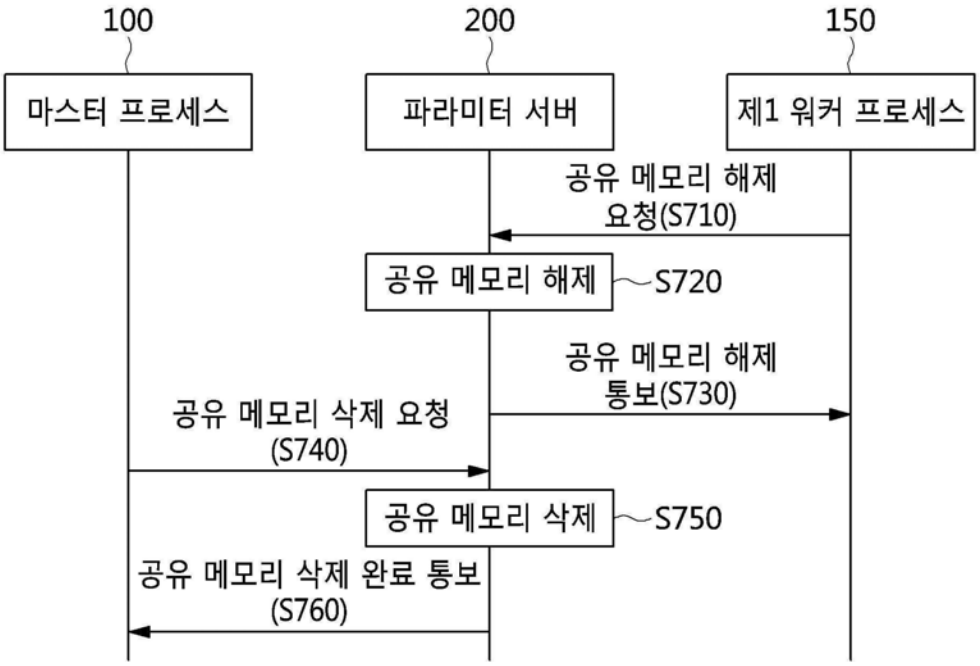
도면6


시 작
공유 메모리 생성 및 할당
S110
마스터 가중치 파라미터 영역 초기화
S120
공유된 분산 딥러닝 파라미터를 이용하여 분산 딥러닝 트레이닝을 수행하도록 함
S130
공유 메모리 해제 및 삭제
S140
종 료

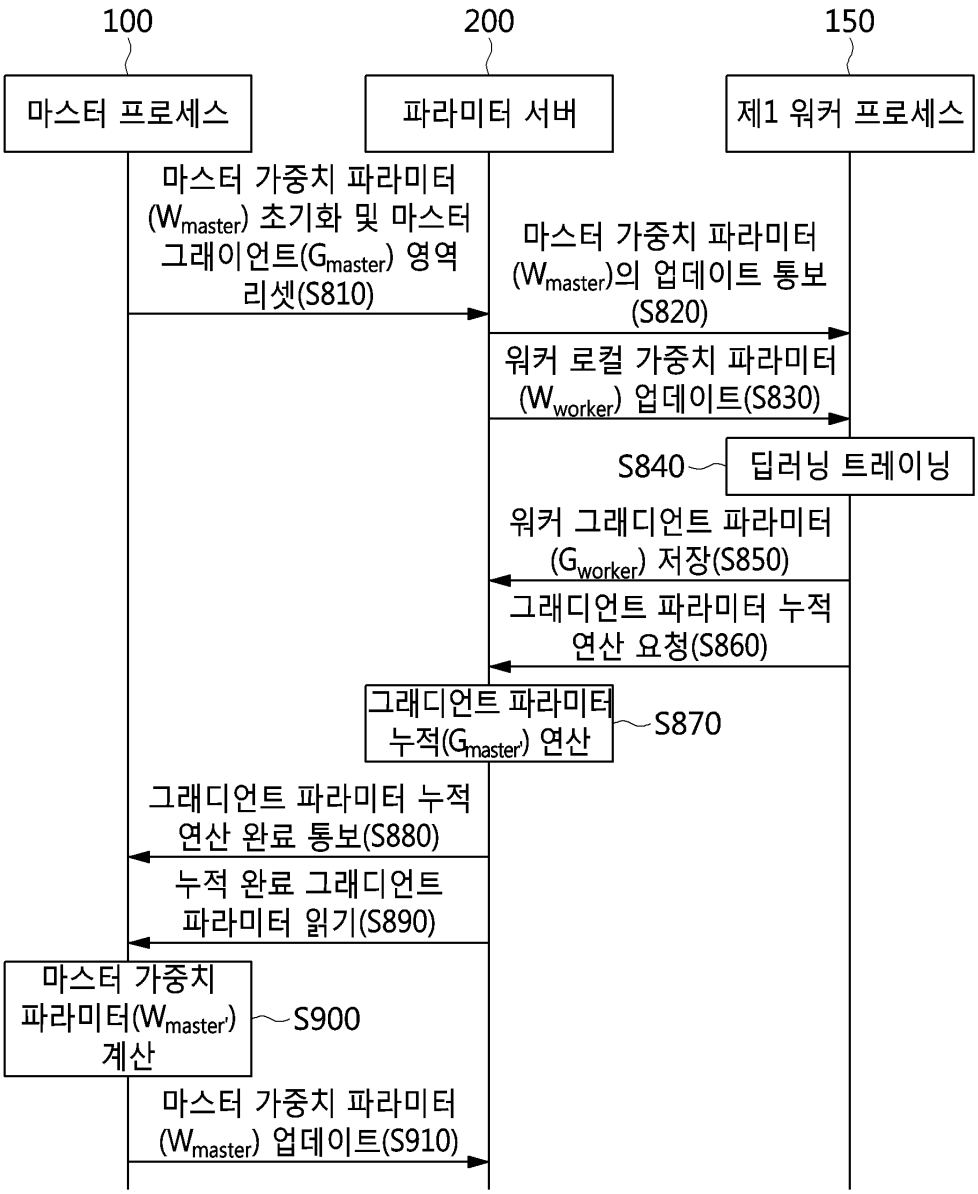
도면7


100
200
150_1
마스터 프로세스
파라미터 서버
제1 워커 프로세스
원격 공유 메모리 생성 요청 (S610)
공유 메모리 생성
S620
공유 메모리 생성키 및 접근 정보(S630)
공유 메모리 이벤트 설정 요청(S640)
공유 메모리 생성키 배포(S650)
공유 메모리 할당요청(S660)
공유 메모리 할당
S670
할당된 공유 메모리 접근정보(S680)

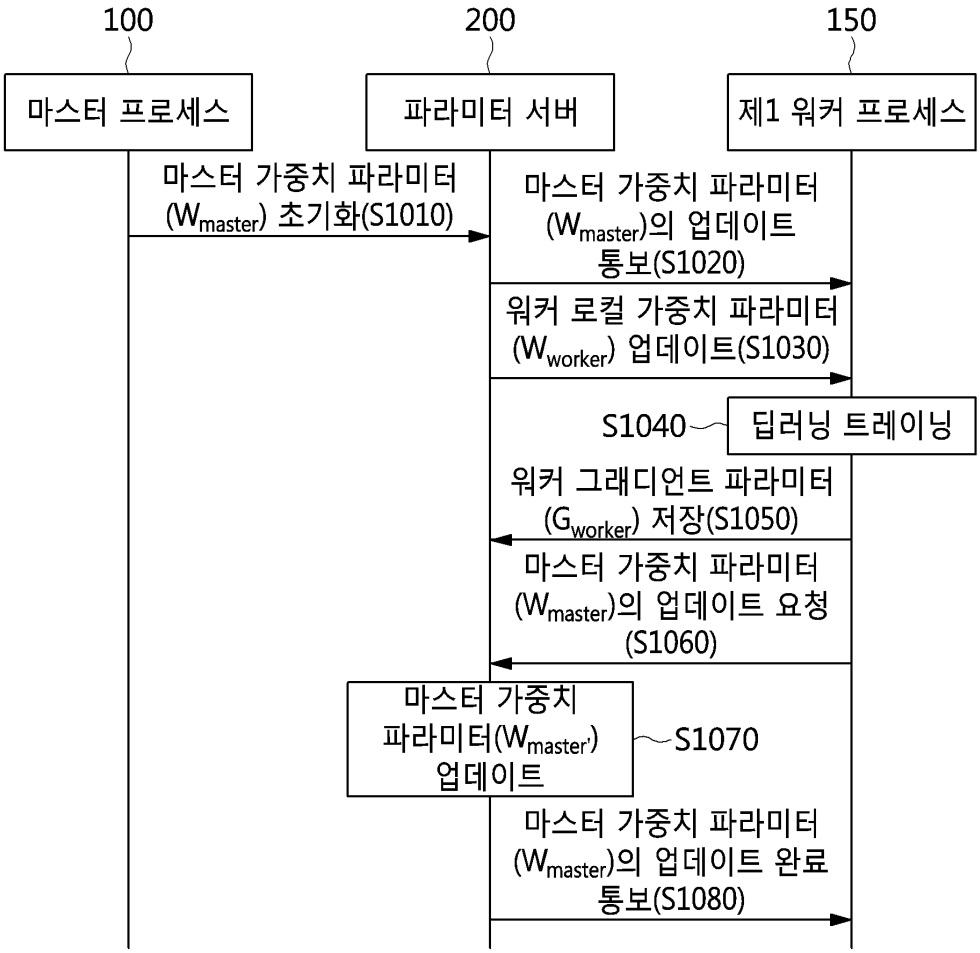
도면8


100
200
150
마스터 프로세스
파라미터 서버
제1 워커 프로세스
공유 메모리 해제 요청(S710)
공유 메모리 해제
S720
공유 메모리 해제 통보(S730)
공유 메모리 삭제 요청 (S740)
공유 메모리 삭제
S750
공유 메모리 삭제 완료 통보 (S760)

도면9


100
200
150
마스터 프로세스
파라미터 서버
제1 워커 프로세스
마스터 가중치 파라미터 (W_{master}) 초기화 및 마스터 그래이언트(G_{master}) 영역 리셋(S810)
마스터 가중치 파라미터 (W_{master})의 업데이트 통보 (S820)
워커 로컬 가중치 파라미터 (W_{worker}) 업데이트(S830)
S840
딥러닝 트레이닝
워커 그래디언트 파라미터 (G_{worker}) 저장(S850)
그래디언트 파라미터 누적 연산 요청(S860)
그래디언트 파라미터 누적($G_{master'}$) 연산
S870
그래디언트 파라미터 누적 연산 완료 통보(S880)
누적 완료 그래디언트 파라미터 읽기(S890)
마스터 가중치 파라미터($W_{master'}$) 계산
S900
마스터 가중치 파라미터 (W_{master}) 업데이트(S910)

도면10



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 1

【변경전】

파라미터 서버에 의해 수행되는 분산 딥러닝 파라미터 공유 방법에 있어서,

마스터 프로세스 및 적어도 하나의 워커 프로세스를 포함하는 분산 딥러닝 프로세스의 요청에 상응하도록, 원격 공유 메모리를 생성 및 할당하는 단계,

상기 원격 공유 메모리의 마스터 가중치 파라미터 영역을 초기화하는 단계,

상기 분산 딥러닝 프로세스들이 상기 원격 공유 메모리를 통해 공유한 분산 딥러닝 파라미터를 이용하여 분산 딥러닝 트레이닝을 수행하는 단계, 그리고

상기 분산 딥러닝 트레이닝의 수행이 완료된 후, 사용이 완료된 상기 원격 공유 메모리를 해제 및 삭제하는 단계를 포함하고,

상기 원격 공유 메모리를 생성 및 할당하는 단계는,

상기 마스터 프로세스의 요청에 따라 원격 공유 메모리를 생성하고, 상기 적어도 하나의 워커 프로세스에게 상기 원격 공유 메모리에 접근하기 위한 접근 정보를 전달하고,

상기 마스터 프로세스와 상기 적어도 하나의 워커 프로세스가, 상기 원격 공유 메모리에 상응하는 각각의 로컬 물리 메모리를 할당하고, 상기 각각의 로컬 물리 메모리를 분산 딥러닝 트레이닝 엔진의 가상 주소 공간에 맵핑

하고,

상기 분산 딥러닝 트레이닝을 수행하는 단계는

상기 파라미터 서버, 상기 마스터 프로세스 및 상기 적어도 하나의 워커 프로세스가, 상기 각각의 로컬 물리 메모리와 상기 원격 공유 메모리의 명시적 동기화를 통해 상기 분산 딥러닝 파라미터를 공유하여 상기 분산 딥러닝 트레이닝을 수행하고,

상기 원격 공유 메모리는

상기 마스터 프로세스에 의해 생성되고, 마스터 가중치 파라미터와 마스터 그래디언트 파라미터를 저장하는 마스터 영역; 및

상기 적어도 하나의 워커 프로세스의 개수에 상응하도록 생성되고, 적어도 하나의 워커 그래디언트 파라미터를 각각 저장하는 적어도 하나의 워커 영역;

을 포함하고,

상기 분산 딥러닝 트레이닝을 수행하는 단계는

상기 마스터 프로세스가, 상기 마스터 영역에 상기 마스터 가중치 파라미터와 상기 마스터 그래디언트 파라미터를 업데이트하고, 상기 적어도 하나의 워커 프로세스에게 상기 마스터 영역에 접근하기 위한 접근 정보를 전송하고,

상기 적어도 하나의 워커 프로세스가, 상기 접근 정보를 이용하여 상기 마스터 영역에 접근하여 상기 마스터 가중치 파라미터를 자신의 워커 가중치 파라미터로 업데이트하고,

상기 원격 공유 메모리의 적어도 하나의 워커 영역 중 자신이 생성한 워커 영역에 상기 분산 딥러닝 트레이닝을 수행한 결과로부터 학습된 워커 그래디언트 파라미터를 업데이트하고,

상기 파라미터 서버가, 상기 적어도 하나의 워커 프로세스로부터 상기 적어도 하나의 워커 영역에 상기 적어도 워커 그래디언트 파라미터가 업데이트되었음을 알림받으면, 상기 적어도 워커 그래디언트 파라미터를 상기 마스터 그래디언트 파라미터에 업데이트하고,

상기 마스터 서버가, 업데이트된 상기 마스터 그래디언트 파라미터를 이용하여 상기 마스터 가중치 파라미터를 업데이트하는 것을 특징으로 하는 분산 딥러닝 파라미터 공유 방법.

【변경후】

파라미터 서버에 의해 수행되는 분산 딥러닝 파라미터 공유 방법에 있어서,

마스터 프로세스 및 적어도 하나의 워커 프로세스를 포함하는 분산 딥러닝 프로세스의 요청에 상응하도록, 원격 공유 메모리를 생성 및 할당하는 단계,

상기 원격 공유 메모리의 마스터 가중치 파라미터 영역을 초기화하는 단계,

상기 분산 딥러닝 프로세스들이 상기 원격 공유 메모리를 통해 공유한 분산 딥러닝 파라미터를 이용하여 분산 딥러닝 트레이닝을 수행하는 단계, 그리고

상기 분산 딥러닝 트레이닝의 수행이 완료된 후, 사용이 완료된 상기 원격 공유 메모리를 해제 및 삭제하는 단계를 포함하고,

상기 원격 공유 메모리를 생성 및 할당하는 단계는,

상기 마스터 프로세스의 요청에 따라 원격 공유 메모리를 생성하고, 상기 적어도 하나의 워커 프로세스에게 상기 원격 공유 메모리에 접근하기 위한 접근 정보를 전달하고,

상기 마스터 프로세스와 상기 적어도 하나의 워커 프로세스가, 상기 원격 공유 메모리에 상응하는 각각의 로컬 물리 메모리를 할당하고, 상기 각각의 로컬 물리 메모리를 분산 딥러닝 트레이닝 엔진의 가상 주소 공간에 맵핑하고,

상기 분산 딥러닝 트레이닝을 수행하는 단계는

상기 파라미터 서버, 상기 마스터 프로세스 및 상기 적어도 하나의 워커 프로세스가, 상기 각각의 로컬 물리 메모리와 상기 원격 공유 메모리의 명시적 동기화를 통해 상기 분산 딥러닝 파라미터를 공유하여 상기 분산 딥러

닝 트레이닝을 수행하고,

상기 원격 공유 메모리는

상기 마스터 프로세스에 의해 생성되고, 마스터 가중치 파라미터와 마스터 그래디언트 파라미터를 저장하는 마스터 영역; 및

상기 적어도 하나의 워커 프로세스의 개수에 상응하도록 생성되고, 적어도 하나의 워커 그래디언트 파라미터를 각각 저장하는 적어도 하나의 워커 영역;

을 포함하고,

상기 분산 딥러닝 트레이닝을 수행하는 단계는

상기 마스터 프로세스가, 상기 마스터 영역에 상기 마스터 가중치 파라미터와 상기 마스터 그래디언트 파라미터를 업데이트하고, 상기 적어도 하나의 워커 프로세스에게 상기 마스터 영역에 접근하기 위한 접근 정보를 전송하고,

상기 적어도 하나의 워커 프로세스가, 상기 접근 정보를 이용하여 상기 마스터 영역에 접근하여 상기 마스터 가중치 파라미터를 자신의 워커 가중치 파라미터로 업데이트하고,

상기 원격 공유 메모리의 적어도 하나의 워커 영역 중 자신이 생성한 워커 영역에 상기 분산 딥러닝 트레이닝을 수행한 결과로부터 학습된 워커 그래디언트 파라미터를 업데이트하고,

상기 파라미터 서버가, 상기 적어도 하나의 워커 프로세스로부터 상기 적어도 하나의 워커 영역에 상기 적어도 하나의 워커 그래디언트 파라미터가 업데이트되었음을 알림받으면, 상기 적어도 하나의 워커 그래디언트 파라미터를 상기 마스터 그래디언트 파라미터에 업데이트하고,

상기 마스터 프로세스가, 업데이트된 상기 마스터 그래디언트 파라미터를 이용하여 상기 마스터 가중치 파라미터를 업데이트하는 것을 특징으로 하는 분산 딥러닝 파라미터 공유 방법.

【직권보정 2】

【보정항목】 청구범위

【보정세부항목】 청구항 8

【변경전】

마스터 프로세스 및 적어도 하나의 워커 프로세스를 포함하는 분산 딥러닝 프로세스의 요청과 관련된 메시지를 송수신하는 통신 처리부,

상기 분산 딥러닝 프로세스의 요청에 상응하도록, 분산 딥러닝 파라미터를 저장하기 위한 원격 공유 메모리를 생성, 할당 및 해제하는 원격 공유 메모리 관리부, 그리고

상기 분산 딥러닝 프로세스가 상기 원격 공유 메모리를 통해 공유한 분산 딥러닝 파라미터를 이용하여 분산 딥러닝 트레이닝을 수행하는 파라미터 연산부를 포함하고,

상기 원격 공유 메모리 관리부는

상기 마스터 프로세스의 요청에 따라 상기 원격 공유 메모리를 생성하고, 상기 적어도 하나의 워커 프로세스에게 상기 원격 공유 메모리에 접근하기 위한 접근 정보를 전달하고,

상기 마스터 프로세스와 상기 적어도 하나의 워커 프로세스는

상기 원격 공유 메모리에 상응하는 각각의 로컬 물리 메모리를 할당하고, 상기 각각의 로컬 물리 메모리를 분산 딥러닝 트레이닝 엔진의 가상 주소 공간에 맵핑하고,

상기 파라미터 연산부는

상기 마스터 프로세스 및 상기 적어도 하나의 워커 프로세스와 함께, 상기 각각의 로컬 물리 메모리와 상기 원격 공유 메모리의 명시적 동기화를 통해 상기 분산 딥러닝 파라미터를 공유하여 상기 분산 딥러닝 트레이닝을 수행하고,

상기 원격 공유 메모리는

상기 마스터 프로세스에 의해 생성되고, 마스터 가중치 파라미터와 마스터 그래디언트 파라미터를 저장하는 마스터 영역; 및

상기 적어도 하나의 워커 프로세스의 개수에 상응하도록 생성되고, 적어도 하나의 워커 그래디언트 파라미터를 각각 저장하는 적어도 하나의 워커 영역;

을 포함하고,

상기 마스터 프로세스는

상기 마스터 영역에 상기 마스터 가중치 파라미터와 상기 마스터 그래디언트 파라미터를 업데이트하고,

상기 적어도 하나의 워커 프로세스에게 상기 마스터 영역에 접근하기 위한 접근 정보를 전송하고,

상기 적어도 하나의 워커 프로세스는

상기 접근 정보를 이용하여 상기 마스터 영역에 접근하여 상기 마스터 가중치 파라미터를 자신의 워커 가중치 파라미터로 업데이트하고,

상기 원격 공유 메모리의 적어도 하나의 워커 영역 중 자신이 생성한 워커 영역에 상기 분산 딥러닝 트레이닝을 수행한 결과로부터 학습된 워커 그래디언트 파라미터를 업데이트하고,

상기 파라미터 연산부는

상기 적어도 하나의 워커 프로세스로부터 상기 적어도 하나의 워커 영역에 상기 적어도 워커 그래디언트 파라미터가 업데이트되었음을 알림받으면, 상기 적어도 워커 그래디언트 파라미터를 상기 마스터 그래디언트 파라미터에 누적 연산하여 업데이트하고,

상기 마스터 서버는

업데이트된 상기 마스터 그래디언트 파라미터를 이용하여 상기 마스터 가중치 파라미터를 업데이트하는 것을 특징으로 하는 파라미터 서버.

【변경후】

마스터 프로세스 및 적어도 하나의 워커 프로세스를 포함하는 분산 딥러닝 프로세스의 요청과 관련된 메시지를 송수신하는 통신 처리부,

상기 분산 딥러닝 프로세스의 요청에 상응하도록, 분산 딥러닝 파라미터를 저장하기 위한 원격 공유 메모리를 생성, 할당 및 해제하는 원격 공유 메모리 관리부, 그리고

상기 분산 딥러닝 프로세스가 상기 원격 공유 메모리를 통해 공유한 분산 딥러닝 파라미터를 이용하여 분산 딥러닝 트레이닝을 수행하는 파라미터 연산부를 포함하고,

상기 원격 공유 메모리 관리부는

상기 마스터 프로세스의 요청에 따라 상기 원격 공유 메모리를 생성하고, 상기 적어도 하나의 워커 프로세스에게 상기 원격 공유 메모리에 접근하기 위한 접근 정보를 전달하고,

상기 마스터 프로세스와 상기 적어도 하나의 워커 프로세스는

상기 원격 공유 메모리에 상응하는 각각의 로컬 물리 메모리를 할당하고, 상기 각각의 로컬 물리 메모리를 분산 딥러닝 트레이닝 엔진의 가상 주소 공간에 맵핑하고,

상기 파라미터 연산부는

상기 마스터 프로세스 및 상기 적어도 하나의 워커 프로세스와 함께, 상기 각각의 로컬 물리 메모리와 상기 원격 공유 메모리의 명시적 동기화를 통해 상기 분산 딥러닝 파라미터를 공유하여 상기 분산 딥러닝 트레이닝을 수행하고,

상기 원격 공유 메모리는

상기 마스터 프로세스에 의해 생성되고, 마스터 가중치 파라미터와 마스터 그래디언트 파라미터를 저장하는 마스터 영역; 및

상기 적어도 하나의 워커 프로세스의 개수에 상응하도록 생성되고, 적어도 하나의 워커 그래디언트 파라미터를

각각 저장하는 적어도 하나의 워커 영역;

을 포함하고,

상기 마스터 프로세스는

상기 마스터 영역에 상기 마스터 가중치 파라미터와 상기 마스터 그래디언트 파라미터를 업데이트하고,

상기 적어도 하나의 워커 프로세스에게 상기 마스터 영역에 접근하기 위한 접근 정보를 전송하고,

상기 적어도 하나의 워커 프로세스는

상기 접근 정보를 이용하여 상기 마스터 영역에 접근하여 상기 마스터 가중치 파라미터를 자신의 워커 가중치 파라미터로 업데이트하고,

상기 원격 공유 메모리의 적어도 하나의 워커 영역 중 자신이 생성한 워커 영역에 상기 분산 딥러닝 트레이닝을 수행한 결과로부터 학습된 워커 그래디언트 파라미터를 업데이트하고,

상기 파라미터 연산부는

상기 적어도 하나의 워커 프로세스로부터 상기 적어도 하나의 워커 영역에 상기 적어도 하나의 워커 그래디언트 파라미터가 업데이트되었음을 알림받으면, 상기 적어도 하나의 워커 그래디언트 파라미터를 상기 마스터 그래디언트 파라미터에 누적 연산하여 업데이트하고,

상기 마스터 프로세스는

업데이트된 상기 마스터 그래디언트 파라미터를 이용하여 상기 마스터 가중치 파라미터를 업데이트하는 것을 특징으로 하는 파라미터 서버.

【직권보정 3】

【보정항목】청구범위

【보정세부항목】청구항 18

【변경전】

제8항에 있어서,

상기 마스터 프로세스 및 워커 프로세스는,

상기 원격 직접 메모리 접근(RDMA)을 지원하는 고속 네트워크를 통하여, 상기 파라미터 서버에 저장한 상기 분산 딥러닝 파라미터를 직접 읽어오거나 쓰는 방식으로 상기 분산 딥러닝 파라미터를 공유하는 것을 특징으로 하는 파라미터 서버.

【변경후】

제8항에 있어서,

상기 마스터 프로세스 및 워커 프로세스는,

원격 직접 메모리 접근(RDMA)을 지원하는 고속 네트워크를 통하여, 상기 파라미터 서버에 저장한 상기 분산 딥러닝 파라미터를 직접 읽어오거나 쓰는 방식으로 상기 분산 딥러닝 파라미터를 공유하는 것을 특징으로 하는 파라미터 서버.