(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년04월21일
(11) 등록번호 10-2524823
(24) 등록일자 2023년04월19일

(51) 국제특허분류(Int. Cl.)
*G06T 7/269* (2017.01) *G06N 3/04* (2023.01)
*G06N 3/08* (2023.01) *G06T 5/20* (2006.01)
*G06T 7/207* (2017.01)
(52) CPC특허분류
*G06T 7/269* (2017.01)
*G06N 3/04* (2023.01)
(21) 출원번호 10-2020-0168004
(22) 출원일자 2020년12월04일
심사청구일자 2020년12월04일
(65) 공개번호 10-2022-0078832
(43) 공개일자 2022년06월13일
(56) 선행기술조사문헌
Heng Wang 등, Video Modeling with Correlation Networks, arXiv:1906.03349v2 (2020.05.27.)*
*(뒷면에 계속)*

(73) 특허권자
**포항공과대학교 산학협력단**
경상북도 포항시 남구 청암로 77 (지곡동)
(72) 발명자
**조민수**
경상북도 포항시 남구 청암로 77, 포항공과대학교 컴퓨터공학과 (지곡동)
**권희승**
경상북도 포항시 북구 새천년대로1076번길 38, 305동 1102호 (두호동, 창포아이파크3차아파트)
*(뒷면에 계속)*
(74) 대리인
**특허법인(유한)아이시스**

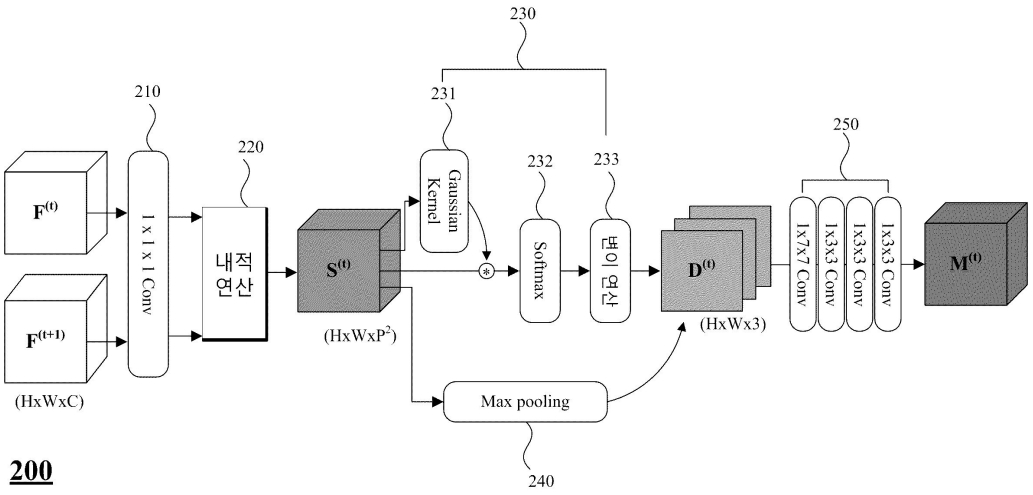
전체 청구항 수 : 총 16 항 심사관 : 이재원

(54) 발명의 명칭 **신경망 모델 기반 비디오의 움직임 특징 정보 추출 방법 및 분석장치**

## (57) 요 약

신경망 모델 기반 비디오의 움직임 특징 정보 추출 방법은 분석장치가 시간상 연속된 2개의 프레임들 각각에 대한 특징 맵을 획득하는 단계, 상기 분석장치가 상기 2개의 프레임들의 특징 맵들 사이의 상관관계를 나타내는 상관관계 텐서를 생성하는 단계, 상기 분석장치가 상기 상관관계 텐서를 기준으로 커널 기반 변이를 추정하여 변이 텐서를 생성하는 단계 및 상기 분석장치가 상기 변이 텐서를 컨볼루션 계층에 입력하여 움직임 특징 맵을 생성하는 단계를 포함한다.

## 대 표 도



(52) CPC특허분류
***G06N 3/08*** (2023.01)
***G06T 5/20*** (2023.01)
***G06T 7/207*** (2017.01)
*G06T 2207/10016* (2013.01)
*G06T 2207/20076* (2013.01)
*G06T 2207/20081* (2013.01)

(72) 발명자
**곽수하**
경상북도 포항시 남구 지곡로 155, 4동 201호 (지곡동, 교수아파트)
**김만진**
전라북도 전주시 완산구 호암로 40, 206동 1002호 (효자동2가, 골든팰리스휴먼시아아파트)

이 발명을 지원한 국가연구개발사업

| | |
|---|---|
| 과제고유번호 | 1711116284 |
| 과제번호 | 2017M3C4A7069369 |
| 부처명 | 과학기술정보통신부 |
| 과제관리(전문)기관명 | 한국연구재단 |
| 연구사업명 | 차세대정보·컴퓨팅기술개발 |
| 연구과제명 | 비전 모델 기반 공간 상황 인지 원천기술 연구 |
| 기 여 율 | 1/1 |
| 과제수행기관명 | 한양대학교 |
| 연구기간 | 2020.04.01 ~ 2020.12.31 |

공지예외적용 : 있음

(56) 선행기술조사문헌
Junghyup Lee 등, SFNet: Learning Object-aware Semantic Correspondence, arXiv:1904.01810v2 (2019.04.05.)*
KR1020200084467 A
KR102020062686 A
KR1020190024689 A
JP2016514867 A
*는 심사관에 의하여 인용된 문헌

# 명 세 서

## 청구범위

### 청구항 1

분석장치가 시간상 연속된 2개의 프레임들 각각에 대한 특징 맵을 획득하는 단계;

상기 분석장치가 상기 2개의 프레임들의 특징 맵들 사이의 상관관계를 나타내는 상관관계 텐서를 생성하는 단계;

상기 분석장치가 상기 상관관계 텐서를 기준으로 커널 기반 변이를 추정하여 변이 텐서를 생성하는 단계; 및

상기 분석장치가 상기 변이 텐서를 컨볼루션 계층에 입력하여 움직임 특징 맵을 생성하는 단계를 포함하되,

상기 변이 텐서를 생성하는 단계는

상기 분석장치는 상기 상관관계 텐서에 커널 기반 변이 추정 기법(kernel-soft-argmax)을 적용하여 채널별 변이 맵을 생성하는 단계;

상기 분석장치는 상기 상관관계 텐서에 풀링 연산을 하여 신뢰 맵을 생성하는 단계; 및

상기 분석장치는 상기 변이 맵과 상기 신뢰 맵을 결합하여 상기 변이 텐서를 생성하는 단계를 포함하는 신경망 모델 기반 비디오의 움직임 특징 정보 추출 방법.

### 청구항 2

제1항에 있어서,

상기 분석장치는 상기 2개의 프레임들의 특징맵들에서 동일 위치를 기준으로 변이에 대한 내적 연산(dot product)을 하여 상기 상관관계 텐서를 생성하는 신경망 모델 기반 비디오의 움직임 특징 정보 추출 방법.

### 청구항 3

제1항에 있어서,

상기 분석장치는 아래 수식을 이용하여 결정되는 상관관계 점수로 구성되는 상기 상관관계 텐서를 생성하는 신경망 모델 기반 비디오의 움직임 특징 정보 추출 방법.

$$s(\mathbf{x}, \mathbf{p}, t) = \mathbf{F}_{\mathbf{x}}^{(t)} \cdot \mathbf{F}_{\mathbf{x}+\mathbf{p}}^{(t+1)}$$

(s(x,p,t)는 상관관계 점수, t는 시간, x는 위치, p는 변이, ·는 내적 연산, $F^{(t)}$는 시간 t 프레임의 특징맵, $F^{(t)}$는 시간 t+1 프레임의 특징맵)

### 청구항 4

제3항에 있어서,

상기 상관관계 점수 연산에서 변이의 최대 범위를 $p \in [-k, k]^2$로 제한하는 신경망 모델 기반 비디오의 움직임 특징 정보 추출 방법.

(k는 정수)

### 청구항 5

제1항에 있어서,

상기 분석장치는

상기 상관관계 텐서에서 2D 가우시안 커널(Gaussian kernel)이 마스킹된 커널 기반 변이(kernel-soft-argmax)를 수행하여 상기 변이 텐서를 구성하는 변이 맵을 생성하는 신경망 모델 기반 비디오의 움직임 특징 정

보 추출 방법.

**청구항 6**

삭제

**청구항 7**

제1항에 있어서,

상기 분석장치는 아래 수식을 이용하여 상기 변이 맵을 생성하는 신경망 모델 기반 비디오의 움직임 특징 정보 추출 방법.

$$d(\mathbf{x},t)=\sum_{\mathbf{p}}\frac{\exp(g(\mathbf{x},\mathbf{p},t)s(\mathbf{x},\mathbf{p},t)/\tau)}{\sum_{\mathbf{p}'}\exp(g(\mathbf{x},\mathbf{p}',t)s(\mathbf{x},\mathbf{p}',t)/\tau)}\mathbf{p}$$

,

$$g(\mathbf{x},\mathbf{p},t)=\frac{1}{\sqrt{2\pi}\sigma}\exp(\frac{\mathbf{p}-\mathrm{argmax}_{\mathbf{p}}s(\mathbf{x},\mathbf{p},t)}{\sigma^2})$$

(d(x,t)는 시간 t 및 위치 x에 대한 변이값, g(x,p,t)는 가우시안 커널, s(x,p,t)는 상관관계 점수, τ는 소프트맥스 분포 조절 인자)

**청구항 8**

분석장치가 연속된 2개의 비디오 프레임들 각각을 컨볼루션 계층에 입력하여 상기 비디오 프레임들에 대한 제1 특징 맵 및 제2 특징 맵을 생성하는 단계;

상기 분석장치가 상기 제1 특징 맵과 상기 제2 특징 맵 사이의 변이(displacement) 나타내는 상관관계 텐서를 이용하여 변이 텐서를 생성하는 단계;

상기 분석장치가 상기 변이 텐서를 컨볼루션 계층에 입력하여 움직임 특징 맵을 생성하는 단계;

상기 분석장치가 상기 움직임 특징 맵과 상기 제1 특징 맵을 결합하여 최종 움직임 특징 맵을 생성하는 단계; 및

상기 분석장치가 상기 최종 움직임 특징 맵을 분류 계층에 입력하여 상기 비디오 프레임 내의 움직임을 추정하는 단계를 포함하되,

상기 변이 텐서를 생성하는 단계는

상기 분석장치는 상기 상관관계 텐서에 커널 기반 변이 추정 기법(kernel-soft-argmax)을 적용하여 채널별 변이 맵을 생성하는 단계;

상기 분석장치는 상기 상관관계 텐서에 풀링 연산을 하여 신뢰 맵을 생성하는 단계; 및

상기 분석장치는 상기 변이 맵과 상기 신뢰 맵을 결합하여 상기 변이 텐서를 생성하는 단계를 포함하는 신경망 모델 기반 비디오의 움직임 정보 추정 방법.

**청구항 9**

제8항에 있어서,

상기 분석장치는 상기 제1 특징 맵과 상기 제2 특징 맵에서 동일 위치를 기준으로 변이에 대한 내적 연산(dot product)을 하여 상기 상관관계 텐서를 생성하는 신경망 모델 기반 비디오의 움직임 정보 추정 방법.

**청구항 10**

제8항에 있어서,

상기 분석장치는 아래 수식을 이용하여 결정되는 상관관계 점수로 구성되는 상기 상관관계 텐서를 생성하는 신경망 모델 기반 비디오의 움직임 정보 추정 방법.

$$s(\mathbf{x},\mathbf{p},t) = \mathbf{F}_{\mathbf{x}}^{(t)} \cdot \mathbf{F}_{\mathbf{x}+\mathbf{p}}^{(t+1)}$$

(s(x,p,t)는 상관관계 점수, t는 시간, x는 위치, p는 변이, ·는 내적 연산, $F^{(t)}$는 시간 t 프레임의 제1 특징 맵, $F^{(t)}$는 시간 t+1 프레임의 제2 특징맵)

**청구항 11**

제8항에 있어서,

상기 상관관계 텐서에서 2D 가우시안 커널(Gaussian kernel)이 마스킹된 커널 기반 변이(kernel-soft-argmax)를 수행하여 상기 변이 텐서를 구성하는 변이 맵을 생성하는 신경망 모델 기반 비디오의 움직임 정보 추정 방법.

**청구항 12**

삭제

**청구항 13**

제8항에 있어서,

상기 분석장치는 아래 수식을 이용하여 상기 변이 맵을 생성하는 신경망 모델 기반 비디오의 움직임 정보 추정 방법.

$$d(\mathbf{x},t) = \sum_{\mathbf{p}} \frac{\exp(g(\mathbf{x},\mathbf{p},t)s(\mathbf{x},\mathbf{p},t)/\tau)}{\sum_{\mathbf{p}'}\exp(g(\mathbf{x},\mathbf{p}',t)s(\mathbf{x},\mathbf{p}',t)/\tau)}\mathbf{p},$$

$$g(\mathbf{x},\mathbf{p},t) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{\mathbf{p} - \mathrm{argmax}_{\mathbf{p}}s(\mathbf{x},\mathbf{p},t)}{\sigma^2}\right)$$

(d(x,t)는 시간 t 및 위치 x에 대한 변이값, g(x,p,t)는 가우시안 커널, s(x,p,t)는 상관관계 점수, τ는 소프트맥스 분포 조절 인자)

**청구항 14**

시간상 연속된 2개의 프레임들 각각에 대한 제1 특징 맵 및 제2 특징 맵을 입력받는 입력장치;

비디오를 구성하는 연속된 프레임들에 대한 특징 맵을 이용하여 비디오의 움직임 특징 맵을 생성하는 신경망 모델을 저장하는 저장장치; 및

상기 제1 특징 맵 및 상기 제2 특징 맵 사이의 상관관계를 나타내는 상관관계 텐서를 생성하고, 상기 상관관계 텐서를 기준으로 커널 기반 변이를 추정하여 변이 텐서를 생성하고, 상기 변이 텐서를 컨볼루션 계층에 입력하여 상기 2개의 프레임에 대한 움직임 특징 맵을 생성하는 연산장치를 포함하되,

상기 연산장치는 상기 상관관계 텐서에 커널 기반 변이 추정 기법(kernel-soft-argmax)을 적용하여 채널별 변이 맵을 생성하고, 상기 상관관계 텐서에 풀링 연산을 하여 신뢰 맵을 생성하고, 상기 변이 맵과 상기 신뢰 맵을 결합하여 상기 변이 텐서를 생성하는 신경망 모델 기반 비디오의 움직임 특징 정보를 추출하는 분석 장치.

**청구항 15**

제14항에 있어서,

상기 연산장치는 상기 제1 특징 맵 및 상기 제2 특징 맵에서 동일 위치를 기준으로 변이에 대한 내적 연산(dot product)을 하여 상기 상관관계 텐서를 생성하는 신경망 모델 기반 비디오의 움직임 특징 정보를 추출하는 분석 장치.

**청구항 16**

제14항에 있어서,

상기 연산장치는 아래 수식을 이용하여 결정되는 상관관계 점수로 구성되는 상기 상관관계 텐서를 생성하는 신경망 모델 기반 비디오의 움직임 특징 정보를 추출하는 분석 장치.

$$s(\mathbf{x},\mathbf{p},t) = \mathbf{F}_{\mathbf{x}}^{(t)} \cdot \mathbf{F}_{\mathbf{x}+\mathbf{p}}^{(t+1)}$$

(s(x,p,t)는 상관관계 점수, t는 시간, x는 위치, p는 변이, ·는 내적 연산, $F^{(t)}$는 시간 t 프레임의 특징맵, $F^{(t)}$는 시간 t+1 프레임의 특징맵)

**청구항 17**

제14항에 있어서,

상기 연산장치는 상기 상관관계 텐서에서 2D 가우시안 커널(Gaussian kernel)이 마스킹된 커널 기반 변이(kernel-soft-argmax)를 수행하여 상기 변이 텐서를 구성하는 변이 맵을 생성하는 신경망 모델 기반 비디오의 움직임 특징 정보를 추출하는 분석 장치.

**청구항 18**

삭제

**청구항 19**

제14항에 있어서,

상기 연산장치는 아래 수식을 이용하여 상기 변이 맵을 생성하는 신경망 모델 기반 비디오의 움직임 특징 정보를 추출하는 분석 장치.

$$d(\mathbf{x},t) = \sum_{\mathbf{p}} \frac{\exp(g(\mathbf{x},\mathbf{p},t)s(\mathbf{x},\mathbf{p},t)/\tau)}{\sum_{\mathbf{p}'}\exp(g(\mathbf{x},\mathbf{p}',t)s(\mathbf{x},\mathbf{p}',t)/\tau)}\mathbf{p}$$

,

$$g(\mathbf{x},\mathbf{p},t) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{\mathbf{p} - \operatorname{argmax}_{\mathbf{p}} s(\mathbf{x},\mathbf{p},t)}{\sigma^2}\right)$$

(d(x,t)는 시간 t 및 위치 x에 대한 변이값, g(x,p,t)는 가우시안 커널, s(x,p,t)는 상관관계 점수, τ는 소프트맥스 분포 조절 인자)

## *발명의 설명*

### *기 술 분 야*

[0001] 이하 설명하는 기술은 비디오에서 움직임 특징 정보를 추출하는 방법에 관한 것이다.

### *배 경 기 술*

[0002] 비디오 동작 인식 기술은 비디오를 통해 사람의 행동에 대해 이해하기 위한 기술이다. 기본적으로 비디오가 초당 25~30프레임의 이미지로 이루어져 있음에 기인해서, 비디오 동작 인식 기술은 이미지 인식 기술에 쓰이는 컴퓨터 비전 및 기계 학습 분야의 방법론을 활용한다.

[0003] 비디오 동작 인식 분야에서도 CNN(Convolutional Neural Network)과 같은 딥러닝 모델을 활용한 연구가 증가하고 있다.

[0004] 초창기 딥러닝 모델은 시공간 피처를 학습하는 데에 삼차원 컨볼루션 인공 신경망(3D CNN)들이 주로 연구되었다. 삼차원 컨볼루션 인공 신경망은 연산량이 매우 많기도 하고 비디오에서 움직임을 추정하는데 효율이 낮다. 현재 비디오 내의 움직임 특징을 학습하고 추정하기 위해서 대표적으로 옵티컬 플로우 (optical flow)가 사용되고 있다. 일반적으로 독립된 컨볼루션 인공 신경망을 두고 추출한 옵티컬 플로우를 입력으로 사용하여 움직임 특징을 학습한다. 이러한 방식은 일반적인 RGB 비디오 프레임을 입력으로 사용하는 인공 신경망의 추정값과의 융합을 통해 최종 결과를 도출하는데, 이를 투-스트림 (two-stream) 방식이라고 부른다.

## 선행기술문헌

### 특허문헌

[0005] (특허문헌 0001) 한국공개특허 제10-2019-0088087호

## 발명의 내용

### 해결하려는 과제

[0006] 투-스트림 방식은 움직임 특징을 효과적으로 학습하지만, 비디오 처리의 효율성을 낮은 편이다. 한편, 시각적 유사성 (visual correspondence)을 학습하는 방식은 컴퓨터 비전의 여러 분야에 적용되고 있는데 계산량이 너무 많거나 미분 연산이 어려워 활용도가 낮다.

[0007] 이하 설명하는 기술은 인공신경망을 활용한 비디오 움직임 특징 추출 방법을 제공하고자 한다. 이하 설명하는 기술은 움직임 정보를 추출하는 엔드투엔드(end-to-end) 모델을 제공하고자 한다.

### 과제의 해결 수단

[0008] 신경망 모델 기반 비디오의 움직임 특징 정보 추출 방법은 분석장치가 시간상 연속된 2개의 프레임들 각각에 대한 특징 맵을 획득하는 단계, 상기 분석장치가 상기 2개의 프레임들의 특징 맵들 사이의 상관관계를 나타내는 상관관계 텐서를 생성하는 단계, 상기 분석장치가 상기 상관관계 텐서를 기준으로 커널 기반 변이를 추정하여 변이 텐서를 생성하는 단계 및 상기 분석장치가 상기 변이 텐서를 컨볼루션 계층에 입력하여 움직임 특징 맵을 생성하는 단계를 포함한다.

[0009] 신경망 모델 기반 비디오의 움직임 특징 정보를 추출하는 분석 장치는 시간상 연속된 2개의 프레임들 각각에 대한 제1 특징 맵 및 제2 특징 맵을 입력받는 입력장치, 비디오를 구성하는 연속된 프레임들에 대한 특징 맵을 이용하여 비디오의 움직음 특징 맵을 생성하는 신경망 모델을 저장하는 저장장치 및 상기 제1 특징 맵 및 상기 제2 특징 맵 사이의 상관관계를 나타내는 상관관계 텐서를 생성하고, 상기 상관관계 텐서를 기준으로 커널 기반 변이를 추정하여 변이 텐서를 생성하고, 상기 변이 텐서를 컨볼루션 계층에 입력하여 상기 2개의 프레임에 대한 움직임 특징 맵을 생성하는 연산장치를 포함한다.

### 발명의 효과

[0010] 이하 설명하는 기술은 옵티컬 플로우 추출보다 훨씬 적은 양의 연산을 통해 움직임 정보를 생성한다. 또한, 이하 설명하는 기술은 기존 인공 신경망의 중간에 삽입된 형태로 활용이 가능하여 엔드투엔드 학습이 가능하고 파라미터의 개수도 매우 적다. 이하 설명하는 기술은 비디오 검색, CCTV 감시 시스템, 의료 영상 진단, 자율주행, 인간-로봇 인터랙션, 지능형 로봇 등의 다양한 분야에 활용될 수 있다.

### 도면의 간단한 설명

[0011] 도 1은 비디오에서 움직임을 추정하는 시스템의 예이다.

도 2는 연속된 특징 맵을 이용하여 움직임 정보를 생성하는 과정에 대한 예이다.

도 3은 비디오의 움직임을 인식하는 신경망 모델에 대한 예이다.

도 4는 분석장치의 구조에 대한 예이다.

### 발명을 실시하기 위한 구체적인 내용

[0012] 이하 설명하는 기술은 다양한 변경을 가할 수 있고 여러 가지 실시례를 가질 수 있는 바, 특정 실시례들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나, 이는 이하 설명하는 기술을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 이하 설명하는 기술의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0013] 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 해당 구성요소들은 상기 용어

들에 의해 한정되지는 않으며, 단지 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 이하 설명하는 기술의 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] 본 명세서에서 사용되는 용어에서 단수의 표현은 문맥상 명백하게 다르게 해석되지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 하고, "포함한다" 등의 용어는 설명된 특징, 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 의미하는 것이지, 하나 또는 그 이상의 다른 특징들이나 개수, 단계 동작 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 배제하지 않는 것으로 이해되어야 한다.

[0015] 도면에 대한 상세한 설명을 하기에 앞서, 본 명세서에서의 구성부들에 대한 구분은 각 구성부가 담당하는 주기능 별로 구분한 것에 불과함을 명확히 하고자 한다. 즉, 이하에서 설명할 2개 이상의 구성부가 하나의 구성부로 합쳐지거나 또는 하나의 구성부가 보다 세분화된 기능별로 2개 이상으로 분화되어 구비될 수도 있다. 그리고 이하에서 설명할 구성부 각각은 자신이 담당하는 주기능 이외에도 다른 구성부가 담당하는 기능 중 일부 또는 전부의 기능을 추가적으로 수행할 수도 있으며, 구성부 각각이 담당하는 주기능 중 일부 기능이 다른 구성부에 의해 전담되어 수행될 수도 있음은 물론이다.

[0016] 또, 방법 또는 동작 방법을 수행함에 있어서, 상기 방법을 이루는 각 과정들은 문맥상 명백하게 특정 순서를 기재하지 않은 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 과정들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.

[0017] 이하 설명하는 기술은 비디오를 분석하여 비디오에 포함된 움직임 정보를 분류하는 기술에 해당한다.

[0018] 입력 비디오를 처리하여 시간의 흐름에 따른 움직임 정보를 분류 내지 추정하는 장치를 분석장치라고 명명한다. 분석장치는 데이터 처리 및 연산이 가능한 컴퓨팅 장치이다. 예컨대, 분석장치는 서버, PC, 스마트기기, 프로그램이 임베디드된 칩 등일 수 있다.

[0019] 분석장치는 인공신경망 모델을 이용하여 입력 비디오에서 움직임 정보를 분류한다.

[0020] 도 1은 비디오에서 움직임을 추정하는 시스템(100)의 예이다. 입력데이터(110)는 복수의 프레임들(연속된 프레임들)이다. 입력데이터(110)는 기본적으로 2개 이상의 연속된 프레임들이다. 분석장치(120)는 입력되는 비디오 프레임들을 분석하여 해당 비디오에 포함된 객체 또는 객체의 동작을 인식하거나 추정할 수 있다. 분석장치(120)는 다양한 형태로 구현될 수 있다. 예컨대, 분석장치(120)는 개인 PC, 스마트기기, 네트워크상의 서버 등과 같은 장치일 수 있다. 분석장치(120)는 일정한 신경망 모델(125)을 이용하여 비디오 움직임을 인식 내지 추정한다. 신경망 모델(125)은 다양한 형태를 가질 수 있고, 연속된 비디오 프레임을 입력받아 해당 비디오 프레임에 포함된 객체 또는 객체의 동작에 대한 추정 정보를 출력한다. 분석장치(120)는 신경망 모델(125)을 이용한 분석 결과를 출력하거나, 외부 객체에 전달할 수 있다.

[0022] 비디오 내의 움직임 특징을 생성하기 위해서, 옵티컬 플로우가 주로 사용된다. 옵티컬 플로우는 움직임 정보를 효과적으로 학습하지만, 비디오 처리 자체는 복잡도가 높다. 이하 비디오에서 움직임 특징을 낮은 복잡도로 효과적으로 추출하는 기법을 설명한다.

[0024] 도 2는 연속된 특징 맵을 이용하여 움직임 정보를 생성하는 과정(200)에 대한 예이다. 도 2는 분석 장치가 연속된 특징 맵들을 분석하여, 움직임 정보를 나타내는 움직임 특징 맵을 생성하는 과정이다. 움직임 특징 맵을 추출하는 구성을 움직임 정보 추출 계층이라고 명명한다. 도 2는 움직임 정보 추출 계층의 동작을 나타낸다. 도 2에 도시한 움직임 정보 추출 계층은 입력되는 특징맵에서 움직임 정보를 출력하는 신경망 모델이다. 따라서, 도 2에 도시한 신경망 모델은 움직임 정보 추출 모델이라고 명명할 수도 있다.

[0025] 움직임 정보 추출 계층의 입력 데이터는 두 개의 인접한 특징 맵 $F^{(t)}$ 및 $F^{(t+1)}$이다. t는 시간을 의미한다. 따라서, $F^{(t)}$ 및 $F^{(t+1)}$는 시간축에서 연속된 프레임에 대한 특징 맵이다. 각 특징맵은 H×W×C 크기의 3D 텐서(tensors)이다. 공간 해상도는 H(Height)×W(Width)이고, C는 채널의 개수이다.

[0026] 분석장치는 입력되는 특징 맵 $F^{(t)}$ 및 $F^{(t+1)}$을 각각 특정 크기의 컨볼루션 필터를 적용하여 크기를 줄일 수 있다(210). 도 2는 입력되는 특징 맵에 1×1×1 컨볼루션 계층을 적용한 예이다.

[0027] 시간 t에서 위치 x의 변이(displacement) p에 대한 상관관계 점수(correlation score) s(x,p,t)는 아래 수학식

1과 같이 표현가능한다. 분석장치는 두 개의 특징 맵에 대하여 내적 연산을 수행한다(220).

*수학식 1*

[0029] $$s(\mathbf{x},\mathbf{p},t)=\mathbf{F}_{\mathbf{x}}^{(t)}\cdot\mathbf{F}_{\mathbf{x}+\mathbf{p}}^{(t+1)}$$

[0030] ·는 내적 연산(dot product)이다. 연산 효율을 위하여 최대 변이를 $p \in [-k, k]^2$로 제한하고, 상관 점수는 위치 x의 주변 P = 2k+1 범위에서 연산한다. k는 정수이다.

[0032] 두 개의 특징 맵 $F^{(t)}$ 및 $F^{(t+1)}$을 입력받아 t 번째 프레임을 기준으로 $H\times W\times P^2$ 크기의 상관관계 텐서(correlation tensor) S(t)가 생성된다. 상관관계 텐서 연산 복잡도는 $P^2$ 커널로 1×1 컨볼루션하는 것과 같다. 상관 연산은 $P^2$ 커널로 t+1 번째 특징 맵을 이용하여 t 번째 특징 맵에 대한 2D 컨볼루션을 하는 것이다.

[0034] 분석 장치는 상관관계 텐서 S(t)로부터, 움직임 정보에 대한 변이 필드(displacement filed)를 추정한다.

[0035] 변이 값을 추정하는 직접적인 방법은 상관관계 텐서에 $\text{argmax}_p s(x,p,t)$를 취해 가장 상관관계 값이 높은 위치를 추출하는 것이다. 그러나 argmax는 미분이 어렵기 때문에, 미분 가능한 변이 추정 연산으로 가중평균 변이 추정 기법을 이용할 수 있다. 예컨대, soft-argmax를 이용할 수 있고, 이는 아래 수학식 2와 같이 표현할 수 있다.

*수학식 2*

[0037] $$d(\mathbf{x},t)=\sum_{\mathbf{p}}\frac{\exp(s(\mathbf{x},\mathbf{p},t))}{\sum_{\mathbf{p}'}\exp(s(\mathbf{x},\mathbf{p}',t))}\mathbf{p}$$

[0039] 수학식 2는 모든 상관관계 값을 사용하기 때문에, 상관관계 텐서에서의 이상값(outliers)에 민감하다. 따라서, 아래의 수학식 3으로 표현되는 커널 기반 변이 추정 기법(kernel-soft-argmax)을 사용할 수 있다. 상관관계 텐서 S(t)를 기준으로 커널 기반 변이 추정하여 변이 맵을 생성한다. kernel-soft-argmax는 상관관계 값에 2D 가우시안 커널(Gaussian kernel)을 마스킹(masking)하여 이상값을 억제한다. 커널은 각 목표 위치의 중심에 위치하여 주변값중 가까운 값일 수록 더 큰 영향을 받게 한다.

*수학식 3*

[0041] $$d(\mathbf{x},t)=\sum_{\mathbf{p}}\frac{\exp(g(\mathbf{x},\mathbf{p},t)s(\mathbf{x},\mathbf{p},t)/\tau)}{\sum_{\mathbf{p}'}\exp(g(\mathbf{x},\mathbf{p}',t)s(\mathbf{x},\mathbf{p}',t)/\tau)}\mathbf{p}$$

[0042] g(x,p,t)는 가우시안 커널을 나타내며 아래 수학식 4와 같이 표현된다. τ는 소프트맥스 분포를 조절하는 인자이다.

*수학식 4*

[0044] $$g(\mathbf{x},\mathbf{p},t)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{\mathbf{p}-\text{argmax}_{\mathbf{p}}s(\mathbf{x},\mathbf{p},t)}{\sigma^2}\right)$$

[0046] 분석장치는 상관관계 텐서 S(t)에 대하여 변이 맵을 생성한다(230). 분석장치는 가우시안 커널(213) 및 소프트맥스(212)를 적용하고, 상기 수학식 3을 이용하여 변이 연산(233)을 할 수 있다.

[0048] 분석장치는 부가적인 움직임 정보로 상관관계 신뢰 맵(confidence map)을 사용할 수 있다. 분석장치는 상관관계

텐서 S(t)에 대한 풀링(pooling) 연산을 하여 신뢰 맵을 생성할 수 있다(240). 예컨대, 아래 수학식 5와 같이 각 위치 x에서 가장 높은 상관관계 값을 추출하는 최대 풀링으로 신뢰 맵을 생성할 수 있다.

*수학식 5*

[0050] $$s^*(\mathbf{x},t) = \max_{\mathbf{p}} s(\mathbf{x},\mathbf{p},t)$$

[0051] 신뢰도 맵은 변이 값의 이상치(outlier)를 식별하고, 움직임 특징에 대한 정보를 학습하는데 도움을 준다.

[0053] 분석장치는 2채널 변이 맵과 1채널 신뢰 맵을 결합(concatenation)하여 H×W×3 크기의 변이 텐서 D(t)를 생성한다.

[0054] 분석장치는 변이 텐서 D(t)를 이용하여 움직임 특징 맵을 생성한다. 분석장치는 변이 텐서 D(t)를 복수의 컨볼루션 계층들로 구성된 움직임 추출 계층에 입력하여 움직임 특징 맵을 생성할 수 있다(250). 복수의 컨볼루션 계층들은 변이 텐서 D(t)를 입력받아 움직임 정보를 나타내는 최종 움직임 특징 $M^{(t)}$를 출력한다. 도 2에서 복수의 컨볼루션 계층들은 4 깊이별 분리 컨볼루션 계층 depth-wise separable convolution layers)이다. 복수의 컨볼루션 계층들은 하나의 1×7×7 계층과 3개의 1×3×3 계층들로 구성된 예이다. 복수의 컨볼루션 계층들은 변이 텐서 D(t)를 입력받아 특징 맵 F(t)와 같은 채널 개수 C 갖는 움직임 특징 맵 M(t)를 출력한다. 각 컨볼루션 계층은 배치 정규화 및 ReLU 연산을 포함한다. 물론 움직임 추출 계층은 도 2와 다른 구조를 가질 수도 있다.

[0056] 도 2의 움직임 정보 추출 계층(움직임 정보 추출 모델)은 비디오의 움직임 정보를 사용하는 다양한 모델에 사용될 수 있다. 예컨대, 비디오의 움직임 정보를 이용하는 다양한 모델에 도 2의 움직임 정보 추출 모델을 삽입하여 활용할 수 있다. 예컨대, 비디오에 포함된 객체의 미래 동작을 추정하는 모델은 움직임 정보 추출 모델에서 출력되는 움직임 특징 맵을 기준으로 미래 동작을 추정할 수 있다.

[0057] 도 3은 비디오의 움직임을 인식하는 신경망 모델(300)에 대한 예이다. 도 3은 도 2의 움직임 정보 추출 계층을 이용한 움직임 인식의 예이다.

[0058] 신경망 모델(300)은 입력 비디오의 특징을 추출하는 특징 추출 계층(310), 움직임 정보 추출 계층(320) 및 움직임 추정 계층(330)을 포함한다.

[0059] 특징 추출 계층(310)은 비디오에서 시간상 연속된 프레임을 입력받아 각 프레임에 대한 특징 맵을 생성한다. 특징 추출 계층은 컨볼루션 계층들로 구성될 수 있다.

[0060] 움직임 정보 추출 계층(320)은 도 2에서 설명한 구성(200)과 같다. 프레임 t를 기준으로 설명하면, 움직임 정보 추출 계층(320)은 시간 t의 특징 맵 $F^{(t)}$ 및 $F^{(t+1)}$을 입력받아 시간 t의 프레임 기준으로 움직임 특징 맵 $M^{(t)}$을 생성한다. 아래 수학식 6과 같이 분석장치는 시간 t의 특징 맵 $F^{(t)}$와 움직임 특징 맵 $M^{(t)}$를 합산(add)하여 결합된 특징 맵 $F'^{(t)}$를 생성한다.

*수학식 6*

[0062] $$\mathbf{F}'^{(t)} = \mathbf{F}^{(t)} + \mathbf{M}^{(t)}$$

[0064] 한편, 분석장치는 특징 맵 $F^{(t)}$와 움직임 특징 맵 $M^{(t)}$에 대한 다른 연산으로 특징 맵 $F'^{(t)}$를 생성할 수도 있다. 예컨대, 분석장치는 결합(concatenation), 곱셈(multiplication) 연산을 이용하여 특징 맵 $F'^{(t)}$를 생성할 수도 있다.

[0066] 움직임 추정 계층(320)은 프레임 t에 대하여 $F^{(t)}$와 $M^{(t)}$를 결합한 특징 맵을 입력받고, 해당 프레임에서의 움직임을 분류 내지 추정한다. 마지막 프레임 T에 대해서는 $M^{(T)} = M^{(T-1)}$로 설정하여 처리할 수 있다. 움직임 추정 계

층(320)은 입력되는 특징 맵을 기준으로 영상에서 객체 분류, 객체 동작 인식 등을 할 수 있다. 움직임 추정 계층(320)은 다양한 구조의 신경망 모델이 이용될 수 있다. 예컨대, 움직임 추정 계층(320)은 컨볼루션 계층 및 최종 특징을 기준으로 객체를 분류하는 활성 함수 등을 포함할 수 있다. 움직임 추정 계층(320)은 전연결계층(FC)을 포함할 수도 있다.

[0068] 도 4는 분석장치(400)의 구조에 대한 예이다. 분석장치(400)는 도 1의 분석장치(120)에 해당한다.

[0069] 분석장치(400)는 저장 장치(410), 메모리(420), 연산장치(420) 및 인터페이스 장치(430)를 포함한다. 나아가, 분석장치(400)는 통신장치(450) 및 출력장치(460)를 더 포함할 수도 있다.

[0070] 저장 장치(410)는 입력되는 비디오를 저장할 수 있다.

[0071] 저장 장치(410)는 비디오 프레임들에 대한 특징 맵 F를 저장할 수 있다.

[0072] 저장 장치(410)는 움직임 정보 추출 모델(움직임 정보 추출 계층)을 저장할 수 있다.

[0073] 저장 장치(410)는 움직임 정보 추출 모델이 출력하는 움직임 특징 맵 M을 저정할 수 있다.

[0074] 저장 장치(410)는 움직임 정보 추출 모델을 포함하는 비디오 움직임 인식 모델을 저장할 수도 있다. 즉, 저장 장치(410)는 움직임 정보 추출 모델을 이용하는 다른 애플리케이션 모델 내지 프로그램을 저장할 수 있다.

[0075] 메모리(420)는 비디오 데이터 처리 과정에서 생성되거나 필요한 정보를 임시로 저장할 수 있다.

[0076] 인터페이스 장치(440)는 분석장치에 데이터 및 명령을 전달하는 구성을 의미한다. 인터페이스 장치(440)는 내부 통신을 위한 물리적 장치 및 통신 프로토콜을 포함할 수 있다. 인터페이스 장치(440)는 입력 비디오를 입력받을 수 있다. 인터페이스 장치(440)는 다른 장치 내지 모델이 추정한 비디오 프레임의 특징 맵을 입력받을 수도 있다. 인터페이스 장치(440)는 움직임 추정을 위한 명령 내지 파라미터를 입력받을 수도 있다.

[0077] 통신장치(450)는 유선 또는 무선 통신을 통해 외부 객체로부터 일정한 정보를 수신할 수 있다. 예컨대, 통신장치(400)는 입력 비디오를 수신할 수 있다. 통신장치(450)는 비디오 프레임의 특징 맵을 수신할 수 있다. 통신장치(450)는 움직임 정보 추출 모델이 출력하는 움직임 특징 맵 M을 외부 객체로 송신할 수 있다. 통신장치(450)는 추정된 움직임 특징 맵을 이용한 애플리케이션이 출력하는 분석 결과를 외부 객체로 송신할 수 있다.

[0078] 연산장치(430)는 주어진 데이터 내지 정보를 처리하는 구성을 의미한다. 연산장치(430)는 프로세서, AP, 프로그램이 임베디드된 칩과 같은 장치일 수 있다.

[0079] 연산장치(430)는 저장 장치(410)에 저장된 프로그램 및 모델을 이용하여 비디오 프레임들의 특징맵으로부터 움직임 정보(움직임 특징 맵)를 생성할 수 있다.

[0080] 연산장치(430)는 연속된 비디오 프레임 t 및 비디오 프레임 t+1 각각에 대한 특징맵을 생성할 수 있다. 연산장치(430)는 비디오 프레임 t 및 비디오 프레임 t+1를 각각 컨볼루션 계층에 입력하여 특징 맵 $F^{(t)}$와 특징 맵 $F^{(t+1)}$을 생성할 수 있다.

[0081] 연산장치(430)는 특징 맵 $F^{(t)}$와 특징 맵 $F^{(t+1)}$을 이용하여 상관관계 맵 $S^{(t)}$를 생성할 수 있다. 연산장치(430)는 수학식 1을 이용하여 상관관계 맵 $S^{(t)}$를 생성할 수 있다.

[0082] 연산장치(430)는 상관관계 맵 $S^{(t)}$를 기준으로 변이 맵 $D^{(t)}$를 생성할 수 있다. 연산장치(430)는 수학식 3을 이용하여 상관관계 맵 $S^{(t)}$으로부터 변이 맵을 생성하고, 수학식 5의 풀링 연산을 통해 생성된 신뢰 맵을 결합하여 최종 변이 맵 $D^{(t)}$를 생성할 수 있다.

[0083] 연산장치(430)는 변이 맵 $D^{(t)}$에서 움직임 특징을 추출하여 움직임 특징 맵 $M^{(t)}$를 생성할 수 있다.

[0084] 또한, 연산장치(430)는 움직임 정보 추출 모델을 포함하는 애플리케이션 모델을 구동하여 입력 비디오에 대한 분석 내지 추정을 할 수도 있다.

[0085] 출력장치(460)는 움직임 정보 추출 모델이 생성한 움직임 정보를 활용한 분석 결과(예컨대, 미래 영상)를 출력할 수 있다.

[0087] 또한, 상술한 바와 같은 움직임 특징 맵 생성 방법 내지 비디오 움직임 정보 추정 방법은 컴퓨터에서 실행될 수

있는 실행가능한 알고리즘을 포함하는 프로그램(또는 어플리케이션)으로 구현될 수 있다. 상기 프로그램은 일시적 또는 비일시적 판독 가능 매체(non-transitory computer readable medium)에 저장되어 제공될 수 있다.

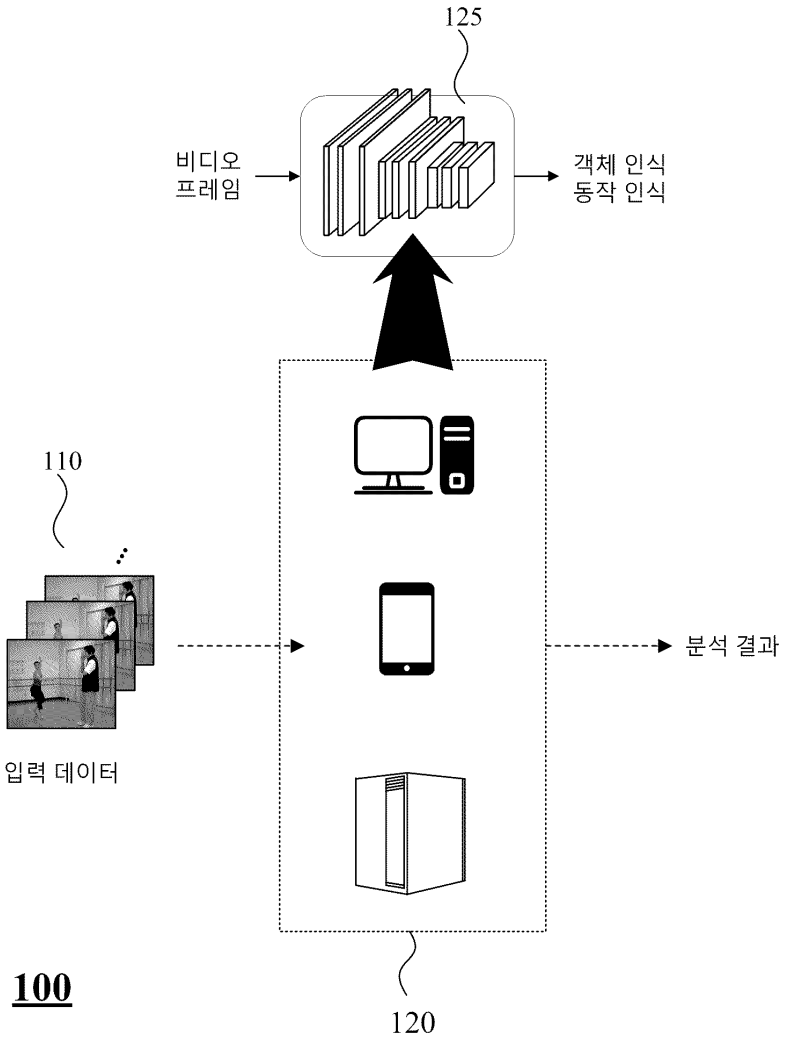
[0088] 비일시적 판독 가능 매체란 레지스터, 캐쉬, 메모리 등과 같이 짧은 순간 동안 데이터를 저장하는 매체가 아니라 반영구적으로 데이터를 저장하며, 기기에 의해 판독(reading)이 가능한 매체를 의미한다. 구체적으로는, 상술한 다양한 어플리케이션 또는 프로그램들은 CD, DVD, 하드 디스크, 블루레이 디스크, USB, 메모리카드, ROM(read-only memory), PROM (programmable read only memory), EPROM(Erasable PROM, EPROM) 또는 EEPROM(Electrically EPROM) 또는 플래시 메모리 등과 같은 비일시적 판독 가능 매체에 저장되어 제공될 수 있다.

[0089] 일시적 판독 가능 매체는 스태틱 램(Static RAM, SRAM), 다이내믹 램(Dynamic RAM, DRAM), 싱크로너스 디램(Synchronous DRAM, SDRAM), 2배속 SDRAM(Double Data Rate SDRAM, DDR SDRAM), 증강형 SDRAM(Enhanced SDRAM, ESDRAM), 동기화 DRAM(Synclink DRAM, SLDRAM) 및 직접 램버스 램(Direct Rambus RAM, DRRAM) 과 같은 다양한 RAM을 의미한다.

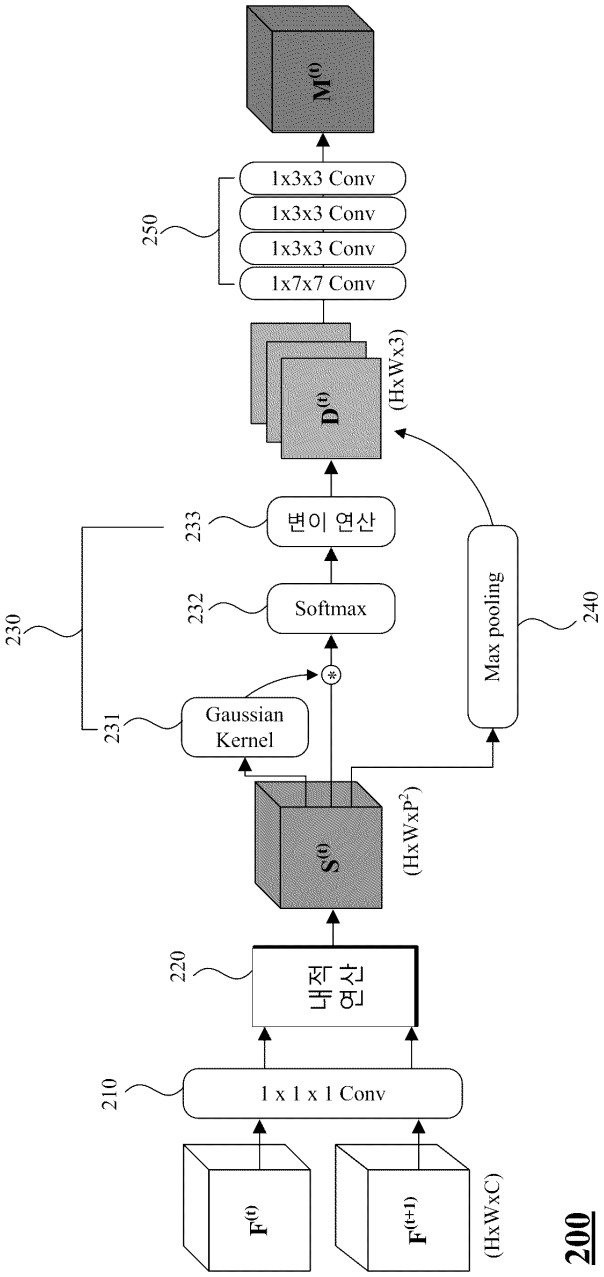
[0090] 본 실시례 및 본 명세서에 첨부된 도면은 전술한 기술에 포함되는 기술적 사상의 일부를 명확하게 나타내고 있는 것에 불과하며, 전술한 기술의 명세서 및 도면에 포함된 기술적 사상의 범위 내에서 당업자가 용이하게 유추할 수 있는 변형 예와 구체적인 실시례는 모두 전술한 기술의 권리범위에 포함되는 것이 자명하다고 할 것이다.

## 도면

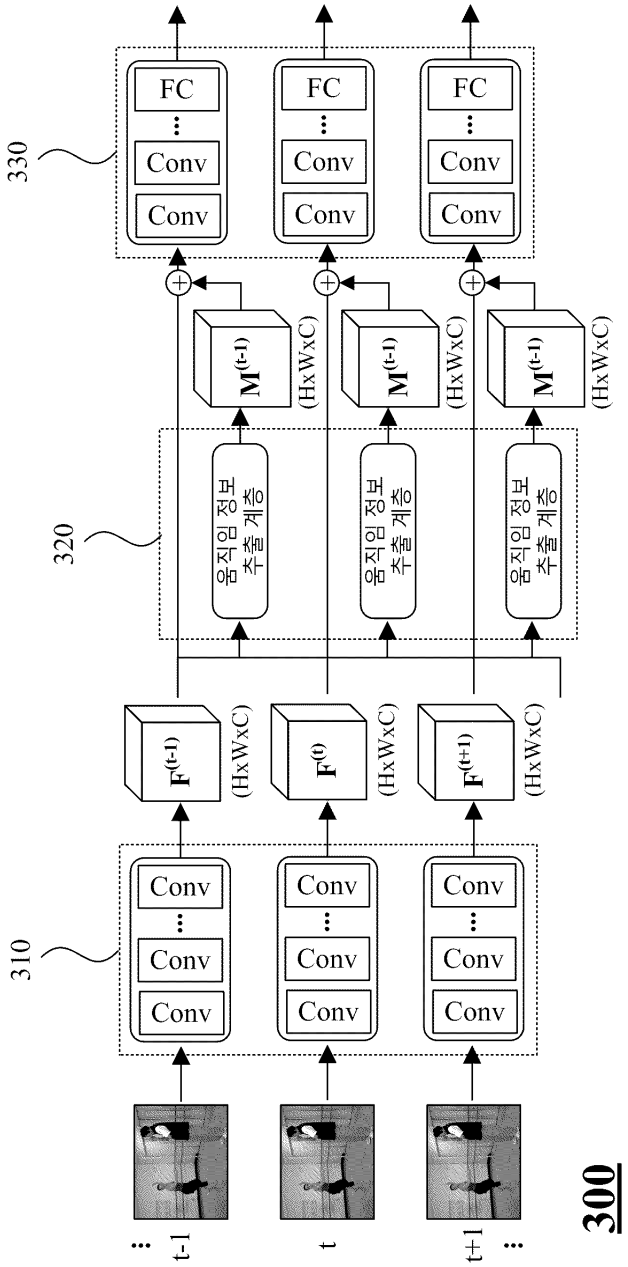
### 도면1



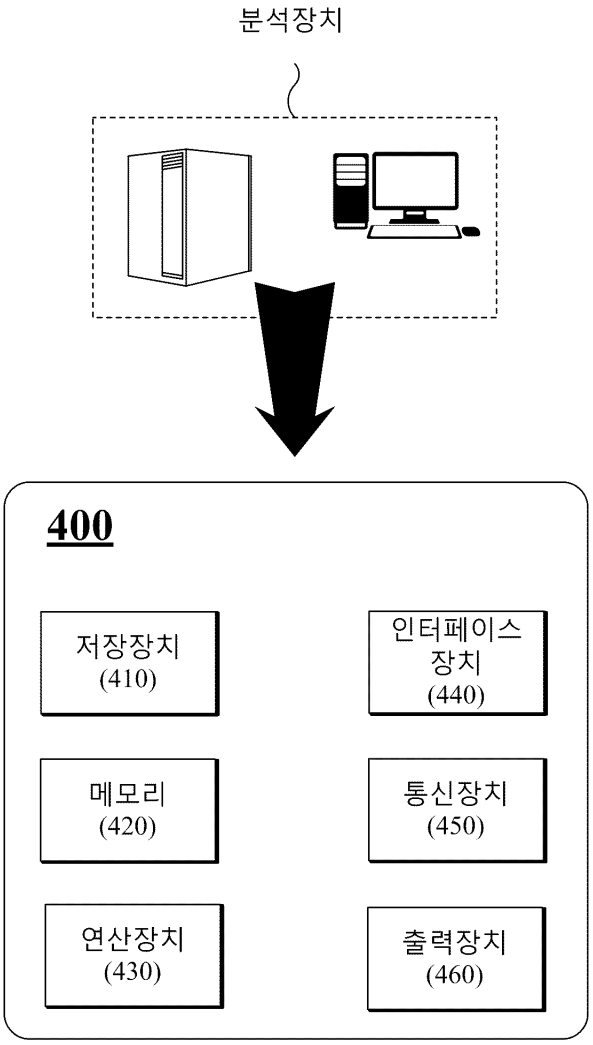
도면2


M(t)
1x3x3 Conv
1x3x3 Conv
1x3x3 Conv
1x7x7 Conv
250
D(t)
(HxWx3)
변이 연산
233
Softmax
232
230
Gaussian Kernel
231
Max pooling
240
S(t)
(HxWxP2)
220
1 x 1 x 1 Conv
210
F(t)
F(t+1)
(HxWxC)
200

도면3


330
FC
Conv
Conv
320
움직임 정보 추출 계층
M(t-1) (HxWxC)
310
F(t-1) (HxWxC)
F(t) (HxWxC)
F(t+1) (HxWxC)
t-1
t
t+1
300

도면4


분석장치
400
저장장치
(410)
인터페이스
장치
(440)
메모리
(420)
통신장치
(450)
연산장치
(430)
출력장치
(460)